Mining Administrative Health Databases for Epidemiological Purposes: A Case Study on Acute Myocardial Infarctions Diagnoses

38

Francesca Ieva, Anna Maria Paganoni, and Piercesare Secchi

Abstract

We present a pilot data mining analysis on the subset of the Public Health Database (PHD) of Lombardia Region concerning hospital discharge data relative to Acute Myocardial Infarctions without ST segment elevation (NON-STEMI). The analysis is carried out using nonlinear semi-parametric and parametric mixed effects models, in order to detect different patterns of growth in the number of NON-STEMI diagnoses within the 30 largest clinical structures of Lombardia Region, along the time period 2000–2007. The analysis is a seminal example of statistical support to decision makers in clinical context, aimed at monitoring the diffusion of new procedures and the effects of health policy interventions.

Keywords

Biostatistics and bioinformatics • Data mining • Generalized linear mixed models • Health service research

38.1 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.

F. Ieva (✉) · A.M. Paganoni · P. Secchi
MOX - Dipartimento di Matematica, Politecnico di Milano, via Bonardi 9, 20133 Milano, Italy
e-mail: francesca.ieva@mail.polimi.it; anna.paganoni@polimi.it; piercesare.secchi@polimi.it

N. Torelli et al. (eds.), *Advances in Theoretical and Applied Statistics*,
Studies in Theoretical and Applied Statistics, DOI 10.1007/978-3-642-35588-2_38,

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 [1, 4, 12]. In this work we focus on growth curves for the number of diagnoses of Acute Myocardial Infarction without ST segment elevation (NON-STEMI), and we explore the question as to whether they had a trend in the time interval 2000–2007. Indeed, clinical best practice maintains that there is no evidence for a greater incidence of NON-STEMI in this period; however, since the early 2000s a new diagnostic procedure—the *troponin* exam—has been introduced and this, by easing NON-STEMI detection, could have produced an increased number of positive diagnoses. Administrative data-banks can be used to check for growth in the number of NON-STEMI diagnoses along with the adoption by clinical institutions of new diagnostic procedures or devices, like the troponin exam.

In this work we will illustrate a pilot data mining case study on hospital discharges data for patients with NON-STEMI diagnosis; data come from the Lombardia Region Public Health Database (PHD), an ongoing 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 Italian 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.

The statistical analysis is conducted along different phases. The visual evidence for growth in the number of NON-STEMI diagnoses is first questioned by fitting a semi-parametric mixed effect model, in order to capture the shape of growth curves and to test the significance of the grouping factor effect. The relevant features emerged with this first analysis are then modeled by means of parametric nonlinear models of decreasing complexity, which are easier to interpret and more suited to inferential purposes.

All the analyses have been performed with the R program [15]; the `mgcv` [18] package and the `nlme` package [14], respectively, for generalized additive mixed models and for nonlinear mixed effects models have been used.

38.2 Data Mining Discharge Data on Acute Myocardial Infarctions

In this section we describe a data mining study of the Hospital Discharge Database (*Database Ricoveri*), which is one of the three main databases belonging to the Star scheme [10] that composes the PHD of Lombardia Region. We focus on the numbers

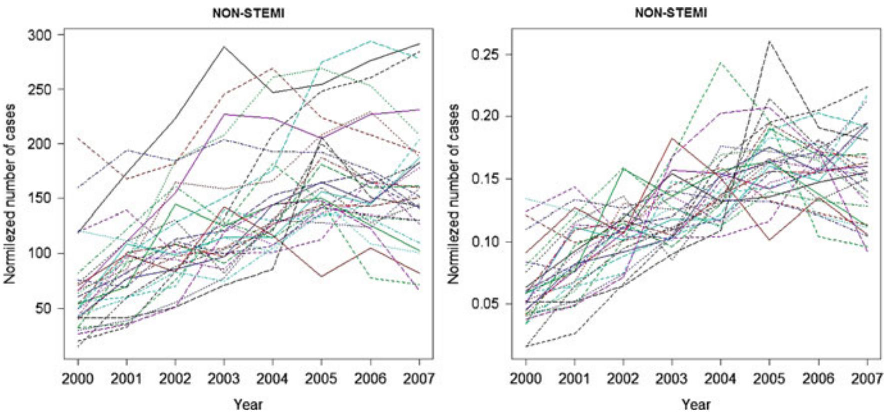

Fig. 38.1 *Left panel*: Number of AMI without ST-elevation diagnoses in the period 2000–2007 in the 30 largest clinical institutions of Lombardia Region. *Right panel*: Standardized number of AMI without ST-elevation diagnoses in the period 2000–2007 in the 30 largest clinical institutions of Lombardia Region. For each hospital the yearly number of diagnoses has been divided by the hospital total number of diagnoses in the time period 2000–2007

of hospital discharges with a diagnosis of NON-STEMI, grouped by hospital and relative to the 30 largest clinical institutions of Lombardia Region, during years 2000–2007. Detection of cases is performed according to the AHQR guidelines [5].

Figure 38.1, left panel, represents the number of acute myocardial infarction without ST-elevation (NON-STEMI) diagnoses, along the time period 2000–2007, for the 30 hospitals. The total number of diagnoses in the time period 2000–2007 has a considerable variability between institutions: in fact it ranges from a minimum value of 715 to a maximum of 1,872. This difference is due to the different exposure of different hospitals; indeed, exposure could be a confounding factor in a statistical analysis focused on the growth trend of the number NON-STEMI cases. Hence, in order to analyze comparable data, for each hospital the yearly number of diagnoses has been standardized by the hospital total number of diagnoses in the time period 2000–2007, thus adjusting for hospital exposure (see Fig. 38.1, right panel).

The high variability between hospitals and the structure of the data grouped by hospital motivate the use of mixed effects models [13] for the analysis of these longitudinal data. A first explorative analysis conducted by means of a simple linear mixed model, where the standardized number of NON-STEMI diagnoses appears as a linear function of time, with hospital as a grouping factor, shows a significant linear trend over time (the p-value of the test on the "year" fixed effect is less than 10^{-14}).

Since the use of a linear parametric model can be quite binding, a further enquire into the growth trend has been conducted by fitting a semi-parametric mixed effect model. Indeed, we set $\tilde{N}_{ij}$ to be the standardized number of NON-STEMI diagnoses for hospital $i = 1, \ldots, 30$ and year $j = 1, \ldots, 8$, where $j = 1$ is for year 2000

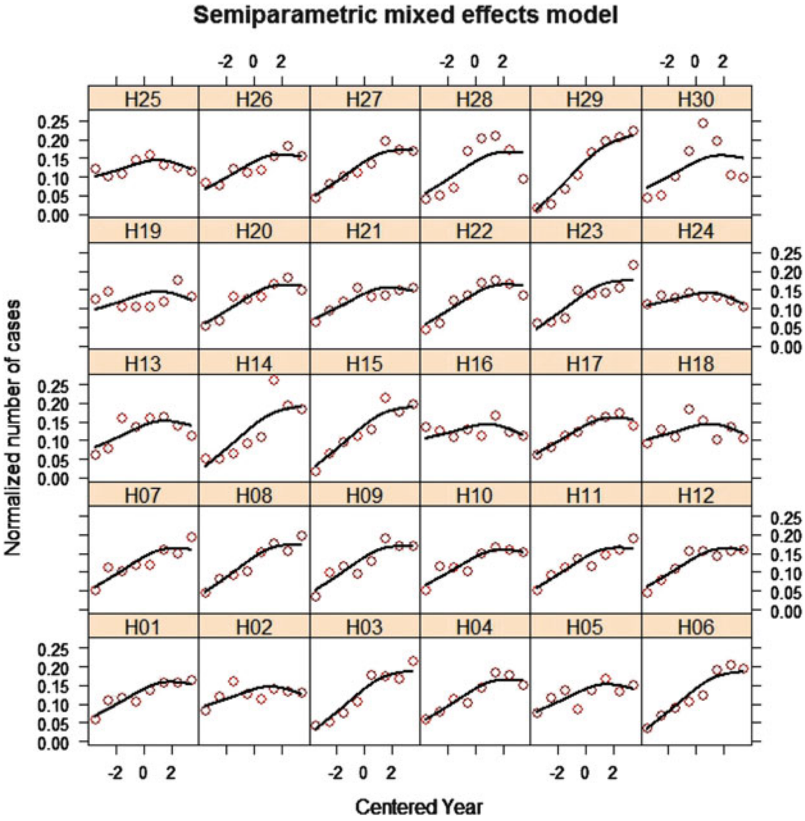

Fig. 38.2 Estimated growth curves through model (38.1) together with the original data

and $j = 8$ is for year 2007, and following [17], we fit the following mixed effects semi-parametric model with respect to time

$$\tilde{N}_{ij} = s(t_j) + b_{0i} + b_{1i} t_j + \varepsilon_{ij} \qquad i = 1, \dots 30, \quad j = 1, \dots, 8, \tag{38.1}$$

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), s is a common cubic regression spline, while b_{0i} and b_{1i} are i.i.d samples of the random variables $b_0 \sim \mathcal{N}(0, \sigma^2_{b_0})$ and $b_1 \sim \mathcal{N}(0, \sigma^2_{b_1})$ respectively, representing gaussian additive independent random effects, grouped by hospital. The quantities ε_{ij} are i.i.d. samples from the random variable $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ representing residual error: ε, b_0 and b_1 are assumed to be independent. Estimates are obtained by maximization of restricted likelihood. Figure 38.2 shows the estimated growth curves together with the original data.

We fitted a semi-parametric mixed effects model in order to catch a common behavior in the growth of normalized number of NON-STEMI diagnoses in the years 2000–2007, smoothing data and taking into account overdispersion due to

Table 38.1 Fixed effects estimates and Anova table for model (38.2)

Fixed effects estimates:

	Value	Std. Error
Asym	0.1544	0.0026
Tmid	−2.7017	0.1368

Anova Table:

	numDF	denDF	F-value	p-value
Asym	1	209	5417.630	$< .0001$
Tmid	1	209	389.845	$< .0001$

the grouping factor. In fact, inspection of Fig. 38.2 suggests a common "S-shaped" growing pattern. Concerning the random effects, the estimated parameters are: $\hat{\sigma}_{b_0} = 2.702 * 10^{-07}$, $\hat{\sigma}_{b_1} = 0.00765$, and $\hat{\sigma} = 0.02297$. The negligible effect of the random variable b_0 suggests that the curves are in fact different only with respect to their growth rate. The greater effect of the random variable b_1 is conducive to a further analysis of these data by means of a model that captures the common growth trend while taking into account overdispersion in the growth rates. Indeed, the following (parametric) logistic mixed effects model accommodates for the "S-shaped" common growing pattern, pointed out by the nonparametric analysis, while enabling the testing of its significance:

$$\tilde{N}_{ij} = \frac{\mathrm{Asym} + \alpha_i}{(1 + \exp(\mathrm{Tmid} + \tau_i - t_j))} + \varepsilon_{ij}, \qquad i = 1, \ldots 30, \quad j = 1, \ldots, 8, \tag{38.2}$$

where t_j is the centered time covariate, 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 $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$ and $\tau \sim \mathcal{N}(0, \sigma_\tau^2)$, respectively, representing gaussian additive random effects, grouped by hospital. The quantities ε_{ij} are i.i.d. samples from the random variable $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and they represent residual error. The two random effects α and τ are assumed to be independent, and independent of ε; all estimates are computed by restricted maximum likelihood. Table 38.1 shows that both fixed effects Asym and Tmid are significant.

Concerning the random effects, the estimated parameters are: $\hat{\sigma}_\alpha = 6.8183 * 10^{-07}$, $\hat{\sigma}_\tau = 0.4821$, and $\hat{\sigma} = 0.0287$. It is then confirmed that the variability of the additive random effect relative to the asymptote is negligible; thus α_i can be removed from model (38.2) without loss in model performance. On the contrary, the variability of the random effect relative to the inflection point is large and implies a very significant effect; this stimulates an interesting interpretation, since, in the logistic model, the inflection point 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.

Inspection of the set of (estimated) random effects $\tau_i, i = 1, \ldots, 30$, related to the inflection point suggests a clustering structure that has been captured by partitioning the set in $k = 1, 2, \ldots,$ clusters by means of the Partitioning Around Medoids procedure (PAM, [11]), implemented with the Euclidean distance, denoted by d. A critical point is the choice of k, the number of groups: an helpful method is the computation of the average silhouette width, and the inspection of the silhouette plot of PAM. For each estimated τ_i, let A be the cluster to which τ_i has been assigned and compute $a(\tau_i)$, the average dissimilarity of τ_i to all other objects in A,

$$a(\tau_i) = \frac{1}{|A| - 1} \sum_{\tau_j \in A,\, \tau_j \neq \tau_i} d(\tau_j, \tau_i).$$

Now, if C is a cluster different from A, denote by

$$d(\tau_i, C) = \frac{1}{|C| - 1} \sum_{\tau_j \in C} d(\tau_j, \tau_i)$$

the average dissimilarity of τ_i from all objects in C and set $c(\tau_i)$ to be the smallest value of $d(\tau_i, C)$ when C is let to range over the set of all clusters different from A. The *silhouette value* $s(\tau_i)$ of τ_i is defined as

$$s(\tau_i) = \frac{c(\tau_i) - a(\tau_i)}{\max\{a(\tau_i), c(\tau_i)\}}.$$

Clearly $s(\tau_i)$ lies between -1 and 1; large values of $s(\tau_i)$ support the fact that the element τ_i is well classified in A. The entire silhouette plot, i.e., the plot of all $s(\tau_i)$, and the Average Silhouette Width, i.e., the average of all silhouette values, are qualitative indexes helpful to judge and compare the results obtained by PAM for different values of k [16].

By inspecting the silhouette plot, represented in Fig. 38.3, the presence of $k = 3$ clusters can be sustained. Indeed, for $k = 3$, the Average Silhouette Width is equal to 0.58 and, as a general rule, it can be asserted that a reasonable clustering structure has been found when the Average Silhouette Width is greater than 0.5. The medoids representative of the three clusters correspond to years $y_A = 2000$, $y_B = 2001$ and $y_C = 2002$. "Cluster A" denotes the institutions for which the estimated time of inflection point Tmid $+ \tau_i$ in model (38.2) is closer to -3.1692, i.e., closer to year $y_A = 2000$. Analogously, "Cluster B" denotes the institutions for which the estimated time of inflection point is closer to -2.6839, i.e., closer to year $y_B = 2001$, and "Cluster C" denotes the institutions for which the estimated time of inflection point is closer to -2.3014, i.e., closer to year $y_C = 2002$.

In the left panel of Fig. 38.4, the curves estimated by model (38.2) are represented, one curve for each hospital, 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., medoids for cluster A, B, and C, respectively.

The particular interest in analyzing the clustering structure of the random effects related to the inflection points derives by the clinical surmise about their

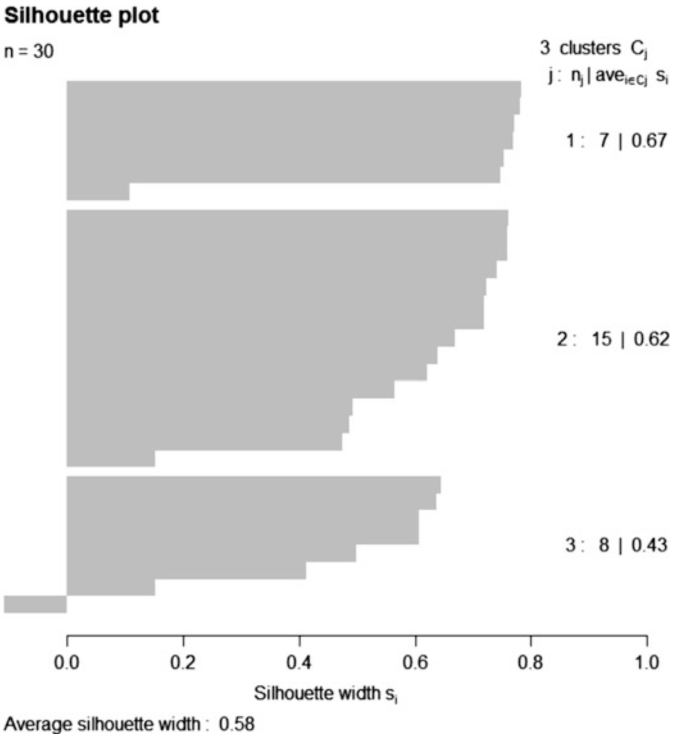

Fig. 38.3 Silhouette plot of PAM procedure on the estimated inflection points with $k = 3$ clusters

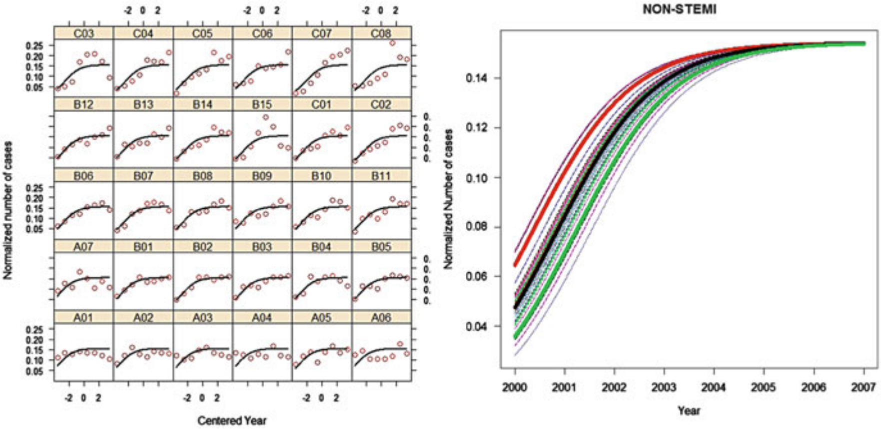

Fig. 38.4 Estimated logistic growth curves for different medical institutions

Table 38.2 Fixed effects estimates for model (38.3)

	Value	Std. Error	p-value
Asym	0.1540	0.0025	$< .0001$
Tmid$_A$	−3.9434	0.2383	$< .0001$
Tmid$_B$	−2.6719	0.1294	$< .0001$
Tmid$_C$	−1.9108	0.1637	$< .0001$

presence. Indeed, it is known that from the early 2000s the troponin exam has been introduced in hospital practices as a diagnostic device to better identify NON-STEMI events; hence, the presence of three clusters for the random effects τ_i could be a consequence of the different hospital timings in the introduction and adoption of this practice. This hypothesis cannot be validated directly since the timings of adoption of the troponin exam by the 30 different hospitals included in the analysis are not available.

The previous analysis suggests a simpler model with fixed effects only, where dummy variables represent the identified cluster structure (clusters A, B, or C). This model is easier to interpret and communicate to clinicians; for instance, it quantifies the statistical evidence of the existence of groups in terms of p-values reported in Table 38.2. The model is:

$$\tilde{N}_{ij} = \frac{\text{Asym}}{(1 + \exp(\text{Tmid}_A \cdot 1_{i \in A} + \text{Tmid}_B \cdot 1_{i \in B} + \text{Tmid}_C \cdot 1_{i \in C} - t_j))} + \varepsilon_{ij}, \tag{38.3}$$

where $i = 1, \ldots 30$, is the institution index, $j = 1, \ldots, 8$, is the year index, and ε is defined as before. Estimates for the effects of model (38.3) appear in Table 38.2; they are all significant. Notice that the fixed effects estimates reported in Table 38.2 are close to the values identifying the inflection points of the three medoids y_A, y_B and y_C generated by the analysis of model (38.2). Testing all possible contrasts between the three different fixed effects related to the inflection point always generates a p-value less than 10^{-4}; there is a strong evidence of different inflection points in the three groups. Diagnostic checks show that normality assumption of residuals can be sustained.

In conclusion, the statistical analysis advocates the presence of three groups of hospitals, possibly distinguished by different timings of introduction and adoption of the troponin test and supports the clinical tenet that in the time period 2000–2007 there has been an apparent increase in the normalized number of NON-STEMI diagnoses that is not due to a real increase in the disease incidence, but to a new diagnostic procedure adopted in hospitals along different timings.

38.3 Conclusions and Further Developments

The study presented in this chapter is a pilot example of an advanced statistical analysis performed on data drawn from a PHD. Administrative health care databases play today a central role in epidemiological evaluation of Lombardia health care system because of their widespread diffusion and low cost of information. Public health care regulatory organizations can assist decision makers in providing information based on available electronic health records, promoting the development and the implementation of the methodological tools suitable for the analysis of administrative databases and answering questions oriented to disease management. The aim of this kind of evaluation is to estimate adherence to best practice (in the setting of evidence-based medicine) and potential benefits and harms of specific health policies. Health care databases can be analyzed in order to calculate measures of quality of care (quality indicators); moreover the implementation of disease and intervention registries based on administrative databases could enable decision makers to monitor the diffusion of new procedures (as was in troponin exam adoption example) or the effects of health policy interventions. 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). More details about the AMI Project and the planned analyses can be found in [2, 3, 6–9].

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. Statistical analysis is conducted in different phases: an explorative analysis of data conducted by means of semi-parametric models to guide the study towards an appropriate parametric model, the fit of a suitable nonlinear parametric model with mixed effects estimated under the usual assumptions of random effects and residuals normality, the criticism of this model assumptions, offered by the presence of a clustering structure in the random effects, and thus the final improvement obtained through the introduction of appropriate dummy variables, taking into account the identified clustering structure, and leading to a simple and significant fixed effect logistic model.

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 tank the Working Group for Cardiac Emergency in Milano and the Cardiology Society.

References

1. 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)
2. 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. "Complex data modeling and computationally intensive statistical methods", Contribution to Statistics, pp. 41–56. Springer, New York (2010)

3. Determinazioni in merito alla “Rete per il trattamento dei pazienti con Infarto Miocardico con tratto ST elevato(STEMI)”: Decreto N^o 10446, 15/10/2009, Direzione Generale Sanità - Regione Lombardia (2009)
4. Every, N.R., Frederick, P.D., Robinson, M. et al.: A comparison of the national registry of myocardial infarction with the cooperative cardiovascular project. J. Am. Coll. Cardiol. **33**(7), 1887–1894 (1999)
5. 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 Qualit (AHRQ) Inpatient Quality Indicators. Med. Care **46**(2), 112–119 (2008)
6. Grieco, N., Ieva, F., Paganoni, A.M.: Performance assessment using mixed effects models: a case study on coronary patient care. IMA J. Manag. Math. **23**(2), 117–131 (2012)
7. 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)
8. Ieva, F., Paganoni, A.M.: Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI^{2}2 survey. Comm. Appl. Ind. Math. **1**(1), 128–147 (2010)
9. Ieva, F.: Designing and mining a multicenter observational clinical registry concerning patients with acute coronary syndromes. In: Grieco, N., Paganoni, A.M., Marzegalli, M. (eds.) Identification and Development of New Diagnostic, Therapeutic and Organizational Strategies for Patients with Acute Coronary Syndromes. Springer (2013, to appear)
10. Inmon, W.H.: Building the Data Warehouse, 2nd edn. Wiley, New York (1996)
11. Kaufman, L., Rousseeuw, P.: Finding Groups in Data. Wiley Series in Probability and Mathematical Statistics. Wiley, New York (1990)
12. 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 Publ. Health **7**, 174–89 (2007)
13. Pinheiro, J.C., Bates, D.M.: Mixed-Effects Models in S and S-Plus. Springer, New York (2000)
14. Pinheiro, J.C., Bates, D.M., DebRoy, S., Sarkar, D. and the R core team: nlme: Linear and Nonlinear Mixed Effects Models (2009)
15. 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
16. Struyf, A., Hubert, M., Rousseeuw, P.J.: Clustering in an object-oriented environment. J. Stat. Software **1** (1996)
17. Wood, S.N.: Modelling and smoothing parameter estimation with multiple quadratic penalties. J. R. Stat. Soc. B **62**(2), 413–428 (2000)
18. Wood, S.N.: Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, Boca Raton, Florida (2006)