



MOX–Report No. 02/2010

**Exploitation, integration and statistical analysis of
Public Health Database and STEMI archive in
Lombardia Region**

**PIETRO BARBIERI, NICCOLO' GRIECO, FRANCESCA IEVA,
ANNA MARIA PAGANONI AND PIERCESARE SECCHI**

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

mox@mate.polimi.it

<http://mox.polimi.it>

Exploitation, integration and statistical analysis of Public Health Database and STEMI archive in Lombardia Region

Pietro Barbieri^a, Niccolò Grieco^b, Francesca Ieva^c, Anna Maria Paganoni^c
and Piercesare Secchi^c

December 9, 2009

^a Ufficio Qualità - A.O. Melegnano
`pietro.barbieri@fastwebnet.it`

^b Centrale Operativa 118 - Ospedale Cà Granda - Milano
`niccolo.grieco@118milano.it`

^c MOX - Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
`francesca.ieva@fastwebnet.it`
`anna.paganoni@polimi.it`
`piercesare.secchi@polimi.it`

Abstract

In this work we describe nature and aims of the Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction". The main goal of the Program is the construction and statistical analysis of data coming from the integration of complex clinical and administrative databases concerning patients with Acute Coronary Syndromes treated in Lombardia Region. Clinical data sets arise from observational studies about specific diseases, while administrative data arise from standardized and on-going procedures of data collection. The linkage between clinical and administrative databases enables Lombardia Region to create an efficient global system for collecting and storing integrated longitudinal data, to check them, to guarantee for their quality and to study them from a statistical perspective.

1 Introduction

The major objective of the two years Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction" (IMA Project) funded by Ministry of Health and by "Direzione Generale Sanità - Regione Lombardia" and

started on January 2009, 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 experience of the Milan network for Cardiac Emergency shows 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 analyse data in order to optimise resources. In 2006 – 2009 a pioneer 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 emergencies) in the urban area of Milano to streamline an optimal care process for patients with STEMI. The statistical analyses of data collected during six time periods lasting from 30 to 60 days (MOMI^{2.1} - MOMI^{2.6}) have supported the clinical best practice that an early pre-alarm of the Emergency Room (ER) is an essential step to improve the clinical treatment of patients. Pre-hospital and in-hospital times have been highlighted as fundamental factors we can act on to reduce the in-hospital mortality and to increase the rate of effective reperfusion treatments of infarcted related arteries. In particular the study proved that, in order to make the Door to Balloon time (DB)¹ lower than 90 minutes - limit suggested by the American Heart Association/American College of Cardiology (AHA/ACC) guidelines - it is fundamental to take and transmit the electrocardiogram as soon as possible.

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, starting from MOMI² experience, in order to construct an archive concerning patients with ACS and involving all the Cardiology Divisions of Hospitals in Lombardia. The innovative idea in this project is not only to guarantee the same procedures to such an extended and intensive-care area, but also to integrate data collected during this observational study with administrative databases (Public Health Databases - PHD) arising from standardized and on-going procedures of data collection; up to now these PHD have been used only for monitoring and managing territorial policies.

The innovative result of this plan is then the construction for each patient of an integrated longitudinal data vector containing both clinical histories and follows up on which advanced statistical analysis can be performed. These analyses are of paramount interest. Indeed, information coming from such data are much more informative and complete than those coming separately from clinical registers or administrative databases. For instance, we now have access to information concerning related pathologies and repeated procedures.

The relevance of the project is also proved by the fact that it will change

¹Door to Balloon time is a time measurement in emergency cardiac care. The interval starts with the patient's arrival in the emergency department, and ends when a catheter guide wire crosses the culprit lesion in the cardiac cath lab.

in the future data collection along a standardized and compulsory procedure for all hospitals in Lombardia²[5]. In the following Section 2 we will present the experience of the MOMI² survey and the statistical analyses performed on it. This seminal experience on the urban area of Milano has been the motivating stimulus for the wider Strategic Program. Then we will illustrate the new register designed for the Strategic Program (Section 3), called STEMI Archive, and the administrative data banks available for data integration (Section 4). Finally in the last Section we will describe the statistical techniques that will be applied for analyzing data generated by the patients involved in the study (Section 5).

2 The MOMI² study

A net connecting the territory to 23 hospitals, by a centralized coordination of the emergency resources has been activated in the urban area of Milano since 2001. Its primary aims are promoting the best utilization of different reperfusion strategies, reducing transport and decisional delays connected with logistics and therapies, and increasing the number of patients undergoing primary Percutaneous Coronary Intervention (PCI) before 90 minutes since the arrival at Emergency Room. Difficulties in reaching these goals are primary due to the fact that the urban area of Milano is a complex territory with high density of population (2.9 million resident and 1 million commuters daily) and a great number of hospitals ($n = 27$). Twenty-three of them have a Cardiology Division and a Critical Care Unit; 18 offer a 24 hour available Cath Lab for primary PCI, 5 are completed with a Cardiac Surgery unit. In order to monitor network activity, time to treatment and clinical outcomes, the data collection MOMI² on data related to patients admitted to hospitals belonging to the net was planned and made, during six periods corresponding to six monthly/bimestral collections (respectively: MOMI².1 from jun 1st to 30th 2006, MOMI².2 from nov 15th to dec 15th 2006, MOMI².3 from jun 1st to jul 30th 2007, MOMI².4 from nov 15th to dec 15th 2007, MOMI².5 from jun 1st to 30th 2008, MOMI².6 from jan 28th to feb 28th 2009). The whole dataset collects data relative to 841 patients.

The experience of the Milano network for Cardiac Emergency shows 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 STEMI and provides the opportunity to collect and analyze data in order to optimize resources. There was a great number of patients treated with reperfusion therapy (82%) with a low hospital mortality (6.7%), an extensive use of PCI (73%), and a continuous attempt to reduce DB time. Almost 62% of patients met the guidelines recommendations with a DB time smaller than 90 minutes.

The analysis of the data collected in the MOMI² surveys show that (see [8, 10, 11]) the DB time is greatly influenced by organizational pre-hospital and in-hospital factors. In particular, we found that timing of the

²Standardized and compulsory procedures for collecting data and sending them to Lombardia data banks are called *Debito Informativo*.

first ECG, means of transport to hospital, pre-alert, direct fast track to the Cath Lab and presentation at the hospital during work time, were all relevant factors for the prediction of a DB time smaller than 90 minutes. Of particular interest was the finding that execution and transmission of pre-hospital ECG (23% of patients) as well as triage within 10 minutes from ER presentation (59% of patients) were the two most important predictive factors in reducing DB time.

The analysis also focused on the dependence between principal performance indicators (in particular times to treatment) and clinical outcomes. Other studies have found conflicting results regarding the relationship between mortality and time to reperfusion with PCI. Some investigators have found a lower mortality for shorter symptom onset-to-balloon times, for all patients or just certain subgroups, such as high-risk patients [4]. Other studies did not find a lower mortality for shorter symptom onset-to-balloon times, but did find a lower mortality for shorter DB time [14]. Finally, some studies failed to find an association between mortality and pre-hospital and in-hospital times [17]. We detected [11] a connection between outcomes and times (both concerning symptoms onset and in-hospital times); in particular our analysis pointed out the dependence between the efficacy of the reperfusion therapy (measured as the 70% reduction of the ST-segment elevation an hour after the PCI) and DB time and symptom onset time.

Details of the analyses are in [11]. Dependence between the DB time and factors we can act on in order to reduce it, has been explored by means of CARTs [2]. Indeed a CART analysis using Gini's impurity index splits groups satisfying or not the limit of 90 min for DB time in terms of time of first ECG within or not 10 minutes (see also [21]), limit suggested by the AHA/ACC guidelines. In fact the distribution of the DB time in the

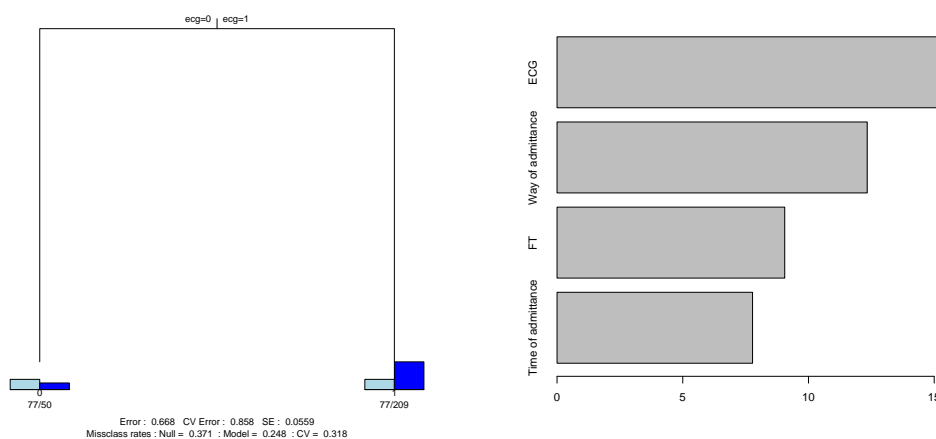


Figure 1: Left panel: CART analysis: the groups satisfying (right path) or not (left path) the limit of 90 minutes for DB time are splitted in terms of first ECG within or not 10 minutes. Right panel: Random Forest on CART predictors (drawn in the left panel) assessing discriminatory power of covariates.

population of patients with the first ECG within 10 minutes is confirmed to be stochastically smaller than the corresponding distribution in patients with the first ECG after 10 minutes; this stochastic order between the two distributions is confirmed by a Mann-Whitney test: p-value $< 10^{-12}$. A random forest analysis [3] applied to CART predictors has been performed in order to assess the discriminatory power of covariates. The message generated by these analyses is always the same: to have a DB time lower than 90 minutes, it is fundamental to take and transmit ECG as soon as possible. In fact, there is a masking effect between covariates detected by the classification analysis: way of in-hospital admittance and time of first ECG. The exact Fisher test, performed on the contingency table of the way of hospital admittance and a variable indicating if the time of first ECG is within or not 10 minutes, shows strong statistical evidence (p-value $< 10^{-10}$) of dependence between these two covariates.

In-hospital survival and reperfusion efficacy have been described through a linear logistic regression as functions of the other variables in the dataset, in order to explore the relevance of the covariates in the improvement of the two performance indicators. In-hospital survival and reperfusion efficiency are both binary independent variables (Fisher exact test: p-value = 0.244). Denoting the outcome variable under study as Y , and the set of p predictors as \mathbf{X} , a linear logistic regression model for the binary response Y can be written as

$$\text{logit}\{P(\mathbf{X})\} \equiv \log \left\{ \frac{P(\mathbf{X})}{1 - P(\mathbf{X})} \right\} = \alpha + \sum_{j=1}^p \beta_j X_j \quad (1)$$

where $P(\mathbf{X}) = \text{pr}(Y = 1|\mathbf{X})$. Clinical best practice and a stepwise model selection procedure based on backward selection with AIC criterion, pointed out Killip class, age and total ischemic time (Symptom onset to Balloon time) as explanatory variables for the survival outcome. On the other hand, DB time and Symptom onset time have been selected as explanatory variables for reperfusion efficacy. Figure 2 and 3 illustrate the results; the surfaces describe the probability of in-hospital survival and of effective reperfusion, respectively. In-hospital survival probability decreases with increasing age and total ischemic time, for both the cases of less and more severe STEMI (measured by the Killip class), but more strongly in the latter case. The probability of effective reperfusion decreases with increasing Symptom onset time and DB time.

These results support the effort of acting on some control variables (such as the reduction of DB time) in order to attain an improvement in performance indicators and thus to increase the probability of a successful treatment. The significance of Symptom onset time in modeling the reperfusion efficacy suggests that it would be strongly important to persuade the population to call the free emergency number as soon as possible after the symptom onset.

Studies such as MOMI² surveys and previous experiences of data collection carried out on the Milano intensive care area, have been seminal for two projects with larger scopes. The first one is *Progetto PROMETEO* (PROgetto Milano Ecg Teletrasmessi ExtraOspedaliero), whose goal is to provide all Basic Rescue Units operating in the urban area of Milano with

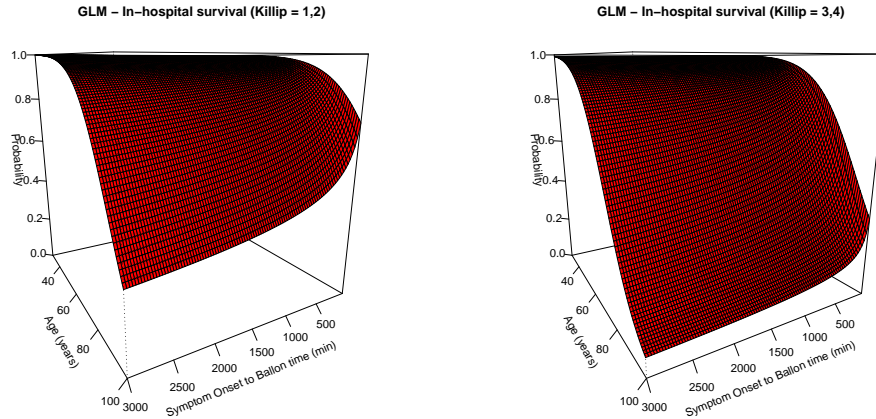


Figure 2: Left panel: In - hospital survival surface estimated by linear logistic regression model (killip class = 1 or 2). Right panel: In - hospital survival surface estimated by linear logistic regression model (killip class = 3 or 4).

the ECG teletransmission equipment. Motivation for the project comes directly out of the evidence provided by MOMI² results on the fundamental role played by an early ECG in improving survival outcome of STEMI, and highlights how the effort of monitoring data from a statistical perspective has a deep social impact. The second project, which represents the main target of the Strategic Program, aims at extending the MOMI² paradigm for collecting and analyzing data to all Cardiology Divisions of hospitals operating in the Lombardia Region. In the following sections we will focus on this second project.

3 The STEMI Archive

In this section we describe aims and contents of the STEMI Archive. The archive has been designed according to the aims of the Strategic Program and will be ready to be tested by the end of 2009. The Archive consists in the collection of clinical information related to all patients admitted in hospitals of the Lombardia Network with STEMI diagnosis, as it was in the MOMI² collection for the Milano urban area context. From the information contained in the Archive it will be possible to construct a data set where each patient will be represented by a profile with the following entries: individual serial number, date of birth, sex, time and type of symptoms onset, time to call for rescue, type of rescue unit sent (advance or basic rescue unit, that is with or without pre-hospital 12d ECG teletransmission), site of infarction on ECG, way of hospital admission, blood pressure and cardiac frequency at presentation, history of cardiac pathology, pre-hospital medication, date and hour of angioplasty (wiring), culprit lesion, ST resolution at 60 minutes, MACE (Major Adverse Cardiovascular Events) and Ejection Fraction at discharge. Personal data is collected so that the patient can be

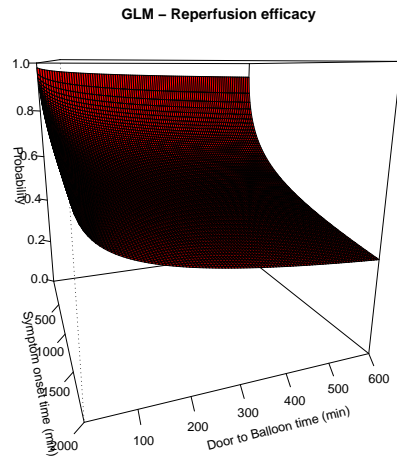


Figure 3: Reperfusion efficacy surface estimated by linear logistic regression model.

identified and a complete follow-up can be recorded. Other data will be reported to evaluate critical times (Symptom Onset time, Door to Balloon time, Door to Needle time, time of first ECG and first medical contact to reperfusion time) with the aim of designing a preferential therapeutic path to reperfusion in STEMI patients, and to direct the patient flow through different pathways according to time (e.g.: on hours vs off hours) or clinical conditions (Killip class 1 or 2 vs others). Finally, information concerning results and outcomes of the procedures will be resumed in records attesting if a subject is alive or not and if the reperfusion procedure has been effective or not. As in the MOMI² study, these data will represent some of the principal outcomes of interest. The STEMI Archive should overcome the difficulties faced with MOMI² data collections related to non-uniformity, inaccuracy of filling and data redundancy. In particular non-uniformity of data collection among different structures, or among successive surveys, and inaccuracy in filling data set fields will cease to be a problem because the Archive procedure for collecting data will become mandatory for all hospitals through a directive issued by the Lawmaker [15, 16]. All centers will fill in the registry along the same protocol and with the same software; all fields of the data bank have been agreed on with opinion leaders and Scientific Societies of cardiologist and a unique data collector was identified in the Governance Agency for Health, that will also be the data owner.

4 The Public Health Database

In this section we describe structure, aim and use of the Lombardia Public Health Database. Up to now, this database has been used for administrative purposes only, since decision makers of health care organizations

need information about efficacy and costs of health services. Randomized Controlled Trials (RCTs) remain the accepted "gold standard" for determining the efficacy of new drugs or medical procedures. Randomized trials alone, however, cannot provide all the relevant information that decision makers need to weigh the implications of particular policies affecting medical therapies. Moreover quality organizations and professional societies need information about applicability of trial findings to the settings and patients of interest. Research using disease and intervention registries, outcome studies using administrative databases and performance indicators adopted by quality improvement methods can all shed light on who is most likely to benefit, what the important tradeoffs are and how policy makers might promote the safe, effective and appropriate use of new interventions.

4.1 Healthcare databases

Administrative health care databases play today a central role in epidemiological evaluation of Lombardia healthcare 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 or the effects of health policy interventions.

The Lombardia Region Data Warehouse, called "BDA" (Banca Dati Assistito), contains a huge amount of data and requires specific and advanced tools and structures for data mining and data analysis. The structure adopted by Lombardia Region is called *Star scheme* [13], since it is centered on three main databases (AMBULATORIALE, FARMACEUTICA, RICOVERI) - containing informations about visits, drugs, hospitalizations, surgical procedure that took place in hospitals in Lombardia - while being supported by secondary databases (Assistibili, Medici, Strutture e Farmacie, Farmaci, Codici Diagnosi e Procedure Chirurgiche) which contain specific information about procedures coding or anagraphical information about people involved in the care process. The star scheme does not allow for repetitions in records entering, i.e. just one record for each person is allowed. Records may be linked in order to achieve the correct information about the basic observation unit (i.e. the individual patient/subject). However each of the above databases has its own dimension and structure, and data are different and differently recorded from one database to another. Suitable techniques are therefore required to make information coming from different databases uniform. Deterministic record linkage is the preferred tool, given the good quality of such data. Alternatively a useful approach to record linkage is given by Probabilistic Record Matching that enables the

reconstruction of links between rows of different tables damaged by data entry errors [6].

There is an increasing agreement among epidemiologists on the validity of disease and intervention registries based on administrative databases; this motivated Lombardia Region to use its own administrative database for clinical and epidemiological aims. The most critical issue when using administrative databases for observational studies is represented by selection criteria of the observation units: several different criteria may be used, and they will result in different images of prevalence or incidence of diseases. It is therefore mandatory to perform sensitivity analysis to evaluate the validity of the estimates. Statistical analysis can be performed by means of multiple logistic regression models for studying outcomes and by means of survival analysis when studying failure times (hospital readmissions, continuity of drug prescriptions, survival times). Multilevel models can also be adopted if structural and organizational variables are measured. When outcomes are the main focus of the observational study, appropriate risk adjustment tools are needed. Hospital discharge records may be analyzed with the indicators developed by Agency for Health care Research and Quality (AHRQ) that include efficient risk adjustment tools within a multiple logistic regression model. In disease management programs the Johns Hopkins Adjusted Clinical Groups (ACG) methodology and the Classification Reasearch Group (CRG) classification system have been proposed. The choice of the most suitable analysis for the Lombardia health information system is currently a topic of discussion.

4.2 Health information system in Lombardia

Health information systems in Lombardia experienced a rapid growth as a consequence of the introduction in the Italian health management of Diagnosis Related Groups (DRGs) in 1995. The development of health care measures for the specific aim of health system financing, gave rise to the availability of information useful for evaluating the efficiency of the providers and the efficacy of their activities. The development of health information systems was particularly pronounced in hospitals, and this extended the possibilities of measuring their activities: from the "classic" indicators (average length of stay, occupancy rate, turnover interval), measuring bare hospitality, to more meaningful evaluations linked to patient classification systems and to the actual opportunity of calculating quality indicators. Several regional and national rules introduced in recent years a large number of indicators in the Italian national health system. However, most of these indicators measure only few aspects of the health system: costs, degree and characteristics of supply, organizational factors, access to health care, population health status. Few indicators measure patient outcomes or evaluate the processes within the hospitals; indications about criteria for the definition of such measures are scanty and research about the validation of the indicators has not been properly developed. On the basis of these considerations the National Agency for Regional Health Services in Lombardia "Agenzia per i Servizi Sanitari Regionali" (ASSR) developed a set of quality measures (outcome and process indicators) in the context

of the Strategic Program founded by the Ministry of Health.

Indeed, one of the main goals of the Strategic Program is finding a set of indicators useful for comparison of health care providers and for the identification of factors which can produce different outcomes. Interpretation of the results will need to be supported by information collected to measure confounding factors due to differences in case-mix or selective health care options. This will be obtained by the STEMI Archive.

5 The statistical perspective

In this section we describe the statistical tools that we will use in order to model in-hospital survival and treatment efficacy outcomes. The identification of principal prognostic factors of outcomes is the main goal of the statistical analysis we will conduct. Some preliminary results obtained in a pilot study [12] support the choice of the methodological approach we will use in the analysis of data available after the linkage between STEMI Archive and hospital discharge data from Public Health Database will take place. Results from these preliminary statistical analysis are very promising; unfortunately we are not yet allowed to discuss them in public, because the matter is covered by a non disclosure agreement with the health governance of Lombardia Region.

We identified suitable statistical techniques to make an effective dimensional reduction of the complex longitudinal data vectors representing patients. *Frailty models* appear to be appropriate for capturing and characterizing information arising from hospital discharge data. Data coming from health databases are usually affected by a huge variability, called overdispersion. The main cause for this phenomenon is the grouped nature of data: each patient is a grouping factor with respect to its own admissions, while hospitals are a grouping factor with respect to admitted patients, and so on. In this study, we will model outcomes using the hospital as grouping factor. The choice is based on clinical considerations and practical evidence. Indeed, a negative outcome can be due to a bad performance of the structure with respect to a patient (without considering the initial conditions of the patient), but also to a good performance with respect to a patient arrived in very bad conditions. If we could split the effect on outcome due to the hospital from the outcome variability due to the different patient initial conditions, we would be in the position to generate health indicators of performance and a benchmark, that will make hospitals aware of their standing in the wider regional context. These goals could be achieved by fitting *Generalized linear mixed models* on data coming from the integrated database.

5.1 Frailty models

First we study the previous clinical history of patients, and in particular the sequence of hospital discharge data coming from Public Health Database. We focus on a general class of semiparametric models for recurrent events, such as hospitalizations, proposed by Peña and Hollander [18, 19]. Consider a patient that is being monitored for the occurrence of a recurrent

event over a time period $[0, \tau]$; τ could be a random time (for example the time registered in the STEMI Archive) following an unknown probability distribution function. Let $0 \equiv S_0 < S_1 < \dots$ be the random times of occurrences. Let $\mathbf{X}(\mathbf{s})$ be a possibly time-varying, observable q -dimensional vector of covariates such as gender, age, concurrent diseases. So we are dealing with the trajectories of the following counting process

$$N(s) = \sum_{j=1}^{+\infty} \mathbf{I}\{S_j \leq s, S_j \leq \tau\},$$

which represents the number of occurrences of the recurrent event (hospitalization) during the period $[0, s]$; in [19] the authors propose a general model for the hazard rate function $\lambda(s)$ of the process N :

$$\lambda(s|Z, \mathbf{X}) = \mathbf{Z}\lambda_0(\mathbf{s} - \mathbf{S}_{\mathbf{N}(s^-)})\rho(\mathbf{N}(\mathbf{s}^-); \alpha)\psi(\beta^t \mathbf{X}(\mathbf{s})). \quad (2)$$

Z is a random variable which represent the unobservable frailty of the patient, λ_0 is a unknown baseline hazard rate function, the function $\rho(\cdot; \alpha)$ incorporates the effect of the accumulating event occurrences and the link function ψ summarizes the covariates contribution. Many authors [9, 22] interpret the frailty as modeling the effect of an unobserved covariate which leads some patients to have more occurrences than others. In particular (2) is a random effect model for time-to-event data where the random effect has a multiplicative effect on the baseline hazard function. By assuming specific forms for the law of Z some elegant mathematical results can be derived; in fact a common choice for the unknown frailty distribution is a gamma distribution with unit mean and variance $1/\xi$ ($Z \sim \Gamma(\xi, \xi)$). Imposing the restriction that the gamma shape and scale parameters are equal guarantees model identifiability (see [19]) and estimation of the model parameters $(\xi, \lambda_0(\cdot), \alpha, \beta)$.

5.2 Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMM) are a natural extension of Generalized Linear Models (GLM). GLM extend ordinary regression by allowing nonnormal responses and a link function of the mean. GLMM is a further extension that permits random effects as well as fixed effects in the linear predictors. An extensive overview on these topics can be found in [1, 7, 20]. In this regression setting, parameters that describe factor effects in ordinary linear models are called fixed effects; they apply to all categories of interest. By contrast, random effects usually apply to a sample. For a study using a sample of hospitals the model treats observations from a given hospital as a cluster, and assumes a random effect for each hospital. Let Y_{ij} be the response of subject i in cluster j , $i = i_1, \dots, i_j$. In our case the responses are not only in-hospital survival and reperfusion efficacy, but also number of re-hospitalizations or re-procedures after the trigger event registered in the STEMI Archive. The number of observations may vary by cluster. Let \mathbf{X}_{ij} denote a vector of explanatory variables, such as age, procedure times, symptom onset times, estimated frailty, for fixed effect model parameters γ . Let \mathbf{U}_j denote the vector of random effects for hospital j (for example

exposure of the hospital). This is common to all observations in the cluster. Let \mathbf{Z}_{ij} be a vector of their explanatory variables. Let $\mu_{ij} = \mathbb{E}(Y_{ij}|\mathbf{U}_j)$. The linear predictor for a GLMM has the form

$$g(\mu_{ij}) = \gamma^t \mathbf{X} + \mathbf{U}_j^t \mathbf{Z}_{ij}$$

where g is the link function. For binary outcomes the link function g is the logit link. The random effect vector \mathbf{U}_j is assumed to have a multivariate Normal distribution $\mathcal{N}_q(\mathbf{0}, \Sigma)$. The covariance matrix Σ depends on unknown variance components and possibly also correlation parameters.

The main goal is the joint estimation of (γ, Σ) . Parameters pertaining to the random effects may be also of interest as a useful summary of the degree of heterogeneity of the population. Indeed the ratio between two components of fixed effects γ is the ratio of partial derivative of the log odds with respect to the corresponding covariates, and could be thought as a measure of the relative covariates strength in modeling the outcome.

5.3 Hierarchical Multilevel Modeling

GLMM are appealing when treating grouped data coming from health databases, but there are some computational problems, connected with the estimation of regression parameters, when they are used with binary data such as the outcomes of interest in our analysis.

The hierarchical model formulation where the outcome Y_{ij} is modeled conditionally on random effects, which are in turn then modeled in an additional step, makes the Bayesian paradigm appealing for interpreting and fitting GLMMs. The complete likelihood is in this case

$$L(\gamma, \Sigma) = \prod_{j=1}^N \int \prod_{i=1}^{n_j} f_{ij}(y_{ij}|\mathbf{u}_j, \gamma) f(\mathbf{u}_j|\Sigma) d\mathbf{u}_j \quad (3)$$

where N is the number of hospitals and n_j is the number of patients in each hospital. The key problem in maximizing (3) is the presence of N integrals over the q -dimensional random effect \mathbf{u}_j ; in the case of Bernoulli outcomes no analytic expressions are available for these integrals and numerical approximations are needed. The most common approach approximates the integrand using the Laplace method. A different approach represents data as a sum of a mean and an error term, and the first one is obtained as Taylor expansion around the fixed effect component (Penalized Quasi Likelihood - PQL) or around the sum of fixed and random component (Marginal Quasi Likelihood -MQL). Unfortunately both these methods produce biased estimates of model parameters when applied to binary or unbalanced data; this is going to be the case for our readings. To overcome this problem, we will consider a direct numerical approximation of the likelihood.

The Bayesian approach to GLMM fitting is also appealing because the distinction between fixed and random effects no longer occurs, since every effect has a probability distribution. Use of a flat prior distribution yields a posterior that is a constant multiple of the likelihood function. Then, Markov chain Monte Carlo (MCMC) methods can be used for approximating intractable posterior distributions and their mode.

6 Conclusions

The creation of an efficient Regional Network to face the ST-segment Elevation Myocardial Infarction is made possible by the design of the STEMI Archive and its integration with the regional Public Health Database; the link between the two databases will generate the primary platform for the study of impact and care of STEMI on the whole territory of Lombardia Region. Previous data gathering and statistical analysis restricted to the urban area of Milano were compelling for the realization of this complex and challenging project. This innovative and pioneering experience should become a methodological prototype for the optimization of health care processes in the Lombardia Region.

Acknowledgement 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] Agresti, A.: Categorical Data Analysis, Wiley (2002)
- [2] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees, Wadsworth & Brooks, Monterey, California (1984)
- [3] Breiman, L.: Random Forest. Machine Learning **45**, 1, 5-32 (2001)
- [4] Cannon, C.P., Gibson, C.M., Lambrew, C.T., Shoultz, D.A., Levy, D., French, W.J., Gore, J.M., Weaver, W.D., Rogers, W.J., 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 **283** (22), 2941-2947 (2000)
- [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] Fellegi, I., Sunter, A.: A Theory for Record Linkage. Journal of the American Statistical Association **64** (328), 1183-1210 (1969)
- [7] Goldstein, H.: Multilevel Statistical Models, Multilevel Models Project (1999)
- [8] 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 Sup-

plemento "Crema Cardiologia 2008. Nuove Prospettive in Cardiologia, **9**, 56 - 62 (2008)

- [9] Hougaard, P.: Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika* **71**, 75-83 (1984)
- [10] 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>
- [11] 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>
- [12] 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 (2009)
- [13] Inmon, W.H.: Building the Data Warehouse. John Wiley & Sons, second edition (1996)
- [14] Jneid, H., Fonarow, G.C., Cannon, C.P., Palacios, I.F., Kilic, T. et al.: Impact of Time of Presentation on the Care and Outcomes of Acute Myocardial Infarction. *Circulation* **117** 2502-2509 (2008)
- [15] Krumholz, H.M., Anderson, J.L., Bachelder, B.L., Fesmire, F.M.: ACC/AHA 2008 Performance Measures for Adults With ST-Elevation and Non-ST-Elevation Myocardial Infarction. *Circulation*. **118** 2596-2648 (2008)
- [16] Masoudi, F.A., Bonow, R.O., Brindis, R.G., Cannon, C.P., DeBuhr, J., Fitzgerald, S., Heidenreich, P.A.: ACC/AHA 2008 Statement on Performance Measurement and Reperfusion Therapy. *Circulation*. **118** 2649-2661 (2008)
- [17] MacNamara, R.L., Wang, Y., Herrin, J., Curtis, J.P., Bradley, E.H. et al.: Effect of Door-to-Balloon Time on Mortality in Patients with ST-Segment Elevation Myocardial Infarction. *Journal of American College of Cardiology* **47**, 2180-2186 (2006)
- [18] Peña, E., Hollander, M.: Models for recurrent events in reliability and survival analysis. In: Soyer, R., Mazzucchi, T. Singpurwalla, N. (eds.) *Mathematical Reliability: An Expository Perspective*, pp. 105-123. Kluwer Academic Publishers, Dordrecht (2004)
- [19] Peña, E., Slate, E.H., González, J.R.: Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference* **137**, 1727-1747 (2007)
- [20] C.Pinheiro, D.M.Bates: *Mixed-Effects Models in S and S-Plus*, Springer (2000).
- [21] Ting, H.H., Krumholtz, H.M., Bradley, E.H., Cone, D.C., Curtis, J.P. et al.: Implementation and Integration of Prehospital ECGs into Sys-

tem of Care for Acute Coronary Syndrome. *Circulation* (2008) Available

<http://circ.ahajournals.org>

- [22] Vaupel, J.W., Manton, K.G., Stallard, E.: The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography* **16**, 439–454 (1979)

MOX Technical Reports, last issues

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

- 02/2010** PIETRO BARBIERI, NICCOLO’ GRIECO, FRANCESCA IEVA,
ANNA MARIA PAGANONI AND PIERCESARE SECCHI:
*Exploitation, integration and statistical analysis of Public Health Database
and STEMI archive in Lombardia Region*
- 01/2010** G.M. PORTA, S. PEROTTO, F. BALLIO:
*Anisotropic Mesh Adaptation Driven by a Recovery-Based Error Esti-
mator for Shallow Water Flow Modeling*
- 35/2009** C. D’ANGELO, P. ZUNINO:
*Robust numerical approximation of coupled Stokes and Darcy flows ap-
plied to vascular hemodynamics and biochemical transport*
- 34/2009** P. F. ANTONIETTI, P. HOUSTON:
*A Class of Domain Decomposition Preconditioners for hp-Discontinuous
Galerkin Finite Element Methods*
- 33/2009** S. VANTINI:
*On the Definition of Phase and Amplitude Variability in Functional
Data Analysis*
- 32/2009** S. MICHELETTI, S. PEROTTO:
*Anisotropic adaptation via a Zienkiewicz-Zhu error estimator for 2D
elliptic problems*
- 31/2009** PAOLA F. ANTONIETTI, PAOLO BISCARI, ALALEH TAVAKOLI,
MARCO VERANI, MAURIZIO VIANELLO:
Theoretical study and numerical simulation of textiles
- 30/2009** Z. C. XUAN, T. LASSILA, G. ROZZA,
A. QUARTERONI:
*On computing upper and lower bounds on the outputs of linear elasticity
problems approximated by the smoothed finite element method*
- 29/2009** P.F. ANTONIETTI, A. PRATELLI:
Finite Element Approximation of the Sobolev Constant
- 28/2009** L. GAUDIO, A. QUARTERONI:
Spectral Element Discretization of Optimal Control Problems