DIPARTIMENTO DI MATEMATICA
"Francesco Brioschi"
POLITECNICO DI MILANO

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

Ieva, F.; Paganoni, A.M.

Collezione dei *Quaderni di Dipartimento*, numero **QDD 112**
Inserito negli *Archivi Digitali di Dipartimento* in data 29-11-2011

Piazza Leonardo da Vinci, 32 - 20133 Milano (Italy)

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

Francesca Ieva and Anna Maria Paganoni

***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 clinical registry that collects clinical indicators, process indicators and outcomes concerning STEMI patients admitted to any hospital of the Regional district, one of the most advanced and intensive-care area in Italy. This registry is arranged to be automatically linked to the Public Health Database, the on going administrative datawarehouse of Regione Lombardia. Aims and perspectives of this innovative project are discussed, together with feasibility and statistical analyses which are to be performed on it, in order to monitor and evaluate the patterns of care of cardiovascular patients.**

***Index Terms*—Clinical registries, Health service research, Biostatistics and bioinformatics, Provider profiling.**

I. Introduction

Assessment of service delivery at the local level of government is not a new enterprise in clinical context, but linking the measures, or indicators, to program mission, setting performance targets and regularly reporting on the achievement of target levels of performance are new features in the performance measurement movement sweeping across healthcare systems all over the world. A performance measure is a quantitative representation of public health activities, measured in order to evaluate, and then improve, performances and services. In fact, in order to improve something you have to be able to change it; in order to change it you have to be able to understand it; in order to understand it you have to be able to measure it. 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 the Regional district for healthcare, namely the "Direzione Generale Sanità - Regione Lombardia". The main goal of this program is to enhance the integration of different sources of health information in order to automate and streamline clinicians' workflow, so that data collected once can be used multiple times for different aims, and especially for measuring performances of healthcare systems, to understand how hospitals work and to increase efficacy of healthcare offer in terms of costs and patterns of care. In fact, integrated systems enable people in charge with healthcare government to obtain data for billing or performance evaluations, as well as they allow clinicians to see trends in the effectiveness of treatments or to compare patterns of care. Finally, they let researchers to analyze the efficacy and efficiency of system on patients' outcomes. In other words, integrated systems play a fundamental role in complex clinical environments.

The STEMI Archive consists of clinical information collection related to patients admitted in all hospitals of Regione Lombardia with STEMI diagnosis. Starting from information contained in the Archive, it is be possible to construct a data set where each patient is 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 treatments (thrombolysis and/or angioplasty), culprit lesion, Major Adverse Cardiovascular Events (MACE), Ejection Fraction and therapy at discharge. As in classical clinical surveys, also clinical data of the STEMI Archive are collected in order to identify subpopulations (in our case patients affected by STEMI). On the other hand, the innovative contents of this survey are represented by process indicators recorded in it: the main idea is 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. In this sense, this survey represents an instrument both for epidemiological enquiries and for organizational optimization of the cardiological healthcare networks. Moreover, personal data are collected so that the patient can be univocally identified also within administrative datawarehouse and a longitudinal electronic record containing his/her previous clinical history and follow-up can be traced, thanks to the potential of Electronic Health Record (Fascicolo Sanitario Elettronico). 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 Regione Lombardia.

F.Ieva is with MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy (e-mail: francesca.ieva@mail.polimi.it).

A. M. Paganoni is with MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy (e-mail: anna.paganoni@polimi.it).

Finally, information concerning outcomes (i.e. if a subject is discharged alive or not, if the reperfusive procedure has been effective or not) are recorded, so that they can be returned to clinicians and institutions appropriately exploited in terms of patient's case-mix and care pattern, in order to support healthcare decisions and clinical policies through monitoring and analyzing data. This latter step may be carried out through suitable statistical monitoring and modeling. Statistical models, in fact, are able to capture complexity, variability and grouped nature of these data, as we will see in Section VI, providing an evidence based decisional support as well as pursuing the optimization of healthcare offer.
The STEMI Archive should overcome the difficulties faced in previous pilot data collections (i.e. MOMI^2, GestIMA, LombardIMA, see [1], [2]) 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 dataset 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 [3]. All centers will fill in the registry along the same protocol and with the same software, thanks to the help of Lombardia Informatica (`http://www.lispa.it`), the Information & Communication Technology (ICT) society which Regione Lombardia leans on for implementation of Electronic Health Record. Opinion leaders and Scientific Societies of cardiology agreed upon all fields to be recorded and a unique data collector was identified in the Governance Agency for Health, that is also the data owner. Moreover, since this registry is designed to be automatically linked with administrative databases, inaccuracy of information will be partially overcome by the fact that, after the linkage, 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 (for further details on record linkage, see [4]).

In the following we describe the health policy and program of Regione Lombardia concerning cardiovascular diseases (Section II), then we move (Section III) to an in-depth examination of the STEMI Archive and of the administrative datawarehouse of Regione Lombardia (Section IV), i.e the Public Health Database (PHD). Details and aims of integration between the considered clinical registry and administrative databanks are described in Section V, whereas in Section VI the role of statistician and statistical analyses is discussed. Finally conclusions and open problems are presented in Section VII.

II. Cardiovascular disease and health policy in Regione Lombardia

The pathology we are interested in is a particular type of Acute Myocardial Infarction, namely ST-Elevation Myocardial Infarction (STEMI). It belongs to the wider class of Acute Coronary Syndromes (ACS), and it is caused by an occlusion of a coronary artery, which causes an ischemia (a restriction in blood supply) and an oxygen shortage. These effects, if left untreated for long, can damage the heart muscle tissue (myocardium) since the interruption of blood supply to the cells make them die (infarction). STEMI, whose incidence and mortality are very high in Italy as well as all over the world, can be diagnosed by observing abnormal elevation of ST segment in the ECG curve. An early reperfusion therapy is one of the most important goal that must be achieved in the treatment of STEMI patients, and can be obtained through thrombolysis and/or Percutaneous Transluminal Coronary Angioplasty (PTCA). The former one consists in a pharmacological treatment which causes a breakdown of the blood clots which obstruct the coronary vessel, 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. This allows the vessel to be opened up and the blood flow to be improved; then balloon is collapsed and withdrawn.
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. 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 thrombolysis 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 [5], [6], [7], [8]).
Regione Lombardia is very sensitive to Cardiovascular topics, as proved by the huge amount of social and scientific enterprises concerning these syndromes which have been carried out during past years (see [1], [9]). With STEMI Archive and Strategic Program, people in charge with healthcare governance intended to adopt new clinical instruments and already existing administrative resources to create new methods for targeting and measuring performances in cardiovascular healthcare. In particular, *Decreto 10446* [3] establishes which are the treatment times to be measured in order to judge the hospitals quality of care service and choose the STEMI Archive as main tool for collecting, analyzing and evaluating the goals achieved by individual hospitals. So STEMI Archive data collection has become a standardized and compulsory procedure for all hospitals in Regione Lombardia, since January 2011.

III. The STEMI Archive

In this section we describe aims and contents of the STEMI Archive. The Archive is a multicenter observational prospective clinical study, designed during the first phase of Strategic Program, i.e. during 2010, thanks to a collaboration among clinicians, people in charge with healthcare governance of Regione Lombardia and statisticians of Politecnico di Milano. Its filling is mandatory by law [3], and three data collections have been planned within the end of Strategic Program (December 2011). The first one has already been performed during the

time slot of January-December 2010, to set, test and calibrate it; the second one, from January 2011 to the end of June 2011, represents the first official period of data collection; finally a third collection period is planned for October-December 2011. We will refer in the following to results of feasibility analysis on integration performed on the first collection test. This had to last enough time to enable all clinical structures which are involved in the project to overcome software and technical hitches and to offer training and technical assistance to all them, especially concerning SISS system (the Italian platform for supporting Electronic Health Record, see `http://www.siss.regione.lombardia.it/`).

The STEMI Archive, as well as every survey on specific disease, enables researchers to point out a subpopulation of interest for clinical and scientific inquiries. Starting from these subpopulations, studies on effectiveness of different patterns of care and then provider profiling can be carried on, adopting models for explaining outcomes by means of suitable process indicators and adjusting for different case mix. In our case, a primary outcome measure is incidence of MACE defined as any one of the following events: in-hospital mortality, Acute myocardial 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 after the treatment: if the reduction is larger than 50% in the case of thrombolysis and 70% in the case of angioplasty we could consider the procedure effective. Process indicators and patients covariates can be resumed in the following four categories:

- 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 value at presentation, site of ST-elevation, number of leads with ST-elevation, pre-hospital hearth failure;
- Therapeutic data: time of thrombolysis (Door to Needle time), time of angioplasty (Door to Balloon time), culprit lesion, Ejection Fraction and therapy at discharge.

The eligible cohort consists then in all patients admitted to any hospitals of the Regione Lombardia Network with STEMI diagnosis.

IV. The Lombardia Region Administrative Datawarehouse

In this section we describe structure, aim and use of the Regione Lombardia Public Health Database (PHD), the datawrehouse the STEMI Archive has been designed to be integrated with. This is an on going datawarehouse, which up to now has been used only for administrative purposes, since decision makers of healthcare organizations need information about efficacy and costs of health services.

The PHD of Regione Lombardia contains a huge amount of data and requires specific and advanced tools for data mining and data analysis. The datawarehouse structure of PHD is called Star scheme (see [10]), since it is centered on three main databases (*Ambulatoriale*, *Farmaceutica*, *Ricoveri*), containing information about visits, drugs, hospitalizations, surgical procedures that took place during admission to hospitals, and it is supported by secondary databases (*Banca Dati Assistito (BDA)*, *Medici*, *Codici Diagnosi* e *Procedure Chirurgiche*) which contain more specific and administrative information about drugs and procedures coding or personal 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 allowed, and each record finishes with patient discharge. An *Event* is the total of admissions and discharges related to the same episode of disease. Inside the PHD, several records may correspond to the same patient over time, even concerning the same event. A patient may have several events during years, and each event could consist of multiple admissions. For each admission/discharge path, one record is produced in PHD. 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 databases described above 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*, *Farmaceutica*, *Ricoveri* and *BDA* of the PHD. Regione Lombardia data manager and owner provide an encrypted code for each patient in order to protect citizen's privacy. This encrypted code represents the key to obtain the deterministic linkage between the databases.

Once different sources of data have been linked, it is possible for clinicians, researchers and people involved in healthcare governance to answer epidemiological questions such: is the trigger event of the STEMI Archive the first cardiological event for the observed patient? If not, how many cardiovascular events have been recorded in the previous history of this patient? These information are provided by the integration of the STEMI Archive with *Ricoveri* administrative database. Moreover, if a patient is already known to the healthcare systems in terms of cardiovascular hospital admissions, was his therapy compliance good, i.e. did he/she assumed correct quantities of drugs and received a convenient treatment in terms of visits and clinical practice? These information are provided by the integration of the STEMI Archive with *BDA* and *Farmaci* administrative databases. Finally, how the previous clinical history of each patient affects his/her outcome observed in the STEMI Archive gathering? These questions

ask for a proper statistical modelling and represent the real and new challenges of the Strategic Program.
All the information coming from the integration let the researchers to point out new prognostic factors to be considered for better explain the main outcomes the hospitals are evaluated on. Then, the longer is the time slot on which the integration can be performed, the richer, the more complete and the more reliable is the information which can be used in order to built the outcome measures. Regione Lombardia enabled us to look at the administrative datawarehouse up to 8 years ago. In such time slot, a single patient could have order of dozens admissions, hundreds of visits, drugs and procedures. Dealing with such complex and high dimensional data is the challenge of the statistical analysis, as presented in Section VI.

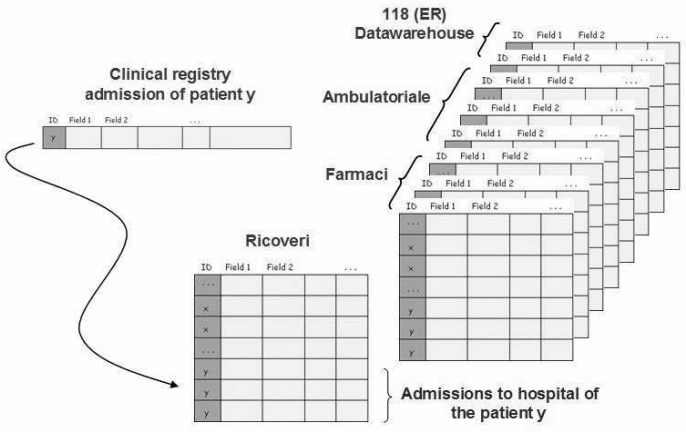

Figure 1. Sketch of integration between STEMI Archive and Public Health Database.

V. Integration with Public Health Database

In this section we discuss the results of integration, in terms of the longitudinal electronic records obtained for each patient inserted in the STEMI Archive.
Over recent years, there has been an increasing agreement among epidemiologists on the validity of disease and intervention registries based on administrative databases (see [11], [12], [13], [14], [15]); this motivated Regione Lombardia to use its own administrative databases for clinical and epidemiological aims. In fact, even if Randomized Controlled Trials (RCTs) remain the accepted "gold standard" for determining the efficacy of new drugs or medical procedures, they cannot provide alone all the relevant information that decision makers need in order to weigh the implications of particular policies affecting medical therapies. 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.
Administrative healthcare databases can be 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 know, 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 healthcare databases play today a central role in epidemiological evaluation of healthcare system because of their widespread diffusion and low cost of information.
When in the PHD we look for events related with a patient belonging to the population selected by the STEMI Archive (see Figure 1), we find all his clinical history in term of healthcare utilization (visits, hospital admissions, drugs, etc). Since we are not interested in all this huge amount of information, but only in cardiovascular events, criteria for adequately choose only the hospital discharge records effectively related to cardiovascular events of the patient of interest are needed. In fact, the most critical issue when using administrative databases within observational studies is represented by the selection criteria of the discharge records: several different criteria may be used, and they will result in different images of prevalence or incidence of diseases. Among the most accepted criteria, those referring to the Agency for healthcare Research and Quality (AHRQ) methodology, the ones of Johns Hopkins Adjusted Clinical Groups (ACG) and Classification Research Groups (CRG) have been considered (for further details, see [16]).
As we said before, integrating clinical surveys on specific diseases with administrative databanks, enable us to select subpopulation of interest for observational studies, focused on answering to specific epidemiological needs. In fact, the main point and the novelty of Strategic Program is the proposal of an epidemiological research for specific subpopulation of interest pointed out by clinical registries, which is different from the classical epidemiological inquiry since it is conducted starting from the Electronic Health Records, then it is faster and cheaper, and moreover it is real time achieving. For this new epidemiology, new methods for inquiry and analysis must be pointed out, and adequate information media must be provided. The STEMI Archive described in Section III and statistical models proposed in Section VI are some of the instruments to be adopted to this aim, and the Strategic Program is the first official set in Italy where they have been considered. For further details on these topics see [12], [13], [16], [17], [18], [19], [20].
As previously mentioned, when integration of different sources of data is performed, attention must be paid to a carefully selection of covariates and data of interest. In this sense, several further problems arise: firstly, as already mentioned, it is necessary to select only cardiovascular events and events in some way related to this pathology; then a dimensional reduction is needed, pointing out just covariates which can be of interest in exploiting outcomes by means of suitable covariates and process indicators. This is the challenge of

the statistician, and it is strongly related with the clinical questions that physicians want to investigate. In this sense, several analyses can performed on such rich and complex data. In the next section, we will give an overview of the potential of a statistical monitoring of data arising from integration of clinical registries and administrative databases.

VI. The role of statistics

As mentioned before, such complex and huge databases ask for continuous monitoring and advanced statistical tools to be applied for evaluating outcomes and expecially for pointing out the relationship between process indicators, patients' case mix, hospitals' exposure and outcomes.

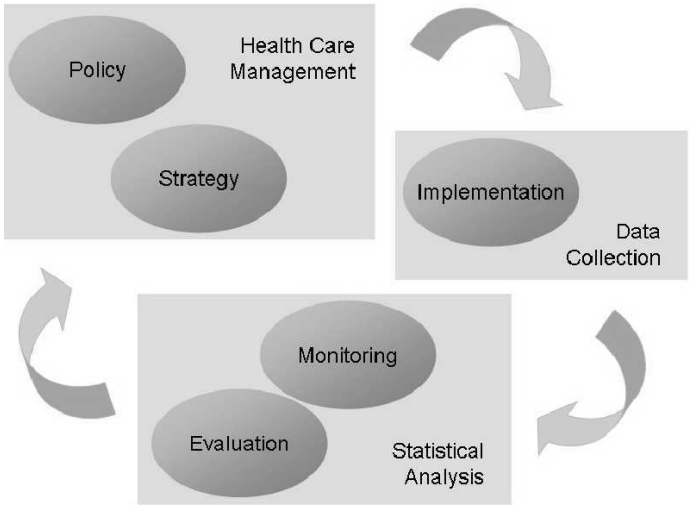

Figure 2. Flow chart of statistical monitoring and evaluation process in healthcare systems.

In fact, as can be evinced by the diagram reported in Figure 2, a statistical analysis is helpful whenever it is necessary to:

- inform policy making at regional level about the health-care system efficiency and efficacy, in order to support their decisions with clear and well defined evaluations procedures;
- improve the quality of care, enhancing those patterns of care which have been proved to be the most effective in improving outcomes of patients;
- identify poor performers;
- provide hospitals information to enable them to set strategies and policies in order to improve their service;
- provide consumer information to facilitate choice of healthcare provider.

In order to do this, it is necessary to point out what causes variation in outcomes between healthcare providers, and this can be obtained from the analysis of data described in the previous sections.

A. Analysis of STEMI Archive data

The preliminary data collection, carried out to test how the new integrated system would have worked, has been performed during the time slot from January to December 2010. It consists of 1087 patients, admitted in 31 hospitals of Regione Lombardia with STEMI diagnosis. The population is stratified according with the expected results which can be found in the literature: in particular, the mean age and standard deviation are respectively 65.75 and 13.08 years (first, second and third quantiles are respectively equal to 56, 66 and 75 years, i.e. more than a quarter of subjects is elder than 70), with men significantly younger than women (63.07 vs 72.74 years, Wilcoxon non parametric test p-value $< 2.2 * 10^{-16}$); concerning sex, males are 786 whereas females are 301 females, i.e. 72.3% vs 27.7%. Moreover, 82% of patients present the less severe infarction, indicated by Killip class I, 11% have Killip class equal to II, 3.4% equal to III and 3.6% are in the most severe class, the Killip class IV. Among these people, 41% are smokers, 18% are affected also by diabetes, 61% are high-blood pressure sufferers. Furthermore, 81.5% of the overall subjects underwent to primary PTCA. Only 4.5% of people have been treated with pharmacological therapy, as we expected since the Cardiological Network of hospitals is oversized with respect to the territorial extension and then it often happens that there is a hospital near enough to make the surgery practice preferable to the pharmacological treatment. Finally, 1039 patients (95.6%) are alive at discharge, and 78% of the overall patients treated surgically had positive outcome in terms of riperfusion (ST-segment resolution greater than 70% after one hour from the intervenction).

For all the patients inserted in the STEMI Archive during this test period, integration with the administrative datawarehouse has been performed (for all patients the deterministic link has been successful). The analyses have been focused on different clinical issues. For example, from integration with the database *Farmaci*, it is possible to check for the compliance to the prescribed therapy for patients labelled as "known" to the cardiovascular events. A patient is defined "known" if the oldest prescription for cardiovascular drugs contained in the administrative datawarehouse is elder than one year from the date of STEMI Archive hospital admission. For these patients it is possible to check if they really bought and then consumed enough quantity of drugs, and then if there is statistical evidence to say that the poorer is the compliance, the higher is the readmission probability. On the other hand, from integration of STEMI Archive with the database *Ricoveri*, information about length and frequency of previous hospital admissions for each patient can be pointed out, in order to perform survival analysis on times to the next admission. In particular, 716 of the overall 1087 patients of the STEMI Archive (i.e. 75.86%) are present in the database *Farmaci*, and 539 of them can be defined "known" in the sense we explained above. On the other hand, 903 patients of the STEMI Archive (83%) has almost a previous hospital admission recorded in the administrative datawarehouse, but only 546 of them have almost one previous admission for cardiovascular diseases. This latter information can be obtained from integration with *BDA* database, where information on category which the admission belongs to can be found, together with codes labelling the category the admission belongs to, as well as information on procedures and diagnoses.

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. In particular effective variable selection is of paramount importance. Non parametric partitioning methods, like CART (Classification and Regression Trees), tests on independence between predictors, explorative data mining will highlight possible dependence patterns between covariates (for example see [2]).

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 his/her own admissions to hospital, while hospitals are a grouping factor with respect to admitted patients, and so on. So one can model the primary and secondary outcomes using hospitals as grouping factor, so that a sort of implicit ranking/classification of providers can be provided directly by the adopted models. In fact, after splitting the effect on outcome due to the hospital from the outcome variability due to the different case-mix, it is possible to generate health performance indicators and benchmarks, that will make hospitals aware of their standing in the wider regional context. Such indexes of performances also enable healthcare governance to rank providers, to evaluate their performances and to plan activities in order to invest in quality improvement. In this sense, integration between different sources of clinical information and complex databases is a good and useful instrument for health governance. In our case, overdispersion is detectable in the outcome variable. This can be due to several different causes. One of the most reasonable to consider is the difference in terms of number of patients yearly treated by the not negligible number of hospitals (31) 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 outcomes, or for the health of patient populations prior to admission. Generalized Linear Mixed Models (GLMMs), i.e. generalized linear models with binary response (in hospital survival) and Generalized Additive Mixed Models (GAMMs) with an additive random effect [21], [22], [23], are suitable statistical methods, both in a frequentist and Bayesian framework, to quantify the effect of the covariates on survival probability, taking the hospital of admission as a grouping factor and assuming it as parametric [24], [25] or non parametric [26], [27] random effect.

Moreover, 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. Then firstly the in-hospital survival rates are to be estimated by fitting parametric or semiparmetric generalized linear mixed effects models (an example of such modelling performed adopting a Dirichlet Process can be found in [28]). Then the comparison among providers' performances can be carried out through unsupervised classification algorithms, like k-means or similar [29]. Finally, decisional criteria for classification can be pointed out minimizing the expected loss arising from misclassification costs, for example using Bayesian optimal decision rules (see [28]).

On the other hand, concerning integrated data, the main focus is the hospitalization's process of patients. In fact, it is possible to model these data as trajectories of a point process (see for example [30]). The great challenge 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 semiparametric model constructed to explain and predict the main outcomes.

VII. Conclusions and Open Problems

In this work we present and describe the STEMI Archive, as an example of multicenter observational clinical registry planned and designed to be integrated with administrative datawarehouse of Regione Lombardia. The link between the two databases will generate the primary platform for the study of impact and care of STEMI on the whole Regional district we are concerned to. Previous data gathering and statistical analysis restricted to the urban area of Milano were compelling for the realization of this complex and challenging project.

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 VI. Moreover, provider profiling can be carried out on performance indicators and they can be used to monitor and control healthcare offer of providers.

This innovative and pioneering experience stands as a candidate to become a methodological prototype for the optimization of healthcare processes in Regione Lombardia, and to be extended in the future to different pathologies of interest, for their incidence and mortality, besides Cardiovascular Diseases.

Acknoledgement

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

[1] 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.

[2] 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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Oltrona, L., Mafrici, A., Marzegalli, M., Fiorentini, C., Pirola, R., Vincenti, A., a nome dei Partecipanti allo Studio GestIMA e della Sezione Regionale Lombarda dell'ANMCO e della SIC (2005). La gestione della fase iperacuta dell'infarto miocardico con sopraslivellamento del tratto ST nella Regione Lombardia (GestIMA), *Italian Heart Journal*, Suppl **6**, 489-497.

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

[11] 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**.

[12] 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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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]

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

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

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

[24] 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*, **1**, 1, 128–147.

[25] Guglielmi, A., Ieva, F., Paganoni, A.M., Ruggeri, F. (2011). A Bayesian random-effects model for survival probabilities after acute myocardial infarction. *Chilean Journal of Statistics*, to appear.

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

[27] Grieco, N., Ieva, F., Paganoni, A.M. (2011). Performance assessment using mixed effects models: a case study on coronary patient care. *IMA Journal of Management Mathematics*, to appear.

[28] Guglielmi, A., Ieva, F., Paganoni, A.M., Ruggeri, F. (2011). Process indicators and outcome measures in the treatment of Acute Myocardial Infarction patients. Accepted for publication on *Statistical Methods in Healthcare*, Wiley (2011).

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

[30] S. Baraldo, F.Ieva, A.M.Paganoni, V.Vitelli (2010). Generalized functional linear models for recurrent events: an application to readmission processes in heart failure patients. Tech. Rep. MOX 42/2010, Dipartimento di Matematica, Politecnico di Milano. [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/42-2010.pdf

Francesca Ieva received a Master Degree in Mathematical Engeneering - Statistical Methods from Politecnico di Milano (Italy) in 2008. Since 2009 she is a PhD student in Mathematical Models and Methods in Engineering program at the Politecnico di Milano (Italy), Department of Mathematics and works on the Strategic Program of Regione Lombardia "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction". She is a member of Società Italiana di Statistica (SIS), Society for Industrial and Applied Mathematics (SIAM), Operating Research (OR) Society, and Royal Statistical Society (RSS). Her readings carried out at MOX concern biostatistics and statistical methods for provider profiling in healthcare.

Anna Maria Paganoni received a Laurea in Physics from the Università di Milano (Italy) in 1994, a Doctorate in Mathematics from the Università di Milano (Italy) in 1998. Since 2010, she is an associate professor of Statistics at the Politecnico di Milano (Italy), Department of Mathematics. She is a member of UMI and SIS. She e coordinates the statistical unit within the Politecnico Unit in the Strategic Program of Regione Lombardia "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction". Her research activities carried out at MOX include adaptive design of experiments and statistical models for biostatistics.