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

Pietro Barbieri, Niccolò Grieco, Francesca Ieva, Anna Maria Paganoni, Piercesare Secchi

Abstract 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 the Ministry of Health and by ”Direzione Gen-

Pietro Barbieri
Ufficio Qualità - Cernusco sul Naviglio, e-mail: pietro.barbieri@fastwebnet.it

Niccolò Grieco
A.O. Niguarda Cà Granda, AAT 118 - Milano, e-mail: niccolo.grieco@118milano.it

Francesca Ieva
MOX - Dipartimento di Matematica, Politecnico - Milano, e-mail: francesca.ieva@mail.polimi.it

Anna Maria Paganoni
MOX - Dipartimento di Matematica, Politecnico - Milano, e-mail: anna.paganoni@polimi.it

Piercesare Secchi
MOX - Dipartimento di Matematica, Politecnico - Milano, e-mail: piercesare.secchi@polimi.it

erale 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 MOMI^2 (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 MOMI^2 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.

Other different experiences encouraged the effort of Lombardia Region in designing and supporting this challenging project. In particular it is worth mention-

[1] Door to Ballon 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.

ing the experience of Strategic Program: ”Detection, characterization and prevention of Major Adverse Cardiac Events after Drug Eluting Stent implantation in patients with Acute Coronary Syndrome”, developed in Emilia Romagna Region, and the implementation of the REAL registry (*REgistro regionale AngiopLastiche dell'Emilia-Romagna*). This is a large prospective web-based multicenter registry designed to collect clinical and angiographic data of all consecutive Percutaneous Coronary Interventions (PCI) performed in a four-million residents Italian region. Thirteen public and private centres of interventional cardiology participate to data collection. Procedural data are retrieved directly and continuously from the resident databases of each laboratory, which share a common pre-specified dataset. In this case, follow-up is obtained directly and independently from the Emilia-Romagna Regional Health Agency through the analysis of the Hospital Discharge Records and the Mortality Registries. This ensures a complete follow-up for 100% of patients resident in the Region [33, 34]. The existence of this parallel register proves the scientific interest and relevancy of these procedures in the health care policy.

The relevance of the project is also proved by the fact that it will change in the future data collection along a standardized and compulsory procedure for all hospitals in Lombardia[2][8].

In the following Section 2 we present the experience of the MOMI^2 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. We then 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 describe the statistical techniques that will be applied for analyzing data generated by the patients involved in the study (Section 5).

2 The MOMI^2 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 27 hospitals, a great number of different healthcare structures. 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

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

clinical outcomes, the data collection MOMI^2 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: MOMI^2.1 from Jun 1st to 30th 2006, MOMI^2.2 from Nov 15th to Dec 15th 2006, MOMI^2.3 from Jun 1st to Jul 30th 2007, MOMI^2.4 from Nov 15th to Dec 15th 2007, MOMI^2.5 from Jun 1st to 30th 2008, MOMI^2.6 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 MOMI^2 surveys show that (see [13, 17, 18]) the DB time is greatly influenced by organizational pre-hospital and in-hospital factors. In particular, we found that timing of the 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 [6]. 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 [23]. Finally, some studies failed to find an association between mortality and pre-hospital and in-hospital times [27]. We detected [18] 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 [18]. Dependence between the DB time and factors we can act upon in order to reduce it, has been explored by means of CARTs [4]. 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 [36]), limit suggested by the AHA/ACC guidelines. In fact the distribution of the DB time in the 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

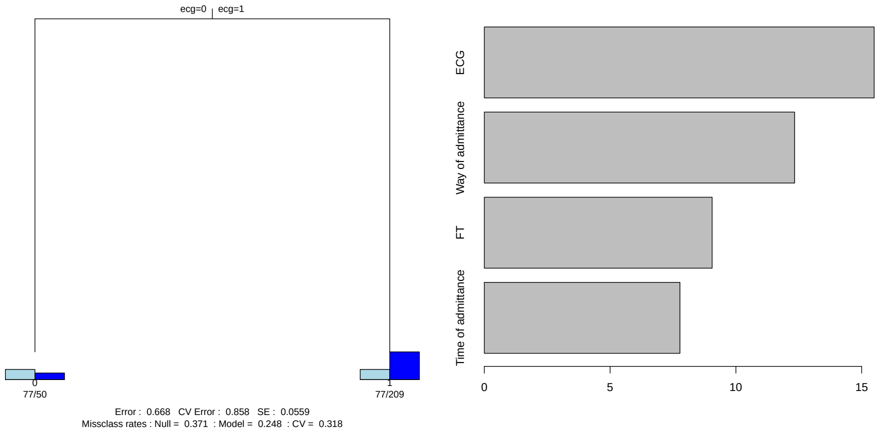

Fig. 1 Left panel: CART analysis: the groups satisfying (right path) or not (left path) the limit of 90 minutes for DB time are splitted in terms of first ECG within or not 10 minutes. Right panel: Random Forest on CART predictors (drawn in the left panel) assessing discriminatory power of covariates.

the two distributions is confirmed by a Mann-Whitney test: p-value $< 10^{-12}$. In order to asses the discriminatory power of covariates we performed a random forest analysis [5] applied to CART predictors. The length of bars in the right panel of Figure 1 is proportional to the discriminatory power of each variable in splitting the groups of patients satisfying or not the limit of 90 minutes for DB time. Time of first ECG and way of admittance are pointed out as the most important covariates to distinguish the two groups. Investigation on the dependency structure between these two covariates showed a masking effect between the covariates detected by the classification analysis. An 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. 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.

The main outcomes (in-hospital survival and reperfusion efficacy) have been described through linear logistic regressions 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

$$\mathrm{logit}\{\mathrm{P}(\mathbf{X})\} \equiv \log\left\{\frac{\mathrm{P}(\mathbf{X})}{1-\mathrm{P}(\mathbf{X})}\right\} = \alpha + \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}} \mathrm{X}_{\mathrm{j}} \tag{1}$$

where $P(\mathbf{X}) = \mathrm{pr}(Y = 1|\mathbf{X})$. 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. Figures 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.

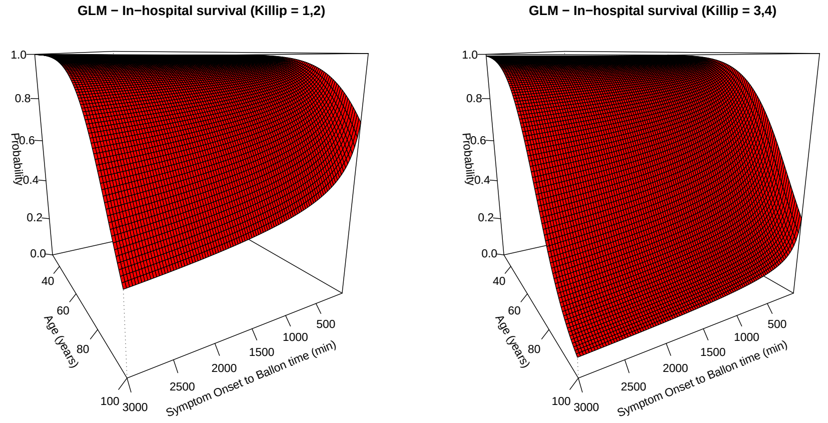

Fig. 2 Left panel: In - hospital survival surface estimated by linear logistic regression model (killip class = 1 or 2). Right panel: In - hospital survival surface estimated by linear logistic regression model (killip class = 3 or 4).

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 very important to persuade the population to call the free emergency number as soon as possible after the symptom onset.

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

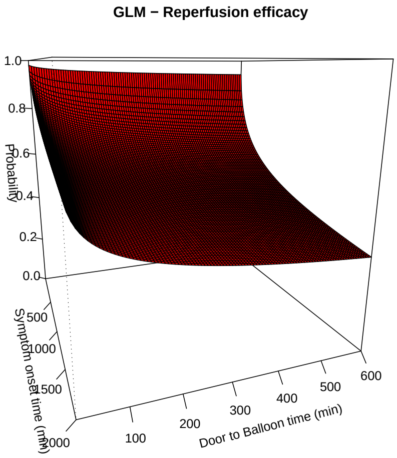

Fig. 3 Reperfusion efficacy surface estimated by linear logistic regression model.

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 MOMI^2 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 identified and a complete follow-up can be recorded. Other data are 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 trough 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 reperfusive procedure has been effective or not. As in the MOMI^2 study, these data will represent some of the principal outcomes of interest. The STEMI Archive should overcome the difficulties faced with MOMI^2 data collections related to non-uniformity, inaccuracy of filling and data redundance. 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 [24, 26]. 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 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 trade-offs 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 [22], 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. The longitudinal data that we will analyze will be generated by deterministic record linkage tools between STEMI Archive and the databases *Ambulatoriale, Ricoveri* of the BDA; and by probabilistic record matching [10] between STEMI Archive and database *Farmaceutica* which is not entirely based on the personal individual code (*Codice Fiscale*).

There is an increasing agreement among epidemiologists on the validity of disease and intervention registries based on administrative databases [9, 25, 3, 14, 38, 2]; 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 the 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. 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 the 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 [11, 35, 16].

4.2 Health information systems 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 [21]. 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 [19] 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. After splitting 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.

In the following subsections we summarize three different tools of complex data statistical modeling: frailty models, generalized linear mixed models and bayesian hierarchical models. Frailty models have been widely and successfully used (see for instance [31, 37]) to model the hazard function of hospitalizations process of patients. The estimated hazard functions, different between patients because of the random effect of the latent frailty variable, could be used as a functional data to cluster different risk subpopulations, or as a prognostic factor for the outcomes of the acute event registered in the STEMI Archive. Generalized linear mixed models seem very promising to model and explain the binary or counting outcomes of interest, not only adjusting the analysis with respect to the traditional fixed covariates, but also taking into account the overdispersion due to the grouped nature of data. Some preliminary results about the use of mixed models on MOMI^2 data are summarized in [20]. Bayesian hierarchical models have been used to study variations in health care utilization for multilevel clustered data, such as patients clustered by hospital and geographic origin (see for instance [7, 29]) and the data collected in this project present exactly a multilevel clusterd structure.

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 [30, 31]. 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 $0 \equiv S_0 < S_1 < \cdots$ 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 [31] the authors propose a general model for the hazard rate function $\lambda(s)$ of the process N:

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

Z is a random variable which represents 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 [15, 37] interpret 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 [31]) and estimation of 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, 12, 32]. 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, $i = i_1,...,i_j$. 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

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

where g is the link function. For binary outcomes the link function g is the logit link. The random effect vector $\mathbf{U}_{\mathbf{j}}$ 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 Bayesian Hierarchical Models

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

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

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 expression is 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 exactly the case we are dealing with in 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. Then, Markov Chain Monte Carlo (MCMC) methods can be used for approximating intractable posterior distributions and their mode [28].

6 Conclusions

Studies such as MOMI^2 surveys and previous experiences of data collection carried out on the Milano intensive care area, have been seminal for two projects with broader 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 the ECG teletrasmission equipment. Motivation for the project comes directly out of the evidence provided by MOMI^2 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 MOMI^2 paradigm for collecting and analyzing data to all Cardiology Divisions of hospitals operating in the Lombardia Region. 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.

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, the Cardiology Society, and the 118 Dispatch Center.

References

1. Agresti, A.: Categorical Data Analysis, Wiley (2002)
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. Barendregt, J.J., Van Oortmarssen, J.G., Vos, T. et al.: A generic model for the assessment of disease epidemiology: the computational basis of DisMod II. Population Health Metrics **1**, (2003)
4. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees, Wadsworth & Brooks, Monterey, California (1984)
5. Breiman, L.: Random Forest. Machine Learning **45**, 1, 5-32 (2001)
6. 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)

7. Daniels, M.J., Gastonis, C.: Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization. Journal of the American Statistical Association **94** (445), 29-42 (1999)
8. 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)
9. 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)
10. Fellegi, I., Sunter, A.: A Theory for Record Linkage. Journal of the American Statistical Association **64** (328), 11831210 (1969)
11. 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. Medical Care **46** (2), 112-119 (2008)
12. Goldstein, H.: Multilevel Statistical Models, Multilevel Models Project (1999)
13. 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)
14. Hanratty, R., Estacio, R.O., Dickinson L.M., et al.: Testing Electronic Algorithms to create Disease Registries in a Safety Net System. Journal of Health Care Poor Underserved, **19** (2), 452-465 (2008)
15. Hougaard, P.: Life table methods for heterogeneous populations: Distributions describing the heterogeneity. Biometrika **71**, 75-83 (1984)
16. Hughes, J.S., Averill, R.F., Eisenhandler, J. et al.: Clinical Risk Groups (CRGs). A Classification System for Risk-Adjusted Capitation-Based Payment and Health Care Management. Medical Care **42** (1), 81-90 (2004)
17. 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
18. 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
19. 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)
20. 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. Submitted (2010)
21. Gli Indicatori per la qualit: strumenti, metodi, risultati. Supplemento al numero 15 di Monitor (2005) Available: http://www.agenas.it/monitor_supplementi.html
22. Inmon, W.H.: Building the Data Warehouse. John Wiley & Sons, second edition (1996)
23. 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)
24. 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)
25. 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)
26. 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)

27. 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)
28. Molenberghs, G.,Verbeke, G.: Linear Mixed Models for Longitudinal Data, Springer, (2000)
29. Normand, S-L. T., Shahian, D.M.: Statistical and Clinical Aspects of Hospital Outcomes Profiling. Statistical Science **22** (2), 206226 (2007)
30. Peña, E., Hollander, M.: Models for recurrent events in reliabilityand survival analysis. In: Soyer, R., Mazzucchi, T. Singpurwalla, N. (eds.) Mathematical Reliability: An Expository Perspective, pp. 105-123. Kluwer Academic Publishers, Dordrecht (2004)
31. 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)
32. Pinheiro, C, Bates, D.M.: Mixed-Effects Models in S and S-Plus, Springer (2000).
33. Saia, F., Piovaccari, G., Manari, A., Guastaroba, P., Vignali, L., Varani, E., Santarelli, A., Benassi, A., Liso, A., Campo, G., Tondi, S., Tarantino, F., De Palma, R., Marzocchi, A.: Patient selection to enhance the long-term benefit of first generation drug-eluting stents for coronary revascularisation procedures. Insights from a large multicentre registry. EuroIntervention **5**(1), 57-66 (2009)
34. Saia, F., Marrozzini, C., Ortolani, P., Palmerini, T., Guastaroba, P., Cortesi, P., Pavesi, P.C., Gordini, G., Pancaldi, L.G., Taglieri, N., Di Pasquale, G., Branzi, A., Marzocchi A.: Optimisation of therapeutic strategies for ST-segment elevation acute myocardial infarction: the impact of a territorial network on reperfusion therapy and mortality. Heart **95**(5), 370-376 (2009)
35. Sibley, L.M., Moineddin, R., Agham,M.M. et al.: Risk Adjustment Using Administrative Data-Based and Survey-Derived Methods for Explaining Physician Utilization. Medical [Epub ahead of print] (2009)
36. Ting, H.H., Krumholtz, H.M., Bradley, E.H., Cone, D.C., Curtis, J.P. et al.: Implementation and Integration of Prehospital ECGs into System of Care for Acute Coronary Sindrome. Circulation (2008) Available: http://circ.ahajournals.org
37. 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)
38. Wirehn, A.B., Karlsson, H.M., Cartensen J.M., et al.: Estimating Disease Prevalence using a population-based administrative healthcare database. Scandinavian Journal of Public Health, **35**, 424-431 (2007)