MOX–Report No. 26/2010

Designing and mining a multicenter observational clinical registry concerning patients with Acute Coronary Syndromes

Francesca Ieva, Anna Maria Paganoni

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

mox@mate.polimi.it http://mox.polimi.it

Designing and mining a multicenter observational clinical registry concerning patients with Acute Coronary Syndromes

Francesca Ieva[a] and Anna Maria Paganoni[a]

July 19, 2010

[a] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
francesca.ieva@mail.polimi.it
anna.paganoni@polimi.it

Keywords: Clinical registries, Health service research, Biostatistics and bioinformatics, Generalized Linear Mixed Models.

Abstract

In this work we describe design, aims and contents of the ST-segment Elevation Myocardial Infarction (STEMI) Archive, which is a multicenter observational clinical registry planned within the Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction". This is an observational disease registry that collects clinical indicators, process indicators and outcomes concerning STEMI patients admitted to any hospital in Lombardia Region. This registry is arranged to be automatically linked to the Public Health Database, the on going administrative databank of Lombardia Region. In this work we also provide an example of statistical tools implemented on a pilot integrated database in order to explore and model such an informative database.

1 Introduction

In this work we present and describe the ST-segment Elevation Myocardial Infarction (STEMI) Archive, a multicenter observational clinical registry planned within the Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction" and

funded by Italian Ministry of Health and by "Direzione Generale Sanitá - Regione Lombardia". We also present the statistical analyses which can be performed on data arising from integration of this new registry with administrative ones already existing in Lombardia Region. These analyses will be performed in order to monitor, control and evaluate health care providers and to give support to decision for people involved in health care governance.

The STEMI Archive consists of clinical information collection related to patients admitted in all hospitals of the Lombardia Region with STEMI diagnosis. Starting from 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, mode of hospital admission, blood pressure and cardiac frequency at presentation, history of cardiac pathology, pre-hospital medication, date and hour of angioplasty, 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 previous clinical history and follow-up can be traced. Other data are reported to evaluate treatment times with the aim of designing a preferential therapeutic path to reperfusion in STEMI patients, and to direct the patient flow trough different pathways according, for example, to on hours vs off hours of working time table, or to clinical conditions such severity of infarction. Finally, information concerning results and outcomes of the procedures will be resumed in records attesting for example if a subject is discharged alive or not and if the reperfusive procedure has been effective or not.

The STEMI Archive should overcome the difficulties faced with previously performed data collections (for example MOMI^2 - MOnth MOnitoring Myocardial Infarction in MIlan, see Barbieri et al. 2010, Ieva and Paganoni 2010 for details) 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 has become mandatory for all hospitals through a directive issued by the lawmaker (Krumholtz et al. 2008, Masoudi et al. 2008). 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 cardiologists and a unique data collector was identified in the Governance Agency for Health, that will also be the data owner. Moreover, since this registry is designed to be linked with administrative databases, inaccuracy of information will be partially overcome by the fact that, once it will be linked, all information contained in it will be checked for coherence with those contained in Public Health Databases (PHD). Then only information of interest will be extracted, avoiding redundance and achieving greater accuracy and reliability.

In fact, 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 (PHD) arising from standardized and on-going procedures of data collection; up to now this PHD has been used only for administrative purposes, i.e. monitoring and managing territorial policies. With integration procedures it will be possible to construct a longitudinal data vector containing both clinical history and follow up of each patient inserted in the clinical registry. On this data, advanced statistical analyses can be performed, since continuous monitoring and statistial analyses are necessary in order to evaluate process indicators (mainly treatment times) and make aware each provider about its own position with respect to guidelines, and to improve quality and efficacy of service.

In the following we will then describe the new clinical registry on STEMI (STEMI Archive) we were asked to design (Section 3) and the administrative database (Public Health Database of Lombardia Region) it is already designed to be linked to (Section 4). Then we will describe the project of statistical analyses to be performed on such kind of complex data (Section 5), and we will present an example of these analyses performed on a pilot database already available from previous studies (Section 6).

2 Cardiovascular disease and health policy in Lombardia Region

The pathology we are interested in (STEMI) belongs to the wider class of Acute Coronary Syndromes. In particular an Acute Myocardial Infarction (AMI), also known as a heart attack, is the interruption of blood supply to part of the heart, causing heart cells to die. This is most commonly due to occlusion of a coronary artery following the rupture of a vulnerable atherosclerotic plaque, which is an unstable collection of lipids and white blood cells in the wall of an artery. The resulting ischemia (restriction in blood supply) and oxygen shortage, if left untreated for a sufficient period of time, can cause damage or death (infarction) of heart muscle tissue (myocardium). STEMI is a particular Myocardial Infarction, which can be diagnosed by ECG observing abnormal elevation of ST segment of the ECG curve itself. It highlights bad patterns, for example those characterized by ST-segment elevation. An early reperfusion therapy is one of the most important goal that must be achieved in the context of STEMI. The way to do this could pass through fibrinolysis or Percutaneous Transluminal Coronary Angioplasty (PTCA). The former one consists in a pharmacological treatment which causes a breakdown of the blood clots, while in the latter one an empty and collapsed balloon on a guide wire, known as *Balloon* catheter, is passed into the narrowed or obstructed vessels and then inflated to a fixed size. The balloon crushes the fatty deposit, so that the vessel can be opened up, the blood flow

improved, and then balloon is collapsed and withdrawn. The success of this technology is mainly due to the possibility to make an early diagnosis and so to reduce the intervention time which is particularly important in the treatment of the myocardial infraction with the ST segment elevation. In fact the main target of the STEMI treatment is to achieve an effective coronary revascularization as soon as possible.

The strategy of the connecting net between territory and hospitals, made by a centralized coordination of the emergency resources, gives the possibility to optimize therapeutic choices and so to reduce the intervention time. Randomized clinical trials have shown that reperfusion therapy provided to eligible patients reduces the risk of death due to all causes. The timeliness of reperfusion therapy is of central importance, because the benefits of therapy decrease rapidly with delays in treatment. Thus, American Hearth Association and American College of Cardiology (ACC/AHA) guidelines recommend that fibrinolysis should be provided within 30 minutes of first medical system contact and that primary PTCA within 90 minutes of first medical system contact for patients presenting with STEMI (see Antman et al. 2008, Masoudi et al. 2008, Ting et al. 2008). These guideline recommendations have been translated into performance measures for provider profiling that are reported to the public by the Centers for Medicare & Medicaid Services (CMS) and the Joint Commission. Measurement of the time to reperfusion therapy involves challenges that have hampered the acceptance of these performance measures among some clinicians and hospitals. In particular the health care governance of Lombardia Region established by the *Decreto 10446* (see D.R.G. 2009) which are the treatment times to be measured in order to judge the hospitals quality of care service and chose the STEMI Archive as main tool for collecting, analyzing and evaluating the goals achieved by individual hospitals. As consequence of this law, Regione Lombardia planned a two months data collection (May-June 2010) test, in order to get clinical institutions used to this practice. In this first experimental phase only 20 medical institutions (including public and private ones) are involved. Data concerning about 400 patients are expected to be gathered. The following planned collection is scheduled for October-November 2010 and we estimate a fourfold increase in the number of hospitals and patients involved. In the future STEMI Archive data collection will become a standardized and compulsory procedure for all hospitals in Lombardia.

3 The STEMI Archive

In this section we describe aims and contents of the STEMI Archive. The Archive is a multicenter observational prospective clinical study. Primary Outcome Measures are Incidences of Major Adverse Cardiovascular Events (MACE) defined as any one of the following: In-hospital mortality, Acute myocardial infarction/reinfarction, Cardiogenic shock, Stroke, Long term Mortality, Major

bleeding. A Secondary Outcome Measure is reperfusion effectiveness measured quantifying the reduction of ST segment elevation one hour later the surgery: if the reduction is larger than 70% we could consider the procedure effective.

The eligible cohort consists in all patients admitted to hospitals of the Lombardia Network with STEMI diagnosis, as it was in the MOMI^2 collection for the Milano urban area context (see Grieco et al. 2007, 2008 and Ieva 2008). With 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:

- Demographic data: *Codice Fiscale* (the alpha-numeric identity code used to identify people who have fiscal residence on Italian territory), date of birth, sex, weight, height, hospital of admission;
- Pre-hospital data: diabetes, smoking, high blood pressure, high cholesterol level, history of cardiac pathology;
- Admission data: time and type of symptoms onset, time of first medical contact, time to call for rescue, type of rescue unit sent (advanced or basic rescue unit, that is with or without pre-hospital 12d ECG teletransmission), time of first ECG, site of infarction on ECG, mode of hospital admittance, Fast Track activation, Killip class (which quantify in four categories the severity of infarction), blood pressure, cardiac frequency, ejection fraction and creatinine at presentation, site of ST-elevation, number of leads with ST-elevation, pre-hospital hearth failure;
- Therapeutic data: time of fibrinolysis (Door to Needle time), time of angioplasty (Door to Balloon time), culprit lesion, ST resolution at 60 minutes, Major Adverse Cardiovascular Events, Ejection Fraction at discharge.

Finally, information concerning results and outcomes of the procedures will be resumed in records attesting if a subject is discharged alive or not and if the reperfusive procedure has been effective or not. As in the MOMI^2 study, these data represent some of the principal outcomes of interest. Moreover, we are also interested in pointing out process indicators to be used in order to explain primary and secondary outcomes, and then in identifying covariates providers could act upon to improve therapeutic results.

4 Integration with Public Health Database

In this section we describe structure, aim and use of the Lombardia Public Health Database (PHD), in order to better understand what we mean with "patient clinical history" we can obtain through integration among STEMI Archive and PHD itself.

Up to now, this databank has been used only for administrative purposes, since decision makers of health care organizations need information about efficacy and costs of health services. Anyway, there is an increasing agreement

among epidemiologists on the validity of disease and intervention registries based on administrative databases (see Barendregt et al. 2003, Every et al. 1999, Hanratty et al. 2008, Manuel et al. 2007, Wirehn et al. 2007); this motivated Lombardia Region to use its own administrative databases for clinical and epidemiological aims.

Administrative health care databases can be easily analyzed in order to calculate measures of quality of care (quality indicators). The importance of this kind of database for clinical purposes depends on the fact that they provide all the relevant information that decision makers need to weigh, in order to evaluate the implications of particular policies affecting medical therapies (information about applicability of a trial findings to the settings and patients of interest, effectiveness and widespread of new surgical techniques, estimation of adherence to best practice and potential benefits/harms of specific health policies, etc). Moreover, 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.

The Lombardia Region PHD contains a huge amount of data and requires specific and advanced tools for data mining and data analysis. The datawarehouse structure of Lombardia PHD is called Star scheme (see Inmon 1996), since it is centered on three main databases (*Ambulatoriale*, *Farmaceutica*, *Ricoveri*) - containing informations about visits, drugs, hospitalizations, surgical procedures that took place in hospitals in Lombardia - and it is supported by secondary databases (*Assistibili*, *Medici*, *Strutture e Farmacie*, *Farmaci*, *Codici Diagnosi* e *Procedure Chirurgiche*) which contain specific information about drugs and procedures coding or anagraphical information about people involved in the care process. The star scheme does not allow for repetitions in records entering, for example just one record for each admission to hospital is admitted, and each record finishes with patient discharge. Inside the PHD, several records may correspond to the same patient over time, even concerning the same event. In fact an "Event" is the total of admissions and discharges related to the same episode of disease. A patient may have several events during years, and each event could consist of multiple admissions. For each admission/discharge, one record of PHD is produced.

Records related to the same subject may be linked in a temporal 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 one. Suitable techniques are therefore required to make information coming from different databases uniform and not redundant. The longitudinal data that we will analyze will be generated by deterministic record linkage between STEMI Archive and the databases *Ambulatoriale*, *Ricoveri* of the PHD; and by probabilistic record matching (see Fellegi and Sunter 1969) between STEMI Archive and database *Farmaceutica* which is not entirely based on *Codice Fiscale*. The Lombardia Region data manager and owner will provide

us an encrypted code for each patient in order to protect citizen's privacy.

Now, when in PHD datawarehouse we look for events related with a patient belonging to the population selected by STEMI Archive, we find all his clinical history in term of health care utiliztion (drugs, hospital admissions, visits, etc). Of course we are not interested in all this huge amount of information, since our interests and studies are concentrated on cardiovascular diseases. The most critical issue when using administrative databases for observational studies, in fact, is represented by the selection criteria of the observation records: several different criteria may be used, and they will result in different images of prevalence or incidence of diseases. Up to now, the most accepted and practical way to select admissions concerned a specific disease is the Diagnosis Related Groups (DRGs), introduced in the Italian health care system in 1995. This identification code is assigned to each record of PHD to identify the class of service the hospital has to be refunded for. The analysis of DRGs of interest for purposes of analyses in still object of debate among statisticians and physicians. For the pilot example we present in the next section, selection of clinical events from PHD concerning patients pointed out by clinical registry has been primarly performed using DRGs codes 516, 517, 122 and 127, which are the main codes related to cardiovascular diseases.

For further details on these topics see Ieva and Paganoni 2009b, Barbieri et al. 2010, Every et al. 1999, Glance et al. 2008, Hanratty et al. 2008, Hughes et al. 2008, Sibley et al. 2009.

4.1 Example of integrated data

Let now focus on the integrated data we will obtain by linkage procedure. Let consider a fictitious case of a patient drawn from a clinical registry. Once he/she is inserted in the STEMI Archive, we have all information (see Section 3) related with this trigger event. Now, we look for all clinical events of this subject contained in PHD database concerning the previous 8 years, using his/her encrypted version of *Codice Fiscale*. We then link the Archive record with all those contained in the PHD, which provide the following information:

- data concerning patients admissions to hopitals during this time period, in terms of kind of admission provider (with or without certain department or devices) and time spent in hospitals (hospital of admission, date of arrival and disharge, cause of discharge) → from databases *Ricoveri*, *Strutture e Farmacie*;

- anagraphical data (sex, age, place of birth and residence, class of income), in order to establish the amount which Lombardia Region has to refund for health practice (different procedures are applied for patient who do not lives in Lombardia) → from databases *Ricoveri*, *Assistibili*;

- department of stay, kind of intervention received (description of surgial

practice, DRG code, classification of pathology class) → from databases *Farmaceutica*, *Ambulatoriale*, *Codici Diagnosi*, *Procedure Chirurgiche*;

- data concerning kind and quantity of drugs prescribed as consequence of practice patient has been subjected to. Information about costs, quantity prescribed (daily or monthly), active ingredients, class of membership and pathology each drug is prescribed for are provided → from databases *Farmaceutica*, *Farmaci*, *Medici*, *Ambulatoriale*.

To be noticed is that for a linkage on 8 years time slot, a single patient could have order of dozens admissions, hundreds of visits, drugs and procedures. This makes the integrated data huge and complex, even for a single patient, and the STEMI Archive is designed to collect order of 400 patients for each bimontly data collection.

Of course some information contained in the linked database are redundant, and attention must be paid to a carefully selection of covariates. In this sense, several further problems arise: firstly, as already mentioned, we must select only cardiovascular events and events in some way related to this pathology; then we should also perform a dimensional reduction, pointing out just covariates which can be of interest. This is very hard work, strongly related with the clinical questions that physicians want to investigate.

5 Statistics

The integrated database we presented in the previous Section will be the object of statistical analyses to be performed starting from autumn 2010, when the first official data collection of STEMI Archive will be implemented. Now essentially we plan to remake, adapt and extend the study conducted on the MOMI^2 survey on STEMI patients in Milano (2006 - 2009). For further details about the MOMI^2 study see Grieco et al. 2007, 2008, 2010, Ieva 2008, Ieva and Paganoni 2009a, 2010.

In particular effective variable selection and suitable data dimensionality reduction are of paramount importance. Non parametric partitioning methods, as CART (Classification and Regression Trees), tests on independence between predictors, explorative data mining will highlight possible dependence patterns between covariates.

Moreover 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 to hospital, while hospitals are a grouping factor with respect to admitted patients, and so on. So we will model our primary and secondary outcomes using the hospital as grouping factor. The choice is based on clinical considerations, practical evidence and provider profiling aims. After splitting the effect on outcome due to the hospital from the outcome variability due to the

different case-mix, we would be in the position to generate health performance indicators and benchmarks, that will make hospitals aware of their standing in the wider regional context. These goals could be achieved by fitting Generalized linear and additive models with parametric or non parametric random effects (see McCullagh and Nelder 1989, Goldstein 2003, Hastie and Tibshirani 1999, Aitkin 1999) on data coming from the integrated database. Aiming at finding a ranking or clustered structure of hospitals we will study estimated random effects of each clinical institution.

As we said before, the administrative data available for our readings will cover a very large time period (2002 - 2010) of the patients clinical history. In particular, data we are interested in mostly concern the hospitalization's process of patients. Therefore we will model these data as trajectories of a marked point process (see Baraldo et al. 2010). The great challange in doing this starting from integrated database and not only from the PHD datawarehouse is that an overcome of main problems concerning observational studies can be reached: in fact, using information of the previous patient history, we can account for case mix, while observational studies in general do not allow researchers to do this. Moreover, the linkage between information coming from registry and administrative data makes possible to insert estimates of clinical history of patients (resumed for example by estimated hazard functions of readmission for each patient) in a wider model constructed to explain and predict the main outcomes.

Then we will adopt a semiparametric method for estimation of hospitalization process hazard functions, for example the one described in Peña et al. 2006. We will assume, for patient i, $i = 1, ..., n$, with covariate vector $\boldsymbol{X_i(s)} = (X_{i1}(s), ..., X_{iq}(s))^T$ (possibly time-dependent) the following form for cumulative hazard function:

$$\Lambda_i(s|\boldsymbol{X}_i, Z_i) = \int_0^s Z_i \lambda_0[w] \alpha^{N_i(w^-)} \exp[\boldsymbol{\beta}^T \boldsymbol{X}_i(w)] dw. \tag{1}$$

where Z_i is an unobservable nonnegative random variable, called *frailty* which represents an unobservable source of variability for the risk of re-hospitalization; Z_i typically is assumed to have a parametric distribution (for example from the Gamma family), the parameters of which are determined by an EM algorithm. The baseline hazard function $\lambda_0(\cdot)$ is assumed to be unknown and represents the instantaneous probability of first event occurrence with null covariates and constant frailty, N_i is the counting process of hospitalizations, α and $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)^T$ are real unknown parameters. These components contribute to the definition of a very general model, which embodies various important aspects that influence the hospitalization process. The resulting estimated hazard function is a functional covariate we will consider in modeling outcomes, following Crainiceanu et al. 2009, and fitting a Generalized Multilevel Functional Linear Model.

6 A pilot case study

In this Section we present some preliminary results on a pilot database that arises from integration of third and fourth MOMI^2 data collections (MOMI^2.3 and MOMI^2.4), performed during June-July 2007 and November-December 2007 respectively, with the administrative database of *Schede di Dimissione Ospedaliera* (SDO), which is a subset of database *Ricoveri*, referred to the time period (2006-2008). We chose these databases in this pilot example because their role is very similar to that of STEMI Archive and PHD respectively.

Using this example, we tried to understand if the integration between such different and complex databases is a good and useful instrument for health governance of Lombardia Region. Moreover, we tried to test capability that this integration could offer to the clinicians, health governance and particulary researcher users, and to use advanced statistical techniques to gain relevant results. The database consists of 240 inpatients, for which we provided the following steps:

- spotting of the MOMI^2 event among records of SDO database, using as matching code for each patient his/her encrypted version of *Codice Fiscale*;
- construction of a longitudinal data, arranging in chronological order the events related to the same anonymous patient in the SDO database;
- integration of this longitudinal data with clinical information coming from MOMI^2.

To build this database, we used a deterministic linkage rule. This example play the role of a simple sparring partner of the database we are constructing. For example the time period to which the SDO are referred to is too narrow; in fact for the 75% of patients the hospitalization recorded in MOMI^2 registry seems to be the first acute myocardial event; this fact is not coherent with the literature statements, and it is only a masking effect. So in this case we use the SDO data mainly to validate the clinical ones, which are often wrong filled. In fact in the 3% of cases the date of MOMI^2 event does not match with the one recorded in the SDO database.

We then repeated the statistical analyses conducted on MOMI^2 data set on the integrated database, whose data are more complete and much more reliable, obtaining coherent results. Particulary, we modelled with logistic regression techniques the in-hospital survival as a function of covariates pointed out by stepwise model selection algorithms; fields like sex, age, in-hospital mortality already existing in MOMI^2 database, are now filled in a more complete and reliable way, so that we have less problems connected to missing or incoherent data.

We detected overdispersion in the outcome variable; this can be due to several different causes; one of the most reasonable to consider is the difference in terms,

for example, of number of patients yearly treated by the not negligible number of hospitals (16) involved in the study. It is known from clinical literature that health outcomes at different institutions could vary for random variation, for systematic influences of institutions or covariates on outcome, or for the health of patient populations prior to admission. It is likely that variability in observed in hospital survival, for example, will depend more on conditions of patients reaching hospitals rather than on the care they receive once admitted. For this reason, we fitted a Generalized Linear Mixed Model (GLMM), i.e. a generalized linear model with binary response (in hospital survival) with an additive random effect (see Goldstein 2003), in order to quantify the effect of the covariates on survival probability, taking the hospital of admission as a grouping factor and assuming it as random factor with Normal prior distribution.

This is just an example of analysis which it will be possible to conduct on the integrated database. For a statistical analysis in a Bayesian perspective of this dataset see Guglielmi et al. 2010. The stepwise backward selection, AIC criterion, and clinical best practice, pointed out the Symptom Onset to Balloon time (OB) in logarithmic scale (p-value = 0.236643), *killip* (p-value = 0.069251) and *age* (p-value = 0.011944) of patient as relevant covariates as significant or near significant factors in order to explain survival probability (see Rathore *et al.* 2009 for discussion on variable selection methods in clinical contexts). Specifically, *killip* is a dichotomic variable equal to 1 for more severe infarctions (Killip class equal to 3 or 4) and equal to 0 for less severe infarctions (Killip class equal to 1 or 2), *age* is the age in years of each patient at admission time and OB, as we said before, is the total ischemic time from symptom onset time to surgery practice (balloon) time.

Therefore, calling Y_{ij} the binary random variable representing in-hospital survival of patient $i = 1, \ldots, 240$ treated in the hospital $j = 1, \ldots, 16$, the models fitted is:

$$logit\left(\mathbb{E}\left[Y_{ij}|b_j\right]\right) = logit(p_{ij}) = \beta_0 + \beta_1 age_i + \beta_2 \log(OB)_i + \beta_3 killip_i + b_j \quad (2)$$

where, in (2), $b_j \sim \mathcal{N}(0, \sigma_b^2)$ is the Normal random effect of the grouping factor (i.e. hospital the $i-th$ patient is admitted to).

In Table 1 estimates of fixed effects coefficients and standard deviation of random effect are reported.

		estimate
Intercept	$\hat{\beta}_0$	15.752
age	$\hat{\beta}_1$	-0.125
log(OB)	$\hat{\beta}_2$	-0.539
killip	$\hat{\beta}_3$	-2.287
Std. Dev.	$\hat{\sigma}_b$	1.636

Table 1: Model parameters estimates.

Using the estimated coefficients reported in Table 1, we can draw the estimated in-hospital survival probability surfaces for the patients (Figure 1).

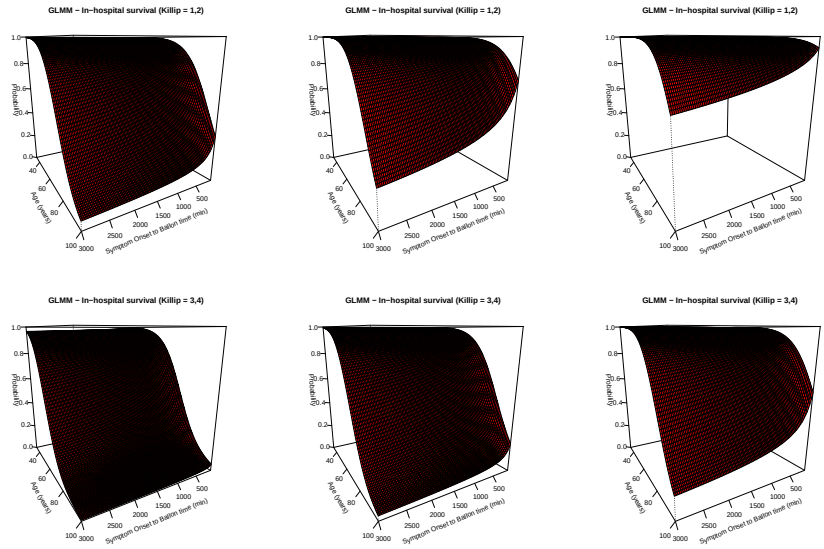

Figure 1: Estimated survival probability surfaces for less severe (upper line) and more severe (lower line) class of Killip: "bad hospital" (left), "standard hospital" (central), "good hospital" (right).

For both killip classes (less or more severe infarction) we represent three different cases: from left to the right, we consider a realization of random effect equal to $-\hat{\sigma}_b$, 0 and $\hat{\sigma}_b$ respectively. We could interpret these three cases as estimates for patients treated in a "bad", "mean" and "good" hospital respectively.

Observing how the hospital behaviour affects the survival probability, we would like to rank providers or to compare their performances with benchmarks gold standards. Procedures for analyzing and comparing health-care providers effects on health services delivery and outcomes have been referred to as provider profiling. In a typical profiling procedure, patient-level responses are measured for clusters of patients treated by different providers. Aiming at finding a clustered structure in hospital's behavior we partition with a k-means clustering algorithm (see Hartigan and Wong 1979) the set of estimated $\hat{b}_j$, $j = 1, ..., 16$. A robustness analysis for the number of clusters using the average silhouette width (Struyf et al. 1996) supported the optimal choice of $k = 2$. Indeed Fig. 2 shows the silhouette plot of this clustering procedure, and the value of average silhouette width equal to 0.69 indicate that a reasonable clustering structure has been found.

The means of the two clusters are -1.555 and 0.427, representing a "Bad" and a "Good" behavior respectively. For a deeper discussion about different techniques to classify medical institutions see Grieco et al. 2010.

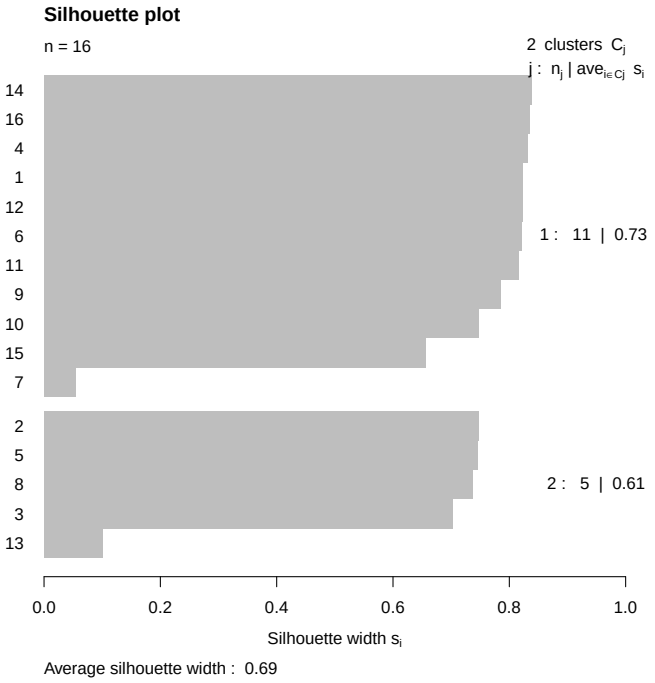

Figure 2: Silhouette plot

7 Conclusions and Open Problems

In this work we presented and described the STEMI Archive, as an example of multicenter observational clinical registry planned and designed to be integrated with administrative datawarehouse of Lombardia Region. Motivations for the project directly come from the evidence provided by results pointed out in previous similar data collections (see MOMI^2 for example). The fundamental role played by a constant monitoring of data from a statistical perspective in managing and optimizing clinical resourches, comes out also from results obtained on pilot integrated database and shown in Section 6.

We showed how 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: in fact this is the first platform for the study of impact and care of STEMI producing longitudinal data containing all the clinical history of patients of interest, which can be studied and resumed with statistical techniques we presented in Section 5, and included in more complex model for main outcomes. Moreover, provider profiling can be performed on performance indicators and they can be used to monitor and control health care offer of providers.

This innovative and pioneering experience stands as a candidate to become a methodological prototype for the optimization of health care processes in the

Lombardia Region, and in the future it could also be extended to further pathologies of interest, for their incidence and mortality, besides Cardiovascular diseases.

Acknowledgements. This work is within 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

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117-128.

Antman, E.M., Hand, M., Amstrong, P.W., Bates, E.R., Green, L.A. et al. (2008). Update of the ACC/AHA 2004 Guidelines for the Management of Patients with ST Elevation Myocardial Infarction. *Circulation* **117**, 269–329.

Baraldo, S., Ieva, F., Paganoni, A.M., Vitelli, V. (2010). Statistical models for hazard functions: a case study of hospitalizations in health failure telemonitoring. Proceedings of the XLVth Scientific Meeting of the Italian Statistical Society 2010.

Barbieri, P., Grieco, N., Ieva, F., Paganoni, A.M., Secchi, P.(2010). Exploitation, integration and statistical analysis of Public Health Database and STEMI archive in Lombardia Region. Complex data modeling and computationally intensive statistical methods - Series "Contribution to Statistics", Springer.

Barendregt, J.J., Van Oortmarssen, J.G., Vos, T. et al. (2003). A generic model for the assessment of disease epidemiology: the computational basis of DisMod II. Population Health Metrics **1**.

Crainiceanu, C.M., Staicu, A.M., Di, C.Z. (2009). Generalized Multilevel Functional Regression. *Journal of the American Statistical Association*, **104** 448, 1550–1561.

Direzione Generale Sanità – Regione Lombardia (2009). 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.

Every, N.R., Frederick, P.D., Robinson, M. et al. (1999). 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

Fellegi, I., Sunter, A. (1969). A Theory for Record Linkage. Journal of the American Statistical Association **64**, 328, 1183-1210

Glance, L.G., Osler, T.M., Mukamel, D.B. et al. (2008). Impact of the present-on-admission indicator on hospital quality measurement experience with the Agency for Healthcare Research and Quality (AHRQ) Inpatient Quality Indicators. Medical Care **46**, 2, 112-119.

Goldstein, H. (2003). Multilevel Statistical Models, Arnolds, London.

Grieco, N., Sesana, G., Corrada, E., Ieva, F., Paganoni, A.M., Marzegalli M. (2007). The Milano Network for Acute Coronary Syndromes and Emergency Services. *MESPE journal*, First Special Issue 2007.

Grieco, N., Corrada, E., Sesana, G., Fontana, G., Lombardi, F., Ieva, F., Paganoni, A.M., Marzegalli, M. (2008). Le reti dell'emergenza in cardiologia : l'esperienza lombarda. *Giornale Italiano di Cardiologia* Supplemento "Crema Cardiologia 2008. Nuove Prospettive in Cardiologia **9**, 56 - 62.

Grieco, N., Ieva, F., Paganoni, A.M. (2010). Provider Profiling Using Mixed Effects Models on a Case Study concerning STEMI Patients *Mox Report n. XX/2010*, Dipartimento di Matematica, Politecnico di Milano. [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/XX-2010.pdf

Guglielmi, A., Ieva, F., Paganoni, A.M., Ruggeri, F. (2010). A Bayesian random-effects model for survival probabilities after acute myocardial infarction *Mox Report n. 17/2010*, Dipartimento di Matematica, Politecnico di Milano. [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/17-2010.pdf

Hanratty, R., Estacio, R.O., Dickinson L.M., et al. (2008). Testing Electronic Algorithms to create Disease Registries in a Safety Net System. Journal of Health Care Poor Underserved, **19**, 2, 452-465.

Hartigan, J.A., Wong, M.A. (1979). A K-means clustering algorithm, *Applied Statistics* **28**, 100-108.

Hastie, T.J., Tibshirani, R.J. (1999). *Generalized Additive Models*, Chapman & Hall/CRC, New York.

Hughes, J.S., Averill, R.F., Eisenhandler, J. et al. (2004). Clinical Risk Groups (CRGs). A Classification System for Risk-Adjusted Capitation-Based Payment and Health Care Management. Medical Care **42**, 1, 81-90.

Ieva, F. (2008). Modelli statistici per lo studio dei tempi di intervento nell'infarto miocardico acuto. *Master Thesis*, Dipartimento di Matematica, Politecnico di Milano.
[Online]:http://mox.polimi.it/it/progetti/pubblicazioni/tesi/ieva.pdf

Ieva, F., Paganoni, A.M. (2009a). 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. [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/05-2009.pdf

Ieva, F., Paganoni, A.M. (2009b). Statistical Analysis of an Integrated Database Concerning Patients With Acute Coronary Syndromes. SCo2009, Sixth Conference - Proceedings, Maggioli editore, Milano.

Ieva, F., Paganoni, A.M. (2010). Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI^2 survey, *Communications in Applied and Industrial Mathematics*. In press.

Inmon, W.H. (1996). Building the Data Warehouse. John Wiley & Sons, second edition.

Krumholz, H.M., Anderson, J.L., Bachelder, B.L., Fesmire, F.M. (2008). ACC/AHA 2008 Performance Measures for Adults With ST-Elevation and Non-ST-Elevation Myocardial Infarction. *Circulation* **118**, 2596-2648.

Manuel, D.G., Lim, J.J.Y., Tanuseputro, P. et al. (2007). How many people have a myocardial infarction? Prevalence estimated using historical hospital data. BMC Public Health **7**, 174-89.

Masoudi, F.A., Bonow, R.O., Brindis, R.G., Cannon, C.P. et al. (2008). ACC/AHA 2008 Statement on Performance Measurement and Reperfusion Therapy A Report of the ACC/AHA Task Force on Performance Measures (Work Group to Address the Challenges of Performance Measurement and Reperfusion Therapy) *Circulation* **118** 2649-2661.

Mc Cullagh, P., Nelder, J.A. (2000). *Generalized Linear Models*, Chapman & Hall/CRC, New York.

Peña , E. A., Slate, E. H., Gonzalez, J.R. (2006). Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference*, **137**, 1727-1747.

Rathore, S.S., Curtis, J.P., Chen, U.J. (2009). Association of door to balloon time and mortality in patients admitted to hospital with ST-elevation myocardial infarction: national cohort study *British Medical Journal*, **338**.

Sibley, L.M., Moineddin, R., Agham,M.M. et al. (2009). Risk Adjustment Using Administrative Data-Based and Survey-Derived Methods for Explaining Physician Utilization. Medical [Epub ahead of print]

Struyf, A., Hubert, M. & Rousseeuw, P.J. (1996). Clustering in an Object-Oriented Environment. *Journal of Statistical Software*, **1**.

Ting, H.H., Krumholtz, H.M., Bradley, E.H., Cone, D.C., Curtis, J.P. et al. (2008). Implementation and Integration of Prehospital ECGs into System of Care for Acute Coronary Sindrome. *Circulation* **118**, 1066-1079.

Wirehn, A.B., Karlsson, H.M., Cartensen J.M., et al. (2007). Estimating Disease Prevalence using a population-based administrative healthcare database. Scandinavian Journal of Public Health **35**, 424-431.

MOX Technical Reports, last issues

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

26/2010 Francesca Ieva, Anna Maria Paganoni:
Designing and mining a multicenter observational clinical registry concerning patients with Acute Coronary Syndromes

25/2010 G. Pena, C. Prud'homme, Alfio Quarteroni:
High Order Methods for the Approximation of the Incompressible Navier-Stokes Equations in a Moving Domain

24/2010 Lorenzo Tamellini, Luca Formaggia, Edie Miglio, Anna Scotti:
An Uzawa iterative scheme for the simulation of floating boats

23/2010 Joakim Baeck, Fabio Nobile, Lorenzo Tamellini, Raul Tempone:
Stochastic Spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison

22/2010 Carlo D'Angelo, Paolo Zunino:
Numerical approximation with Nitsche's coupling of transient Stokes'/Darcy's flow problems applied to hemodynamics

21/2010 Niccolo' Grieco, Francesca Ieva, Anna Maria paganoni:
Provider Profiling Using Mixed Effects Models on a Case Study concerning STEMI Patients

20/2010 Fabio Nobile, Alfio Quarteroni, Ricardo Ruiz Baier:
Numerical solution of an active strain formulation for the electro-mechanical activity in the heart

19/2010 Loredana Gaudio, Alfio Quarteroni:
hN-adaptive spectral element discretization of optimal control problems for environmental applications

18/2010 Paola F. Antonietti, Nur Aiman Fadel, Marco Verani:
Modelling and numerical simulation of the polymeric extrusion process in textile products

17/2010 Alessandra Guglielmi, Francesca Ieva,
Anna Maria Paganoni, Fabrizio Ruggeri:
A Bayesian random-effects model for survival probabilities after acute myocardial infarction