

A hierarchical random-effects model for survival in patients with Acute Myocardial Infarction

Alessandra Guglielmi, Francesca Ieva, Anna M. Paganoni and Fabrizio Ruggeri

Abstract Studies of variations in health care utilization and outcome involve the analysis of multi-level clustered data. These analyses involve estimation of a cluster-specific adjusted response, covariates effect and components of variance. Beyond reporting on the extent of observed variations, these studies examine the role of contributing factors including patients and providers characteristics. In addition, they may assess the relationship between health-care process and outcomes. In this article we present a case-study, considering firstly a Hierarchical Generalized Linear Model (HGLM) formulation, then a semi-parametric Dirichlet Process Mixtures (DPM), and propose their application to the analysis of MOMI² (MOnth MOnitoring Myocardial Infarction in Milan) study on patients admitted with ST-Elevation Myocardial Infarction diagnosis. We develop a Bayesian approach to fitting data using Markov Chain Monte Carlo methods and discuss some issues about model fitting.

Keywords: Hierarchical models, Multilevel data analysis, Statistical modeling, Biostatistics and bioinformatics

1 Introduction

Over recent years there has been a growing interest in the use of performance indicators in health-care; they may measure some aspects of the health-care process, clinical outcomes or disease incidence. The purpose of the present work is to show how advanced statistical methods can be employed in the analysis of complex data coming from such clinical registers. Several examples, available in clinical literature (see for example [15]), make use of clinical registers to evaluate performances of medical institutions, because they enable people concerned with the health-care governance to plan activities on real epidemiological evidence and needs and evaluate performance of structures they manage.

The disease we are interested in is the Acute Myocardial Infarction with ST-segment Elevation (STEMI): it consists of a stenotic plaque detachment, which causes a coronary thrombosis and a sudden critical reduction of blood flow in coronary vessels. This process causes a widespread necrosis of myocardial tissues and leads to an inadequate feeding of myocardial muscle itself. STEMI is characterized by a great incidence (650 - 700 events per month have been estimated only in Lombardia Region) and serious mortality (Italy 8%). Concerning infarction, both survival and quantity of myocardial tissue saved from damage depend strongly on

Alessandra Guglielmi, Francesca Ieva, Anna Maria Paganoni
Dipartimento di Matematica, Politecnico di Milano, via Bonardi 9, 20133 Milano (Italy)
e-mail: alessandra.guglielmi@polimi.it; francesca.ieva@mail.polimi.it; anna.paganoni@polimi.it
Fabrizio Ruggeri
CNR IMATI, via Bassini 15, 20133 Milano (Italy) e-mail: fabrizio@mi.imati.cnr.it

time saved during the process. In this work, we will focus on survival outcome. Time has indeed a fundamental role in the STEMI health-care process. Let us call “Symptom Onset to Door time” the time since symptoms onset up to the arrival at ER, and “Door to Balloon time” (DB time) the time since the arrival at ER up to the surgical practice of Percutaneous Transluminal Coronary Angioplasty (PTCA). Clinical literature strongly stresses the connection between in-hospital survival and procedures time (see [3], [12], [14]): 90 minutes for Door to Balloon time in case of primary PTCA is the actual gold standard limit suggested by the American Heart Association (AHA)/American College of Cardiology (ACC) guidelines (see [2]). Anyway, the presence of substantial variations in the delivery of outcome of health-care has been documented extensively in recent years. In subsequent steps, researchers study the effects of variations in health-care use on patients outcomes; they examine the relationship between measure of the process, which define regional or hospital practice patterns, and measure of outcome, such as mortality and so on.

Data on health-care use have a multilevel structure, usually with patients at first level and hospitals forming the upper-level clusters. A key-analytic goal is to provide cluster-specific estimates of a particular response, adjusted for patients characteristic. Another key-goal is to derive estimates of covariates effects, such as differences between patients of different gender or between hospitals. Hierarchical regression modelling provides a general analytic approach that can accomplish these goals. The main purpose of the article is to set in a Bayesian framework the problem of modelling the outcome by means of relevant covariates, taking into account overdispersion induced by the grouping factor. We illustrate the analysis on a subset of data collected in the MOMI² survey on patients admitted with ST-Elevation Myocardial Infarction (STEMI) diagnosis in one of the structures belonging to the Milan Cardiological Network. For this analysis, patients are grouped by the hospital they are admitted to for their infarction. We use Markov Chain Monte Carlo simulation of posterior distribution of parameters and discuss model checking proposed by Albert and Chib [1]. Finally, we are actually considering the application of Dirichlet Process Mixture to the model, in order to find out a hospital classification based on Bayesian non parametric techniques (for further details see [8]).

2 The dataset

A net connecting the territory to hospitals, by a centralized coordination of the emergency resources, has been activated in the Milan urban area since 2001. The aim of this organizational policy is the setting of a register on STEMI to collect also process indicators (Symptom Onset time, first ECG time, DB time and so on), in order to identify and develop new diagnostic, therapeutic and organizational strategies to be applied by Lombardia Region, hospitals and 118 organization (the National free number for medical emergencies). To reach this goal, it is necessary to understand which organizational aspects can be considered as predictive in reducing time to treatment. In fact, organizational policies in STEMI health-care process include both 118 and hospital, since a subject affected by an infarction can move to the hospital by himself or can be moved to the hospital by 118 rescue units. In order to monitor Milan

Cardiological Network activity, times to treatment and clinical outcomes, the data collection MOMI² was planned and made, during six periods corresponding to six monthly/bimstral collections, on STEMI patients admitted in one of the hospitals belonging to the Network. For these units, information concerning way of admission, demographic features (sex, age), clinical appearance (presenting symptoms, Killip class at admittance, which quantify in four categories the severity of infarction), received therapy, Symptom onset times, in-hospital times (first ECG times, DB times), hospital organization (Fast Track, admission during On/Off hours) and clinical outcome (in-hospital survival) have been listed and studied.

The whole MOMI² survey consists of 840 statistical units, but in this work we only focus on patients undergone primary PTCA (i.e. PTCA without any previous pharmacological treatment, which are 537) and belonging to the third (Jun 1st - Jul 31st 2007, 154 pt.) and fourth (Nov 15st - Dec 15th 2007, 93 pt.) collections only. Among the resulting 247 PTCA-patients, we selected those who had their own hospital admission registered also in the Public Health Database of Lombardia Region, in order to confirm the reliability of information collected in the MOMI² register. Therefore, the dataset finally considered consists of 240 patients.

Previous analyses on MOMI² survey (for further details see , [9], [10] and [11]) pointed out age, the logarithm of Symptom Onset to Balloon time (logOB) and Killip of patient as significant factors in order to explain survival probability from a statistical and clinical point of view.

3 Model and Prior Elicitation

We considered a generalized mixed-effects model for binary data from a Bayesian viewpoint. For any patient $i = 1, \dots, 240$, Y_i is a Bernoulli random variable with mean p_i , which represents the probability that the i -th patient survived after a STEMI event. In the rest of the paper, y denotes the vector of all responses $(y_1, \dots, y_n)$. The p_i 's are modelled through a logit regression with covariates, $\mathbf{x}_i := (1, x_{i1}, x_{i2}, x_{i3})$ which represent the age, the Symptom Onset to Ballon time in the log scale (logOB) and the Killip class, respectively, of the i -th patient in the dataset. The Killip class is binary here, corresponding to 0 for less severe (Killip class equal to 1 or 2) and 1 for more severe (Killip class equal to 3 or 4) infarction. Moreover, the age and the logOB have been centered. This model has been elicited in [10]. Since the patients comes from $J = 17$ different hospitals, we assume the following multilevel model, with the hospital as grouping factor with a random effect on it:

$$Y_i \stackrel{ind}{\sim} \text{Be}(p_i), \quad i = 1, \dots, 240, \quad (1)$$

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + b_{k[i]}, \quad (2)$$

where $b_{k[i]}$ represents the hospital effect of the i -the patient in hospital $k[i]$. We will denote by β the vector of regression parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$. It is well-known that (1)-(2) has a latent variable representation, which can be very useful in computing

the Bayesian inferences, as well as in providing medical significance: conditioning on the latent variables $Z_1, \dots, Z_n, Y_1, \dots, Y_n$ are independent, and, for $i = 1, \dots, n$,

$$Y_i = \begin{cases} 1 & \text{if } Z_i \geq 0 \\ 0 & \text{if } Z_i \leq 0 \end{cases}, \quad (3)$$

$$Z_i = \mathbf{x}_i^T \boldsymbol{\beta} + b_{k[i]} + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} f_\varepsilon, \quad (4)$$

being $f_\varepsilon(t) = e^{-t}(1 + e^{-t})^{-2}$ the standard logistic density function.

Concerning prior elicitation, we make the random-effects assumption for the hospitals the patients were admitted to, that is the hospital effect parameters b_j 's are drawn from a common distribution; moreover, we assume symmetry among the hospital parameters, *i.e.* $b_1, \dots, b_J$ can be considered as (the first part of an infinite sequence of) exchangeable random variables. Via Bayesian hierarchical models, not only we will model dependence among the random-effects parameters $\mathbf{b}$, but it will also be possible to use the dataset to make inferences on hospitals which do not have patients in the study. As usual in the hierarchical Bayesian approach, the regression parameter $\boldsymbol{\beta}$ and the hospital parameter vector $\mathbf{b} := (b_1, \dots, b_J)$ are assumed a priori independent, with $\boldsymbol{\beta} \sim MN(\mu_\beta, V_\beta), b_1, \dots, b_J | \boldsymbol{\sigma} \stackrel{iid}{\sim} \mathcal{N}(\mu_b, \sigma^2), \boldsymbol{\sigma} \sim U(0, \sigma_0)$. Here the uniform prior on $\boldsymbol{\sigma}$ is set as an assumption of ignorance/symmetry of the standard deviation of each hospital effect. We fixed $\mu_b = 0$ regardless of any information, since, by the exchangeability assumption, the different hospital have the same prior mean (fixed equal to 0 to avoid confounding with β_0). The Gaussian prior for $\boldsymbol{\beta}$ is standard, but its hyperparameters, as well as the hyperparameter of the prior distribution for $\boldsymbol{\sigma}$, will be given informatively, using available information from other MOMI² collections. We have in fact enough past data to be relatively informative through prior hyperparameters: they were fixed after having fitted model (1)-(2), with non-informative priors for model parameters ($\boldsymbol{\beta}, \mathbf{b}$ and $\boldsymbol{\sigma}$), to "similar" data (*i.e.* the remaining four MOMI² collections MOMI².1, MOMI².2, MOMI².5, MOMI².6). Therefore we fixed $\mu_\beta = (3, 0, 0.1, -0.7)^T$, which are the posterior means of the regression parameters under the preliminary analysis. The matrix V_β was assumed diagonal, $V_\beta = \text{diag}(2, 0.04, 0.5882, 3.3333)$, which, except for the second value, are about 10 times the posterior variances of the regression parameters under the preliminary analysis (0.04 is 100 times the posterior variance, in order to consider a vaguer prior for β_1). The prior hyperparameter σ_0 was fixed equal to 10, a value compatible with the support of the posterior distribution for $\boldsymbol{\sigma}$ in the preliminary analysis. Posterior estimates of $\boldsymbol{\beta}, \mathbf{b}$ and $\boldsymbol{\sigma}$ proved to be robust with respect to μ_β and V (we fixed a non-diagonal matrix, assuming prior dependence through the regression parameters via the preliminary analysis).

4 Data Analysis and Results

In this section we illustrate the Bayesian analysis of the dataset described in Section 2, but first we give some details on computations. All estimates were computed via a Gibbs sampler algorithm; the model was implemented in WinBUGS (see

<http://www.mrc-bsu.cam.ac.uk/bugs/> and [13]). The first 100,000 iterations were discarded, retaining parameter values each 80 iterations to decrease autocorrelations, with a final sample size equal to 5,000; we run the chains much longer (for a final sample size of 10,000 iterations), but the gain in the MC errors was relatively small. Some convergence diagnostics (Geweke and Heidelberger-Welch) were checked, together with traceplots, autocorrelations and MC error/posterior standard deviation ratios for all the parameters, indicating the MCMC algorithm converged. Code is available from the authors upon request.

Summary inferences about regression parameters and σ can be found in Table 1. It is clear that the marginal posterior of β_1 and β_3 are concentrated on the negative numbers, confirming the naïve interpretation that an increase in age, OB or Killip class decreases the survival probability. The negative effect of the $\log(\text{OB})$ is weaker, even if the posterior median of β_2 is -0.16 ; anyway it is necessary to be included, as suggested by clinical knowhow. Moreover the posterior mean of $\beta_0 + b_j$ is between 2.90 and 4.78, yielding to a high posterior estimates of the survival probability from any hospital, as we could expected since the unbalanced composition of the dataset.

intercept	β_0	3.8160	0.5704
age	β_1	-0.0792	0.0324
$\log(\text{OB})$	β_2	-0.1527	0.3326
killip	β_3	-1.5090	0.8159
rand. eff.	σ_b	1.1770	0.7417

Table 1 Posterior means and standard deviations of the regression parameters and σ .

In Table 2 are reported

$$\tilde{p}_j = \min\{\mathbb{P}(b_j > 0|\mathbf{y}), \mathbb{P}(b_j < 0|\mathbf{y})\}, \quad j = 1, \dots, J, \quad (5)$$

together with the signum of the posterior median of the b_j 's. Low values of $\tilde{p}_j$ denote the posterior distribution of b_j is far from 0, so that the j -th hospital significantly contributes to the (estimated) regression intercept $\beta_0 + b_j$. Credibility Intervals corresponding to $\tilde{p}_j$ can be constructed, and highlight that hospital 9 has a positive effect, while hospital 10, 11 and 15 have a negative effect on the survival probability.

b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}	b_{16}	b_{17}
0.27	0.40	0.32	0.25	0.44	0.41	0.49	0.49	0.18	0.17	0.12	0.28	0.28	0.44	0.17	0.26	0.29
+	+	+	+	-	+	+	+	+	-	-	+	-	+	-	-	+

Table 2 Values of $\tilde{p}_j$ and the signum of the posterior median of each hospital parameters.

We are interested in predictions, too. This implies (i) considering the posterior predictive survival probability of a new patient coming from an hospital already included in the study, or (ii) the posterior predictive survival probability of a new patient coming from a new $(J + 1)$ th hospital. We have, for $j = 1, \dots, J$,

$$\mathbb{P}(Y_{n+1} = 1|\mathbf{y}, \mathbf{x}, b_j) = \int_{\mathbb{R}^5} \mathbb{P}(Y_{n+1} = 1|\boldsymbol{\beta}, b_j, \mathbf{x}) \pi(\boldsymbol{\beta}, b_j|\mathbf{y}) d\boldsymbol{\beta} db_j, \quad (6)$$

for a new patient with covariate vector $\mathbf{x}$ coming from the j -th hospital, and

$$\mathbb{P}(Y_{n+1} = 1 | \mathbf{y}, \mathbf{x}) = \int_{\mathbb{R}^5} \mathbb{P}(Y_{n+1} = 1 | \beta, b_{J+1}, \mathbf{x}) \pi(\beta, b_{J+1} | \mathbf{y}) d\beta db_{J+1}, \quad (7)$$

$$\pi(\beta, b_{J+1} | \mathbf{y}) = \int_{\mathbb{R}^+} \pi(b_{J+1} | \sigma) \pi(\beta, \sigma | \mathbf{y}) d\sigma, \quad (8)$$

where $\pi(b_{J+1} | \sigma)$ is the prior *population* conditional distribution. Figure 1 displays medians and 95% credibility intervals for the posterior predictive survival probabilities (7) of four *benchmark* patients coming from an hospital already in the study: (a) mean age (64 years), mean OB (303 min.) and less severe infarction; (b) age and OB as patient (a), but with severe infarction; (c) elder age (80 years), mean OB and less severe infarction; (d) age and OB as patient (c), but with severe infarction. The last credibility interval (in red) corresponds to the posterior predictive survival probability 7 of a benchmark patient coming from a *new random* ($J+1$)-th hospital. We can see that Killip influence survival more strongly even than age does: in fact, from left to right panels (same age, killip increased) the CIs get longer and longer, more than they does observing panels up to down (same killip, age increased).

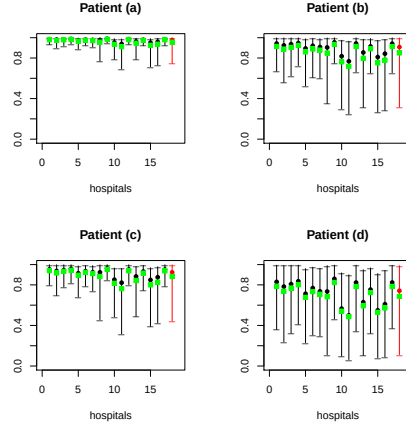


Fig. 1 Posterior median and 95% credibility intervals of the posterior predictive survival probabilities for 4 *benchmark* patients from each hospital in the study and from a *new random* hospital (the 18-th red CI). The green square represents the posterior mean.

Finally we computed the *latent* Bayesian residuals for binary data as suggested in [1]. Thanks to the latent variable representation (3)-(4) of the model, we can consider the *realized* errors

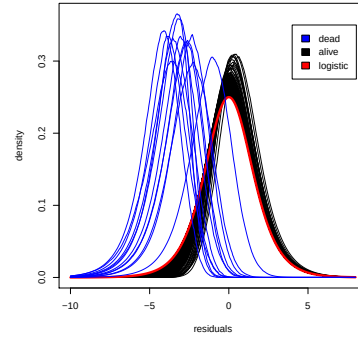
$$e_i = Z_i - (\mathbf{x}_i' \beta + b_{k[i]}) \quad i = 1, \dots, n,$$

obtained solving (4) w.r.t. ε_i . Then the posterior distribution of e_i is obtained following the procedures proposed in [1]. We have

$$\pi(e_i|\mathbf{y}, \boldsymbol{\beta}) = \begin{cases} \frac{\phi(e_i)}{\Phi(x_i^T \boldsymbol{\beta})} \mathbb{I}(e_i > -x_i^T \boldsymbol{\beta}) & \text{if } y_i = 0 \\ \frac{\phi(e_i)}{-\Phi(x_i^T \boldsymbol{\beta})} \mathbb{I}(e_i < -x_i^T \boldsymbol{\beta}) & \text{if } y_i = 1 \end{cases},$$

and for each patient the posterior distribution is compared with the logistic prior for indications of possible departures from the assumed model and the presence of outliers (see also [4]). Therefore, it is sensible to plot credibility intervals for the marginal posterior of each e_i , comparing them to the marginal prior credibility intervals (of the same level). Figure 2 display the posterior distributions of the Bayesian residuals for each observations, where the red line in the plot denotes the prior marginal distribution (logistic).

Fig. 2 Posterior distributions of the latent Bayesian residuals against the fitted probabilities. The *blue* and *black* lines correspond to observations $y_i = 0$ and $y_i = 1$, respectively. The *red* line is the marginal prior distribution (logistic).



From the picture, it seems that there are no outlier among the patients who survived, while we see variability among the dead patients as far as posterior location and dispersion is concerned.

5 Conclusions

In the present work we suggested how hierarchical models can be employed in the analysis of clinical data with a multilevel structure. These data arise from clinical registers integrated with administrative databanks, and analyses carried out on them could provide a decisional support to people concerned with cardiovascular health-care governance. We set in a Bayesian framework the problem of modelling survival outcome by means of relevant covariates, taking into account overdispersion induced by the grouping factor, i.e. the hospital where each patient has been admitted to. The model has been employed to study the effects of variations in health-care use on patients outcomes, since it points out the relationship between measure of the process and measure of outcome. We also provided cluster-specific estimates of survival, adjusted for patients characteristic, and derived estimates of covariates effects, using Markov Chain Monte Carlo simulation of posterior distribution of parameters and discussing model goodness of fit through latent residuals method proposed by Albert and Chib [1].

Finally, as further step of analysis, we are actually considering the application of Dirichlet Process Mixture of Normal to the random effect distribution, in order to find out a hospital classification based on Bayesian non parametric techniques (for further details see [8]).

Acknowledgements This work is a part of the Strategic Program “Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction” supported by “Ministero del Lavoro, della Salute e delle Politiche Sociali” and by “Direzione Generale Sanità - Regione Lombardia”. The authors wish to thank the Working Group for Cardiac Emergency in Milano, the Cardiology Society, and the 118 Dispatch Center.

References

1. Albert, J. and Chib, S.: Residual Analysis for Binary Response Regression Models. *Biometrika* **82**, 4: 747–759 (1995), <http://www.jstor.org/stable/2337342>
2. Antman, E.M., Hand, M., Armstrong, P.W., Bates, E.R., Green, L.A. et al.: Update of the ACC/AHA 2004 Guidelines for the Management of Patients with ST Elevation Myocardial Infarction. *Circulation*, **117**: 269–329 (2008)
3. C.P.Cannon, C.M.Gibson, C.T.Lambrew, D.A.Shultz, D.Levy, W.J.French, Gore J.M., Weaver W.D., Rogers W.J., Tiefenbrunn A.J.: Relationship of Symptom-Onset-to-Balloon Time and Door-to-Balloon Time with Mortality in Patients undergoing Angioplasty for Acute Myocardial Infarction. *Journal of American Medical Association*, Giugno, **283**, 22: 2941-2947 (2000)
4. Chaloner, K. and Brant, R.: A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **31**: 651–659 (1988).
5. Daniels, M.J. and Gatsonis, C.: Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization. *JASA* **94**, 445: 29-42 (1999), <http://www.jstor.org/stable/2669675>
6. Duncan, C., Kelvin, J. and Moon, G.: Context, Composition and Heterogeneity: using Multi-level Models in Health Research. *Soc. Sci. Med.* **46**, 1: 97-117 (1998)
7. Gelman, A. and Hill, J.: Data analysis using regression and multilevel/hierarchical Models. Cambridge (2007)
8. Guglielmi A., Ieva F., Paganoni A.M., Ruggeri F.: A Bayesian random-effects model for survival probabilities after acute myocardial infarction. *Work in Progress* (2010)
9. Ieva F.: Modelli statistici per lo studio dei tempi di intervento nell'infarto miocardico acuto. Master Thesis (2008), Dipartimento di Matematica, Politecnico di Milano. [Online] <http://mox.polimi.it/it/progetti/pubblicazioni/tesi/ieva.pdf>
10. Ieva, F. and Paganoni, A.M.: A case study on treatment times in patients with ST-Segment Elevation Myocardial Infarction. In: MOX Report n. 05/2009, Dipartimento di Matematica, Politecnico di Milano (2008). [Online] <http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/05-2009.pdf>
11. Ieva, F. and Paganoni, A.M.: Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI2 survey. Submitted. (2010)
12. Jneid H., Fonarow G., Cannon C., Palacios I., Kilic T. et al.: Impact of time of presentation on the care and outcomes of acute myocardial infarction. *Circulation*, **117**: 2502-2509 (2008)
13. Ntzoufras, I.: Bayesian modeling using WinBUGS. Wiley (2009)
14. Rathore S.S., Curtis J.P. and Chen J.: Association of door to balloon time and mortality in patients admitted to hospital with ST-elevation myocardial infarction: national cohort study. *British Medical Journal*, **338** (2009)
15. Saia F., Marzocchi A., Manari G., Guastaroba P., Vignali L., Varani E. et al.: Patient selection to enhance the long-term benefit of first generation drug-eluting stents for coronary revascularization procedures: insights from a large multicenter registry. *Eurointervention*, **5**, 1: 5766 (2009).