

Data mining the Lombardia Public Health Database: a pilot case study on hospital discharge data for Acute Myocardial Infarctions

Francesca Ieva, Anna Maria Paganoni and Piercesare Secchi

Abstract Clinical registers and the Public Health Database are the main data banks of the Lombardia Region; they house information about health-care processes, epidemiological incidence and prevalence of diseases, therapeutic procedures and clinical outcomes. We present a pilot data mining analysis, performed with generalized mixed effects models, on the subset of the Public Health Database concerning hospital discharge data relative to Acute Myocardial Infarctions.

Key words: Health service research, Biostatistics and bioinformatics, Generalized Linear Mixed Models, Data Mining.

1 Introduction

Recent years have witnessed 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 epidemiological incidence and prevalence of diseases. In response, a sizeable literature has emerged questioning the right use of such indicators as a measure of quality of care, as well as stating more specific criticism of the statistical methods used to obtain estimates adjusted for patient case-mix.

Health-care service scheduling is strictly connected with a deep knowledge of current health needs; with respect to this, a total novelty is represented by the potential offered by the statistical data mining of administrative data-banks for collecting clinical and epidemiological information. Indeed, there is an increasing agreement among epidemiologists on the validity of disease and intervention registries based on administrative databases [2, 6, 14]. In this work we will illustrate a pilot data mining case study on hospital discharges data for patients with a diagnosis of Acute My-

Francesca Ieva, Anna Maria Paganoni, Piercesare Secchi
MOX - Dipartimento di Matematica, Politecnico di Milano, via Bonardi 9, 20133 Milano (Italy)
e-mail: francesca.leva@mail.polimi.it, anna.pagannoni@polimi.it, piercesare.secchi@polimi.it

ocardial Infarction; data come from the Lombardia Region Public Health Database (PHD), an on going collection of data used, up to now, only for administrative purposes.

The study is part of the Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction" (AMI Project) funded by the Ministry of Health and by "Direzione Generale Sanità - Regione Lombardia" and started in January 2009. The major objective of this project is the identification of new diagnostic, therapeutic and organizational strategies to be applied to patients with Acute Coronary Syndromes (ACS), in order to improve clinical outcomes. To achieve this goal Regione Lombardia authorized the extraction from the PHD database of data concerning patients with Acute Coronary Syndromes.

In the next section will describe the case study while in the third section of the paper we will put it in the context of the AMI Project.

All the analyses have been performed with the R program (version 2.10.1) [17]; for nonlinear mixed effects models, the `nlme` package [16] has been used.

2 Data mining discharge data on Acute Myocardial Infarctions

In this section we describe a data mining exercise on the Hospital Discharge Database (*Database Ricoveri*), which is one of the three main databases belonging to the Star scheme [12] that composes the PHD of Lombardia Region. We focus on the numbers of hospital discharges with a diagnosis of Acute Myocardial Infarction, grouped by hospital, during years 2000 – 2007. Cases detection is performed according to the AHQR guidelines [3]. We explore the question as to whether there has been a trend in the growth of the number of Acute Myocardial Infarction (AMI) diagnoses in the time interval 2000 – 2007.

Figure 1 represents, along the time period 2000 – 2007, the number of ST-elevation Acute Myocardial Infarction (STEMI) diagnoses for the largest 30 hospitals in the database as well as the number of Acute Myocardial Infarction without ST-elevation (NON-STEMI) diagnoses in the same time period and for the same institutions.

The high variability between hospitals and the structure of the data grouped by hospital, motivate the use of mixed effects models [15] for the analysis of these longitudinal data. A first explorative analysis conducted by means of a simple linear mixed model, where the number of diagnoses appears as a linear function of time, with hospital as a grouping factor, shows that there is no statistical evidence for a linear trend over time on the growth of the number of STEMI diagnoses (the p-value of the test on the "year" fixed linear effect is equal to 0.4931). However, the same model fitted on the number of NON-STEMI diagnoses shows a significant linear trend over time (the p-value of the test on the "year" fixed effect is equal to 3.11×10^{-14}).

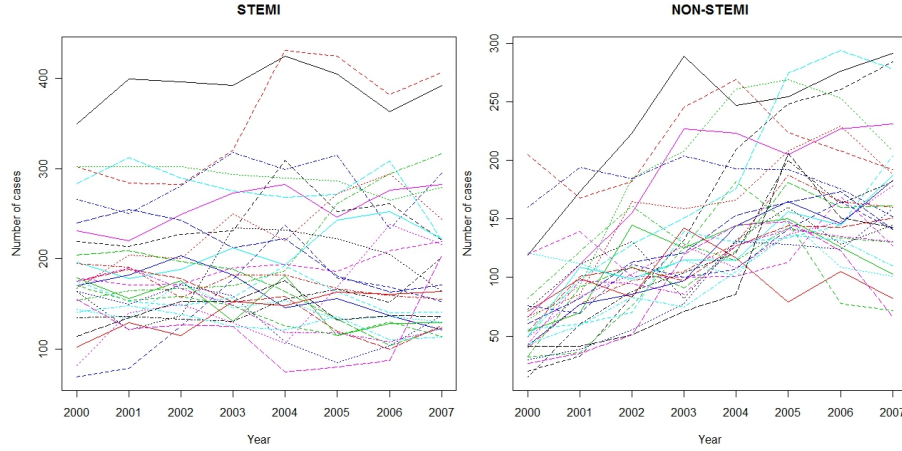


Fig. 1 Left panel: Number of AMI with ST-elevation diagnoses in period 2000 - 2007 in 30 clinical institutions of Lombardia Region. Right panel: Number of AMI without ST-elevation diagnoses in period 2000 - 2007 in 30 clinical institutions of Lombardia Region

In order to elicit benchmark hospital growth curves for the number of NON-STEMI diagnoses, we extend the analysis of the NON-STEMI data by modeling the growth of the number of diagnoses over time by means of a mixed effects logistic model. Indeed we set N_{ij} to be the number of NON-STEMI diagnoses for hospital $i = 1, \dots, 30$ and year $j = 1, \dots, 8$, where $j = 1$ is for year 2000 and $j = 8$ is for year 2007, and we fit the following mixed effects logistic model with respect to time

$$N_{ij} = \frac{\text{Asym} + \alpha_i}{(1 + \exp(\text{Tmid} + \tau_i - t_j))} + \varepsilon_{ij}, \quad i = 1, \dots, 30, \quad j = 1, \dots, 8, \quad (1)$$

where t_j is the centered time covariate (i.e. $t_0 = 2000 - 2003.5 = -3.5$, $t_1 = 2001 - 2003.5 = -2.5$ and so on), the fixed effects Asym and Tmid represent respectively the asymptote and the inflection point of the logistic curve, while α_i and τ_i are i.i.d samples of the random variables $\alpha \sim \mathcal{N}(0, \sigma_\alpha)$ and $\tau \sim \mathcal{N}(0, \sigma_\tau)$, respectively, representing gaussian additive random effects, grouped by hospital. The term $\varepsilon \sim \mathcal{N}(0, \sigma)$ is the residual error. The two random effects α and τ are assumed to be independent, and independent of ε , and the corresponding variances are estimated by restricted maximum likelihood maximization algorithms. Table 1 shows that both fixed effects Asym and Tmid are significant.

Aiming at benchmark hospital growth curves, we inspect Figure 2, right panel, where each of the 30 different institutions is represented in terms of the asymptote and inflection point of the respective fitted logistic growth curve. The variability of the asymptote estimates is a direct consequence of the variability among hospitals with respect to size. The variability of the estimated inflection points stimulates more interesting interpretations, since, in the logistic model, the inflection point

Fixed effects estimates:

	Value	Std. Error
Asym	165.8232	8.954670
Tmid	-2.7050	0.229744

Anova Table:

	numDF	denDF	F-value	p-value
Asym	1	209	366.9734	< .0001
Tmid	1	209	138.6263	< .0001

Table 1 Fixed effects estimates and Anova table for model (1)

indicates the time of maximum growth speed and this, in turn, is directly related to the timing of a growth speed significantly different from zero.

In order to catch some clustering structure, for $k = 1, 2, \dots$, we partition the set of estimated inflection points in k clusters by means of the Partitioning Around Medoids procedure (PAM, [13]), implemented with the euclidean distance. By inspecting the boxplots of the distances between the estimated inflection points and the medoid of the cluster they belong to, obtained for different values of the number k of clusters (Figure 2, left panel), the presence of $k = 3$ clusters can be sustained. Indeed, for $k = 3$, the Average Silhouette Width is equal to 0.57 [19] and the three representative inflection points (medoids) correspond to years 1999, 2001, 2002. In Figure 2, right panel, the point representing each institution has been colored according to the cluster it belongs to. "Cluster A" denotes the institutions colored in red; they are closer to the medoid with earliest inflection point. "Cluster B" denotes the institutions colored in black; they are closer to the medoid with medium inflection point. Finally, "Cluster C" denotes the institutions colored in green; the medoid for this cluster has the latest inflection point.

The clustering structure of the 30 inflection point estimates motivates a further extension of the analysis of the NON-STEMI data. Indeed we improve on model (1) by admitting a different inflection point fixed effect for the three different clusters A, B, or C. The extended model now becomes:

$$N_{ij} = \frac{\text{Asym} + \alpha_i}{(1 + \exp(\text{Tmid}_A \cdot 1_{i \in A} + \text{Tmid}_B \cdot 1_{i \in B} + \text{Tmid}_C \cdot 1_{i \in C} + \tau_i - t_j))} + \varepsilon_{ij}, \quad (2)$$

where $i = 1, \dots, 30$, is the institution index, $j = 1, \dots, 8$, is the year index, and α, τ and ε are defined as before.

The extended model (2) improves on model (1) both in terms of AIC and likelihood ratio test:

	Model	AIC	Test	L.Ratio	p-value
Model with grouping	1	2305.6			
Model without grouping	2	2363.6	1 vs 2	62.001	<.0001

Estimates for the fixed effects of model (2) appear in Table 2. Testing all possible contrasts between the three different fixed effects related to the inflection point,

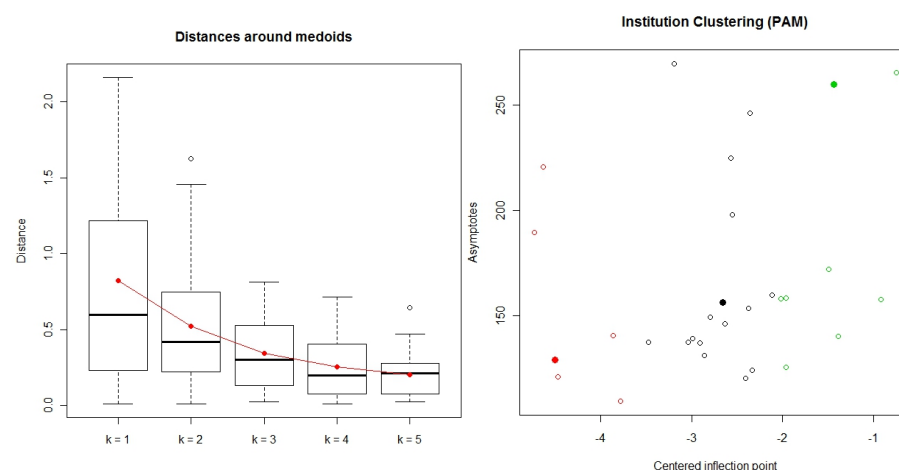


Fig. 2 Left panel: Flanked boxplots of the distances between estimated inflection points and medoids of the corresponding cluster (detected with PAM procedure), for different number k of clusters, and means of distances distribution (red line). Right panel: Parameters of institutions clustered by a Partitioning Around Medoids (PAM) procedure on the estimated inflection points with 3 clusters

Fixed effects estimates:		
	Value	Std. Error
Asym	165.55874	8.898537
Tmid _A	-5.33993	0.561800
Tmid _B	-2.72262	0.106235
Tmid _C	-1.20932	0.146632

Table 2 Fixed effects estimates for model (2)

always generates a p-value less than 10^{-5} ; there is a strong evidence of different inflection points in the three groups.

In the left panel of figure 3, the curves estimated by model (2), one curve for each hospital, are represented together with the real data; the right panel shows the estimated logistic growth curves. The thick red, black and green curves represent the three benchmarks growth curves, i.e. the medoids of the cluster A, B and C, respectively.

The clinical interpretation of this statistical analysis is related to the introduction of the *troponin test* as a procedure to diagnose NON-STEMI Acute Myocardial Infarctions. In fact, certain subtypes of troponin, which are regulatory proteins, are very sensitive and specific indicators of damage to the heart muscle in the case of NON-STEMI Acute Myocardial Infarction, and only in this case. The variability in estimated inflection points among clinical institutions could be due to different timings of introduction and adoption of the troponin test by the different hospitals. Indeed the statistical analysis conducted in this section supports the clinical tenet that in the time period 2000 – 2007 there has been no increase in the number of

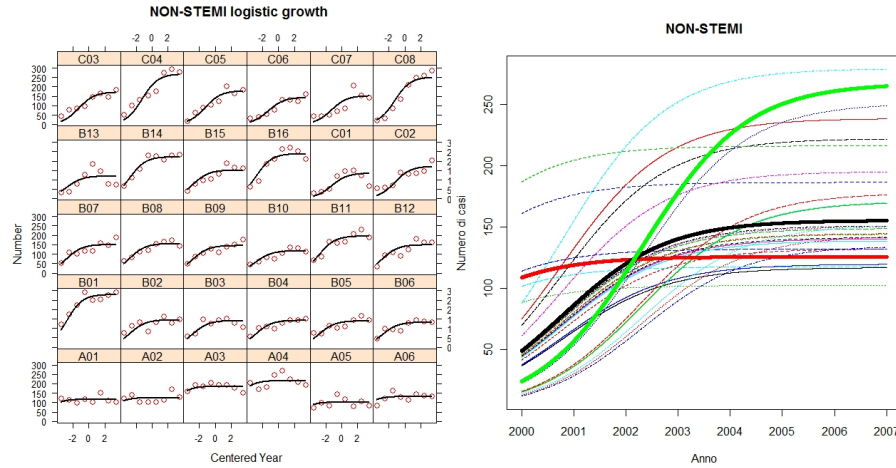


Fig. 3 Estimated logistic growth curves for different medical institutions

STEMI Acute Myocardial Infarctions, while the apparent increase in the number of NON-STEMI diagnoses is not due to a real increase in the disease incidence, but to a new diagnostic procedure adopted in hospitals along different timings.

3 The AMI Project

The case study illustrated in the previous section is an example of the potential offered by the statistical analysis of an administrative database for clinical and epidemiological purposes. The unassailable benefit of the use of the PHD is the high data quality, and the real time data availability without costs increase. This innovative perspective was a paramount motivation for the Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction" (AMI Project). The major objective of the AMI Project is the identification of new diagnostic, therapeutic and organizational strategies to be applied to patients with Acute Coronary Syndromes (ACS), in order to improve clinical outcomes.

The AMI Project was also motivated by a previous experience of the Milan network for Cardiac Emergency, showing how a networking strategy that coordinates territory, rescue units and hospitals in a complex urban area with high technological and medical resources, improves health care of patients with ST-segment Elevation Myocardial Infarction (STEMI) and provides the opportunity to collect and analyze data in order to optimize resources. In fact, in 2006 – 2009 a pioneering pilot study, called MOMI² (MOnth MOnitoring Myocardial Infarction in Milan), has been conducted by The Working Group for Cardiac Emergency in Milano, the Cardiology Society, and the 118 Dispatch Center (national free number for medical emergen-

cies) in the urban area of Milano to streamline an optimal care process for patients with STEMI. Pre-hospital and in-hospital times have been highlighted as fundamental factors we can act upon to reduce in-hospital mortality and to increase the rate of effective reperfusion treatments of infarcted related arteries. (Details about design, study and statistical analysis of the MOMI² clinical register can be found in [7, 8, 9, 11].)

The results of these pilot studies indicated that a structured and efficient network of transport (118) and hospitals makes the difference in reaching best clinical results. This has driven Lombardia Region to design a wider plan, in order to construct an archive concerning patients with Acute Coronary Syndromes and involving all the Cardiology Divisions of Hospitals in Lombardia. This will be a standardized and compulsory procedure for all hospitals in Lombardia [5], in order to improve the therapeutic path of AMI patients. Indeed, the innovative idea of the AMI Project is to devise methods for the joint analysis of data coming from clinical registers, like MOMI², and information obtainable from the PHD, in order to guarantee the same procedures to such an extended and intensive-care area, and also to integrate data collected during observational studies with data contained in the PHD. The ultimate goal is to generate for every patient an integrated longitudinal data vector containing both clinical histories and follows up. More details about the AMI Project and the planned analyses can be found in [4, 10]

4 Conclusions and further developments

The study presented in this article is a pilot example of an advanced statistical analysis performed on data drawn from a Public Health Database; we plan to extend this seminal work with the inclusion of further information concerning related pathologies and repeated procedures for AMI patients. As suggested by an anonymous referee, more compelling explorative analyses of the same data could be obtained through the use of non parametric random effects models, as proposed in [1], as well as by means of Generalized Additive Models for Location, Scale and Shape, as proposed in [18]. We plan to test these models in our future work.

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. Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55** 117128 (1999)
2. Balzi, D., Barchielli, A., Battistella, G. et al.: Stima della prevalenza della cardiopatia ischemica basata su dati sanitari correnti mediante un algoritmo comune in differenti aree italiane. *Epidemiologia e Prevenzione* **32**(3) 22-29 (2008)
3. Glance, L.G., Osler, T.M., Mukamel, D.B. et al.: Impact of the present-on-admission indicator on hospital quality measurement experience with the Agency for Healthcare Research and Quality (AHRQ) Inpatient Quality Indicators. *Medical Care* **46** (2), 112-119 (2008)
4. Barbieri, P., Grieco, N., Ieva, F., Paganoni, A.M., Secchi, P.: Exploitation, integration and statistical analysis of Public Health Database and STEMI archive in Lombardia Region. *MOX-Report*, n. 02/2010, Dipartimento di Matematica, Politecnico di Milano (2010) Available: <http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/02-2010.pdf>
5. Determinazioni in merito alla "Rete per il trattamento dei pazienti con Infarto Miocardico con tratto ST elevato(STEMI)": Decreto N° 10446, 15/10/2009, Direzione Generale Sanità - Regione Lombardia (2009)
6. Every, N.R., Frederick, P.D., Robinson, M. et al.: A Comparison of the National Registry of Myocardial Infarction With the Cooperative Cardiovascular Project. *Journal of the American College of Cardiology* **33** (7), 1887-1894 (1999)
7. Grieco, N., Corrada, E., Sesana, G., Fontana, G., Lombardi, F., Ieva, F., Paganoni, A.M., Marzegalli, M.: Le reti dell'emergenza in cardiologia : l'esperienza lombarda. *Giornale Italiano di Cardiologia Supplemento "Crema Cardiologia 2008. Nuove Prospettive in Cardiologia"*, **9**, 56 - 62 (2008)
8. Ieva, F.: Modelli statistici per lo studio dei tempi di intervento nell'infarto miocardico acuto. *Master Thesis*, Dipartimento di Matematica, Politecnico di Milano (2008) Available: <http://mox.polimi.it/it/progetti/pubblicazioni/tesi/ieva.pdf>
9. Ieva, F., Paganoni, A.M.: A case study on treatment times in patients with ST-Segment Elevation Myocardial Infarction. *MOX-Report*, n. 05/2009, Dipartimento di Matematica, Politecnico di Milano (2009) Available: <http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/05-2009.pdf>
10. Ieva, F., Paganoni, A.M.: Statistical Analysis of an integrated Database concerning patients with Acute Coronary Syndromes, S.Co.2009 - Sixth conference - Proceedings, MAGGIOLI, Milano, 223 - 228 (2009)
11. Ieva, F., 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. Inmon, W.H.: Building the Data Warehouse. John Wiley & Sons, second edition (1996)
13. Kaufman, L., Rousseeuw, P.: Finding Groups in Data. Wiley Series in Probability and Mathematical Statistics (1990)
14. Manuel, D.G., Lim, J.J.Y., Tanuseputro, P. et al.: How many people have a myocardial infarction? Prevalence estimated using historical hospital data. *BMC Public Health* **7**, 174-89 (2007)
15. Pinheiro, J.C., Bates, D.M.: Mixed-Effects Models in S and S-Plus, Springer (2000)
16. Pinheiro, J.C., Bates, D.M., DebRoy, S., Sarkar, D. and the R core team: nlme: Linear and Nonlinear Mixed Effects Models (2009)
17. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (2009) Available: <http://www.R-project.org>
18. Rigby, R. A., Stasinopoulos D. M.: Generalized Additive Models for Location, Scale and Shape, (with discussion). *Applied Statistics*. **54** 507-554 (2005)
19. Struyf, A., Hubert, M., Rousseeuw, P.J.: Clustering in an Object-Oriented Environment. *Journal of Statistical Software*, **1** (1996)