MODELLISTICA E CALCOLO SCIENTIFICO

MOX

MODELING AND SCIENTIFIC COMPUTING

MOX–Report No. 42/2010

Generalized functional linear models for recurrent events: an application to re-admission processes in heart failure patients

Stefano Baraldo, Francesca Ieva,
Anna Maria Paganoni, Valeria Vitelli

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

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

Generalized functional linear models for recurrent events: an application to re-admission processes in heart failure patients

Stefano Baraldo, Francesca Ieva, Anna Maria Paganoni and Valeria Vitelli♯

November 18, 2010

♯ MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica “F. Brioschi”
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
stefano1.baraldo@mail.polimi.it
francesca.ieva@mail.polimi.it
anna.paganoni@polimi.it
valeria.vitelli@mail.polimi.it

Keywords: public health databases, point processes, functional data analysis, generalized linear models.

AMS Subject Classification: 62P10, 62M09

Abstract

An effective methodology for dealing with data extracted from clinical heart failure databases and the Public Health Database is proposed. A model for recurrent events is used for modeling the occurrence of hospital readmissions in time, thus deriving a suitable way to compute individual cumulative hazard functions. Estimated cumulative hazard trajectories are then treated as functional data, and their relation to clinical relevant responses is studied in the framework of generalized functional linear models.

1 Introduction

Heart failure is a degenerative disease known worldwide as one of the most important causes of hospitalization among the eldest in the population. Since the frequency of crises undergone by a given patient increases along time, a growing employment of health care resources in terms of money, structures and personnel is needed. The necessity of a cost-effective solution for the care of this and other chronic pathologies has led to the use of telemedicine as a possible srategy (see Capomolla et al. (2004), Giordano et al. (2008) and Scalvini et al. (2004)).

The basic idea of telemonitoring is to keep the patient at home and to instruct her/him about the use of monitoring instruments, which send registered information (ECG, body weight, heart frequency, etc.) to the health institution by a network connection. The physician in charge evaluates received data to properly manage the home care program, for example by modifying drug doses and by scheduling visits.

Telemonitoring databases contain information, like duration of the telemonitoring period, number of ECGs transmitted to the hospital, NYHA[1] class of the patient, clinical parameters at starting and ending times, etc., mostly regarding the telemonitoring period itself. Telemonitoring outcome, i.e. the conclusion of the planned period (usually 6 months) without interruption by adverse events, should be related to the patients' clinical history to get some insight into the effectiveness and applicability of this strategy. For this reason, we consider Hospital Dimission Forms (*Schede di Dimissione Ospedaliera*, briefly SDO) extracted from Public Health Databases, which gather detailed information about hospitalization periods. The use of hospitalization information to study telemonitoring outcome is an innovative approach, since no standard methodology exists to exploit this kind of data. Heart failure is a pathology that alternates phases of stability to sudden worsenings of the patient's condition; for this reason it is not possible to assume a stationary pattern for these events. Dealing with time dependent observations of localized events, a natural modeling approach, yet new to the field of telemonitoring, is to consider each patient's hospitalizations as points of a non stationary, doubly stochastic counting process. The model we consider derives from the class of models introduced in Limnios and Nikulin (2000), and applied in Peña et al. (2007) to the study of intervention effects after cancer relapse. This class of models is very general, and allows us to take into account many aspects that influence hospitalization risk. Moreover, it enables us to compute the realized trajectories of the cumulative hazard process underlying the hospitalizations counting process, constructing longitudinal data that summarize complex characteristics of the patient's clinical history. Cumulative hazard processes are then studied in the light of functional data analysis techniques (see Ramsay and Silverman (2005) for a general presentation of the subject), and used to construct a generalized functional linear model.

The paper is structured as follows. Section 2 describes the theoretical and methodological framework: the model for recurrent events is introduced; then smoothing of cumulative hazard functions obtained by realized trajectories of the recurrent event processes and dimensional reduction performed via functional principal components are detailed; moreover, generalized functional linear mod-

[1] The New York Heart Association (NYHA) classification divides in four classes the extent of heart failure: 1 is the less severe, 4 the most.

els are presented. Section 3 presents the motivating application, practical issues and results. Finally, Section 4 contains some concluding remarks and discussion of future works.

2 Theoretical framework

2.1 Model for recurrent events

Let $(\mathcal{F}_t)_{t\in I}$ be a filtration associated to the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $I = [0, \tau]$. We define the counting process $(N(t))_{t\in I}$ adapted to $(\mathcal{F}_t)_{t\in I}$ as follows:

$$N(t) = \sum_{j=0}^{\infty} I\{S_j \leq t, S_j \leq \tau\}, \tag{1}$$

where S_j represents the calendar time of the j-th occurrence of the observed event and τ represents a random censoring time for the process.

N is a submartingale such that, for every stopping time T, $N(T)$ is uniformly integrable, then the Doob-Meyer decomposition theorem states that there exists a unique predictable, non decreasing, cadlag and integrable compensator (or *cumulative hazard*) process $(\Lambda(t))_{t\in I}$ such that

$$M = N - \Lambda \tag{2}$$

is a zero-mean, uniformly integrable martingale (see, for example, Andersen et al. (1993)). Hence the distribution of event times is completely characterized by the knowledge of process Λ, on which modeling efforts should then be focused. We assume that

$$\Lambda(t) = \int_0^t C(s)\lambda(s)ds, \tag{3}$$

where $C(s) = I\{s \leq \tau\}$ is the *at-risk process*, and $(\lambda(s))_{s\in I}$ is called *hazard function*, or *intensity process*.

A wide variety of models for the intensity process can be found in literature, ranging from Poisson processes to the Cox model (1972), additive models, frailty and dynamic models (see for instance Aalen et al. (2006) and Andersen et al. (1993) for presentation and discussion of various possibilities). Our choice for the target problem is the following form of intensity: for $i = 1, ..., n$ subjects with covariate vector $\boldsymbol{X}_i(t) = (X_{i1}(t), ..., X_{iq}(t))^T$ (eventually time-dependent), we have

$$\lambda(t|\boldsymbol{X}_i) = \lambda_0[\mathcal{E}_i(t)]\alpha^{N_i(t^-)}e^{\boldsymbol{\beta}^T\boldsymbol{X}_i(t)}, \tag{4}$$

where $\lambda_0(\cdot)$ is an unknown baseline hazard function, $\mathcal{E}_i$ is a time warping function, called *effective age*, α is a real parameter and $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)^T$ a q-dimensional vector of real coefficients.

The model assumed in equation (4) is a specification of the general class of models proposed in Peña et al. (2007), which is very flexible, and thus capable of capturing many different possible behaviors of recurrent events processes. With respect to the general model presented in Peña et al. (2007), we choose to account for unobserved heterogeneity by using the dynamic component $\alpha^{N_i(t^-)}$ instead of a frailty variable, i.e. a multiplicative random effect, which usually masks the entity of population hazard when frail subjects get right censoring for decease (see Aalen et al. (2006) for a discussion on frailty and dynamic models). It is worth mentioning also that computations without random effects are more stable and faster to carry out. A dependence of intensity from process state of the form $\alpha^{N_i(t^-)}$ has been chosen because of its clear interpretation: in fact, values of α higher than 1 indicate that a new event implies a worsening of the patient's condition, increasing future rehospitalization risk, viceversa for α values lower than 1. Moreover, in absence of information about the shape of $\mathcal{E}_i$ two natural choices can be made: *perfect repair*, which corresponds to $\mathcal{E}_i(t) = t - t_{N_i(t^-)}$, with $t_{N_i(t^-)}$ being the last process jump time before time t, and *minimal repair*, which is the identity function. Since a perfect restoring of health status after a hospitalization, modeled with perfect repair, seems too optimistic, we use $\mathcal{E}_i(t) = t$ for each $i = 1, ..., n$.

Adding a censoring variable to account for different observation times, the model for cumulative hazard can be written as follows, for patients $i = 1, ..., n$

$$\Lambda_i(t|\boldsymbol{X}_i) = \int_0^t C_i(s)\lambda_0(s)\alpha^{N_i(s^-)} \exp[\boldsymbol{\beta}^T \boldsymbol{X}_i(s)]ds, \tag{5}$$

where $C_i(s) = I\{s \leq \tau_i\}$ (i.e., subjects have different censoring times τ_i, assumed to be mutually independent). Independent censorship as defined in Kalbfleisch and Prentice (1980) can be reasonably assumed for the considered problem, as we will deepen in the following.

2.2 Cumulative hazard smoothing and reconstruction

Semiparametric estimation of cumulative hazard, as proposed in Peña et al. (2007), produces a step function estimate of baseline hazard $\Lambda_0(t)$, which has the following expression: defining t_j as the j-th observed jump time of the aggregated process $N_\bullet(t) = \sum_{i=1}^n N_i(t)$ and $\tau = max_{i=1,...,n}\tau_i$

$$\widehat{\Lambda}_0(t) = \sum_{t_j \leq t} \frac{1}{\sum_{i=1}^n C_i(t_j)\hat{\alpha}^{N_i(t_j^-)} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{X}_i(t_j)}}, \qquad t \in (0, \tau],$$

where $\hat{\alpha}$ and $\hat{\beta}$ are maximum likelihood estimates of α and β.

Assuming the real Λ_0 function to be absolutely continuous, we deal with the issue of smoothing its estimate $\widehat{\Lambda}_0$, successively moving on to the reconstruction of cumulative hazard process realizations for each patient.

The function $\Lambda_0(t)$ has two a priori characteristics that we want to the smoothing procedure to preserve: increasing monotonicity and $\Lambda_0(0) = 0$. A fast and efficient way of smoothing functional data while enforcing desired constraints has been proposed in He and Ng (1999). The method consists in a minimum absolute deviation estimate of coefficients for a B-spline basis: given a set of observations $\{(x_i, y_i)\}_{i=1,...,m}$ from function $y(x)$ to be smoothed, a set of knots $\{u_0 = 0, u_1, ..., u_k = \tau\}$ and a fixed polynomial degree d, find $\boldsymbol{a}^* = (a_0^*, ..., a_{k+d-1}^*)^T$ such that

$$\boldsymbol{a}^* = argmin_{\boldsymbol{a}\in\mathbb{R}^{k+d}} \sum_{i=0}^{m} |y_i - \sum_{j=0}^{k+d-1} a_j B_j^{(d)}(x_i)|; \tag{6}$$

$B_0^{(d)}(t), ..., B_{k+d-1}^{(d)}(t)$ is the B-spline basis of degree d on the chosen set of knots. If basis functions of polynomial degree $d = 1, 2$ are used, then monotonicity, convexity and pointwise constraints can be written as linear constraints, and since the quantity to minimize can also be written as a linear objective function the problem can be solved with linear programming techniques, whose efficiency and reliability are ascertained. Using $(0, \widehat{\Lambda}_0(0)), (t_1, \widehat{\Lambda}_0(t_1)), (t_2, \widehat{\Lambda}_0(t_2)), ...)$ as observations, the application of this method provides the smooth desired estimate $\widetilde{\Lambda}_0$.

We then need to reconstruct the realizations of processes $\Lambda_i(t)$ for every patient $i = 1, ..., n$ under the chosen model, since in the following we will treat cumulative hazard functions as functional data. Given the particular formulation of our model for cumulative hazard, we can rewrite it in a form that allows to use directly the smoothed estimate $\widetilde{\Lambda}_0(t)$ instead of an estimate of $\lambda_0(t)$. For $i = 1, ..., n$, we set $0 = t_0^{(i)}$ and let $(t_1^{(i)}, ..., t_{N_i(t)}^{(i)})$ be the jump times for patient i; then

$$\begin{aligned}\Lambda_i(t) &= \int_0^t \lambda_0(s) e^{N_i(s^-)\log\alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i(s)} ds \\ &= \sum_{k=0}^{N_i(t)} \int_{t_k^{(i)}}^{t_{k+1}^{(i)}} \lambda_0(s) e^{k\log\alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i(s)} ds. \end{aligned} \tag{7}$$

Here we consider the case of a covariate vector $\boldsymbol{X}_i^T(t) = (\boldsymbol{X}_i^{d^T}, \boldsymbol{X}_i^{cT})$, $i = 1, ..., n$, where $\boldsymbol{X}_i^d = (X_{i1}(t), ..., X_{in_d}(t))^T$ is a vector of derivable functions, while $\boldsymbol{X}_i^c = (X_{i(n_d+1)}(t), ..., X_{i(n_c+n_d)}(t))^T$ is a vector of stepwise constant functions with discontinuities corresponding to jumps of $N_i(t)$; hence we split also the parameter vector $\boldsymbol{\beta}$ using $\boldsymbol{\beta}_d = (\beta_1, ..., \beta_{n_d})^T$ and $\boldsymbol{\beta}_c = (\beta_{n_d+1}, ..., \beta_{n_d+n_c})^T$, so that $\boldsymbol{\beta} = (\boldsymbol{\beta}_d^T, \boldsymbol{\beta}_c^T)^T$. Defining $P_{\boldsymbol{X}}(t) = \int_0^t \lambda_0 e^{\boldsymbol{\beta_d}^T \boldsymbol{X}_i^d(s)} ds$ and integrating by parts we obtain

$$P_{\boldsymbol{X}}(t) = \Lambda_0(t) e^{\boldsymbol{\beta}_d^T \boldsymbol{X}_i^d(s)} - \int_0^t \Lambda_0(s) \boldsymbol{\beta}_d^T [\boldsymbol{X}_i^d(s)]' e^{\boldsymbol{\beta}_d^T \boldsymbol{X}_i^d(s)} ds, \tag{8}$$

where $[\boldsymbol{X}_i^d(s)]' = \left(\frac{dX_{i1}(s)}{ds}, ..., \frac{dX_{in_d}(s)}{ds}\right)^T$.

Plugging $P_{\boldsymbol{X}}(t)$ into (7) leads to the expression

$$\Lambda_i(t) = \sum_{k=0}^{N_i(t)} e^{k\log\alpha + \beta_c^T \boldsymbol{X}_i^c(t_k^{(i)})} \Big[P_{\boldsymbol{X}}(t_{k+1}^{(i)}) - P_{\boldsymbol{X}}(t_k^{(i)}) \Big]. \tag{9}$$

This form allows us to perform only one integration to obtain (8), which is computed substituting $\Lambda_0(t)$ with its smoothed estimate $\widetilde{\Lambda}_0(t)$, and to reconstruct the realizations $\widetilde{\Lambda}_i(t)$ by adding process jumps information.

As a validation of the coherence of the employed model, it is possible to perform a comparison of the average functions of counting and cumulative hazard processes: taking conditional expectation in (2) we notice that $\mathbb{E}[\Lambda_i(t)|\boldsymbol{X}_i(t)] = \mathbb{E}[N_i(t)|\boldsymbol{X}_i(t)]$, for $i = 1, ..., n$. The comparison is not straightforward when curves have different censoring times; in particular, if faster growing curves have higher probability of earlier censoring (this is common for risk curves, as frailer patients die earlier), the naive pointwise sample mean is not monotone and it underestimates expected values for large times. Let $(f(t))_{t \in I}$ be a stochastic process and let τ be a stopping time for this process; then if $\mathbf{f} = (f_1, ..., f_n)^T$ is a set of trajectories of the process f, and $\{\tau_1, ..., \tau_n\}$ a set of realizations of τ, we can define the pointwise sample mean function as

$$\mu_n[\mathbf{f}](t) = \frac{1}{n(t)} \sum_{i=1}^{n} f_i(t) C_i(t), \qquad \forall t \in [0, \tau], \tag{10}$$

where $n(t) = \sum_{i=1}^n C_i(t)$, being $C_i(t) = I(t \le \tau_i)$ the censoring process for subject i and $\tau = max_{i=1,...,n}\tau_i$. Instead of using this estimator, we will use the following

$$\tilde{\mu}_n[\mathbf{f}](x_k) = \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{C_i(x_j)}{n(x_j)} \Big[f_i(x_j) - f_i(x_{j-1}) \Big], \qquad k = 1, ..., m, \tag{11}$$

with $x_j \in \{x_0, ..., x_m\}$, a given set of time points which include $\tau_1, ..., \tau_n$, and $\tilde{\mu}[\mathbf{f}]_n(x_0) = \sum_{i=1}^n f_i(x_0) C_i(x_0) = \sum_{i=1}^n f_i(x_0)$, since we are considering only right censored processes. This estimator, applied to $\{\widetilde{\Lambda}_i\}_{i=1,...,n}$ and $\{N_i\}_{i=1,...,n}$ enforces monotonicity by definition if all sample curves are monotone; moreover, as pointed out in Crowell (1992), this estimator is unbiased and consistent, and in the case of highly positively correlated increments it is likely that $\mathbb{V}ar\{\tilde{\mu}_n[\mathbf{f}](t)\} < \mathbb{V}ar\{\mu_n[\mathbf{f}](t)\}$.

2.3 Functional principal component analysis

A common strategy to deal with complex or high-dimensional data is to perform a dimensional reduction. In the case of functional data, this can be done by representing data on a functional basis, and choosing only relevant components of the expansion.

Consider a functional ANOVA decomposition of data, as suggested in Müller and Stadtmüller (2005)

$$\widetilde{\Lambda}_i(t) = \mu(t) + D_i(t) + \varepsilon_i(t), \qquad i = 1, ..., n \tag{12}$$

where $\mu(t) = \mathbb{E}[\widetilde{\Lambda}(t)]$, $D_i(t)$ is a residual for subject i and $\varepsilon_i(t)$ a noise term. One of the possibilities for representing $\widetilde{\Lambda}_i(t)$ is to use Karhunen-Loève decomposition, which states that functional principal components of a set of functions defined on domain T form a complete orthonormal basis of $L^2(T)$ (see Ferraty and Vieu (2006) for some theoretical results and Ramsay and Silverman (2005) for details on the implementation of functional principal components analysis, briefly FPCA). At this point we will assume that functional data are known on a common support T, thus enabling us to estimate a common Karhunen-Loève basis.

Given the covariance operator

$$G(t,s) = \mathbb{E}\left[\left\{\widetilde{\Lambda}(t) - \mathbb{E}\big[\widetilde{\Lambda}(t)\big]\right\}\left\{\widetilde{\Lambda}(s) - \mathbb{E}\big[\widetilde{\Lambda}(s)\big]\right\}\right] \quad \text{for } (t,s) \in I \times I,$$

the eigenvalue problem to be solved in order to obtain principal components is to find the couples $\{(\psi_k, \nu_k)\}_{i\in\mathbb{N}}$ such that

$$\int_T G(t,s)\psi_k(s)ds = \nu_k \psi_k(t). \tag{13}$$

Once eigenfunctions $\{\psi_k\}_{k\in\mathbb{N}}$ and eigenvalues $\{\nu_k\}_{k\in\mathbb{N}}$ have been found, we can express the functional ANOVA decomposition (12) through the following representation

$$\widetilde{\Lambda}_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t) + \varepsilon_i(t), \qquad i = 1, ..., n,$$

where $\xi_{ik} = \int_T D_i(s)\psi_k(s)ds$ is the k-th score for subject i.

Eigenfunction-eigenvalue couples $\{(\psi_k, \nu_k)\}_{k\in\mathbb{N}}$ completely explain modes of variation in the data, in the sense that eigenfunctions represent orthonormal directions of decreasing variability with respect to the explained variances expressed by corresponding eigenvalues. Thanks to the basis expansion given by principal components, it is possible to represent data using just the first K elements of $\{\psi_k\}_{k\in\mathbb{N}}$, the linear combination of which will naturally be a good approximation for the original curves. The interpretation of eigenvalues as variances is useful also to determine a criterion of choice for most relevant modes. Since $\sum_{k=1}^{K} \nu_i$ represents variance captured by the first K components, we can choose K so that the proportion of variance described by these components is higher than a threshold c, i.e.

$$\frac{\sum_{k=1}^{K} \nu_k}{\sum_{k=1}^{m} \nu_k} \geq c,$$

where m is the number of abscissa values on which functional data are known, which is an upper bound to the number of components that can be estimated.

We then use the following approximation

$$\widetilde{\Lambda}_i^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\psi_k(t) + \varepsilon_i(t), \qquad i = 1, ..., n.$$

For the sake of notation simplicity, from now on we will write $\widetilde{\Lambda}_i(t)$ even when its truncated basis expansion $\widetilde{\Lambda}_i^K(t)$ will be used.

2.4 Generalized functional linear models

Dimensional reduction allows to catch the characterizing features of functional data and to describe them with few variables, i.e. the principal components scores, which can be used as explanatory variables in subsequent model components. The methodology described in the following consists in formulating a generalized functional linear model that can be interpreted as a classical GLM in which FPCA scores and other time independent variables are exploited as covariates.

Let us consider a response variable Y such that $Y_i \sim EF(\theta_i, \eta)$, i.e. Y_i for $i \in 1, ..., n$ belongs to the exponential family

$$\begin{aligned}
&f_{Y_i}(y|\theta_i, \eta) = \exp\left(\frac{y\theta_i - b(\theta_i)}{\eta} + c(y, \eta)\right)\\
&\mathbb{E}[Y_i] = b'(\theta_i)\\
&\mathbb{V}ar[Y_i] = \eta b''(\theta_i)
\end{aligned}$$

with b and c given functions; the *link function* g is s.t. $\mathbb{E}[Y_i] = g^{-1}(\theta_i)$ (i.e. $g^{-1} = b'$). The dependence on observable functions and variables is assumed to be linear and is given by

$$\begin{aligned}
\theta_i &= \int_T D_i(t)\delta(t)dt + \mathbf{z}_i^T\boldsymbol{\gamma}\\
&\approx \int_T \delta(t)\sum_{k=1}^{K}\zeta_{ik}\psi_k(t)dt + \mathbf{z}_i^T\boldsymbol{\gamma}.
\end{aligned}$$

where $\delta : T \mapsto \mathbb{R}$ is a functional parameter, γ a vector of time-independent parameters to be estimated and $\mathbf{z}$ is a vector of time-independent covariates. Notice that we used the K most relevant principal components to represent $D_i(t)$. If $\delta(\cdot)$ is also represented with respect to the principal components basis, i.e. $\delta(t) = \sum_{j=1}^{K}\delta_j\psi_j(t)$, for the orthonormality of $\{\psi_k\}_{k\in\mathbb{N}}$ we obtain

$$\theta_i = \sum_{k=1}^{K}\zeta_{ik}\delta_k + \mathbf{z}_i^T\boldsymbol{\gamma}.$$

We notice that in this formulation the first K FPCA scores can be used to summarize the features of hazard functions with a finite dimensional vector, thus providing a powerful methodology to use functional data in many different classical models for multivariate data. Thanks to this formulation, we reduce the functional estimation problem to the multivariate estimation of parameter vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_K)^T$.

3 Application in telemonitoring data analysis and results

In Lombardia region an experimentation of heart failure telemonitoring was started in 2003, involving 34 health care institutions (see Nuove Reti Sanitarie, `http://ftp.cefriel.it/nrs/` for an overview of program and protocols). Four studies (*Criteria*, *Piano Urbano*, *Nuove Reti Sanitarie* and *Telemaco*) were devoted to collect, under prior informed consent, information about telemonitoring periods, then gathered in a comprehensive database. Each record of this database refers to a telemonitoring period, and contains anagraphic data of the involved patient, number of transmitted electrocardiograms, NYHA class (describing the extent of heart disease), number of occurred hospitalizations and other relevant clinical quantities.

The enrollment protocol adopted during the period 2004–2008 includes adult citizens of Lombardia with a NYHA class of III or IV who have experienced at least one hospitalization for heart failure during the 6 months preceding the beginning of telemonitoring. The telemonitoring period is planned for a 180 days duration, with possible re-enrollment under particular conditions. It period may be interrupted, by protocol, if a hospitalization lasting more than 8 days occurs, or because of the need of an intervention; however, other "external" events may force interruption, such as a change of residence, or the decision by the patient herself/himself to stop the therapy (*drop-out*), or decease.

Since data regarding telemonitoring periods are not sufficient to operate an observational study about effectiveness of this care strategy, we requested an interrogation of regional administrative databases, to obtain hospital discharge data (SDOs) stored during the five years of interest. Each one of these records contains extensive information about a single hospitalization, such as date, duration, DRG[2], drugs received and other data of clinical and economic interest. Each subject contained in the telemonitoring database has been identified by a code, derived from an anonymizing procedure applied to his/her identity number, and used to retrieve from the SDO database the hospitalization histories of these patients in period 2004–2008. The crossing and matching of information between this two databases resulted in the constitution of an initial sample of 1081 patients.

[2] Diagnosis Related Group is a system for the classification of patients discharged from hospital, based on the type of resources used during the stay.

Since we decided to use the period before telemonitoring to predict telemonitoring outcome, a new dataset was built including hospitalizations happened between 1st January 2004 and the start date of the telemonitoring period. Moreover, a further selection of 747 patients is considered in the second part of the analysis, to include only subjects whose telemonitoring period started at least in 2006; in this way, a 2 years time window before telemonitoring is available for all of them for predictive tasks.

The risk of hospital readmission is obviously null during each hospital stay, which typically lasts some days (mean hospitalization length $= 17.18 \pm 28.03$). We deal with this issue by removing the hospital stay period from the process time count, and merging consecutive hospitalization periods, which have to be considered as a single one. The time variable is expressed in days passed from 1st January 2004.

The following analyses have been carried out using the statistical software R (R Development Core Team (2009)). For hazard estimation package `gcmrec` (González et al. (2009)) has been used, while package `cobs` (Ng and Maechler (2009)) has been used for constrained smoothing.

3.1 Hazard estimation

The first step of the analysis is the estimation of model (5) for cumulative hazard functions, using the procedure explained in section 2.2.

The beginning of telemonitoring is introduced as a censoring time τ_i, $i = 1, ..., n$, for the hospitalization counting processes, assuming that this event does not influence preceding hospitalizations; this assumption seems reasonable, on the basis of the enrollment protocol for telemonitoring.

Since cumulative hazard processes are intended to provide a synthesis of time dependent variables, subject age is included as covariate $X_i(s)$ in (5), providing the following model for patients $i = 1, ..., n$

$$\Lambda_i(t|\boldsymbol{X}_i) = \int_0^t C_i(s)\lambda_0(s)\alpha^{N_i(s^-)}\exp[\beta X_i^{\text{age}}(s)]ds,$$

with $C_i(s) = I\{s \leq \tau_i\}$.

Estimated baseline cumulative hazard $\widehat{\Lambda}_0(t)$ is represented in Figure 1 (dashed line), while parameter estimates are shown in Table 1. We notice that parameter α, describing the effect of a new event on the risk of future rehospitalizations, is significantly higher than 1, according to a one-sided hypothesis test with null hypothesis $\alpha \leq 1$; this means that a new event represents an increase of rehospitalization risk. Parameter β, related to the age covariate, is surprisingly negative, meaning that the risk of rehospitalization is slightly lower for older patients; this could be explained by the fact that in the old population considered (the subjects' mean age is 67.82 ± 11.19) subjects survived up to a higher age are the less frail ones. In Figure 1 we can also note that the cumulative baseline hazard function $\widehat{\Lambda}_0(t)$ has a convex behavior, describing a gradual increase of instantaneous risk due to the disease, and common to the whole population.

	estimate	std. dev.	p-value
α	1.21	0.00887	$< 2 \cdot 10^{-16}*$
β	-0.00336	0.00172	0.051

Table 1: Results of hazard parameters estimation. p-value $*$ refers to a test with null hypothesis $\alpha \leq 1$. Both tests are carried out with a normal approximation for maximum likelihood estimators.

Since the nonparametric estimator used for $\Lambda_0(t)$ produces a step function, we perform a smoothing of this estimate with the method exposed in section 2.2; for the B-Spline basis, we choose order 2 and 20 equally spaced knots. A comparison between the nonparametric estimate and the B-spline smoothed estimate is shown in Figure 1.

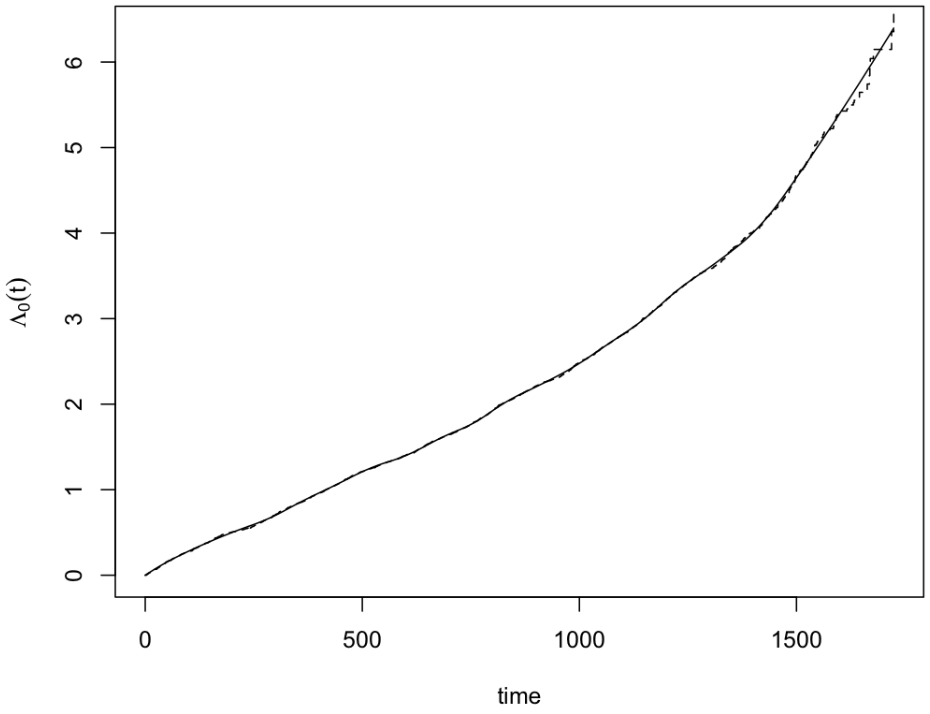


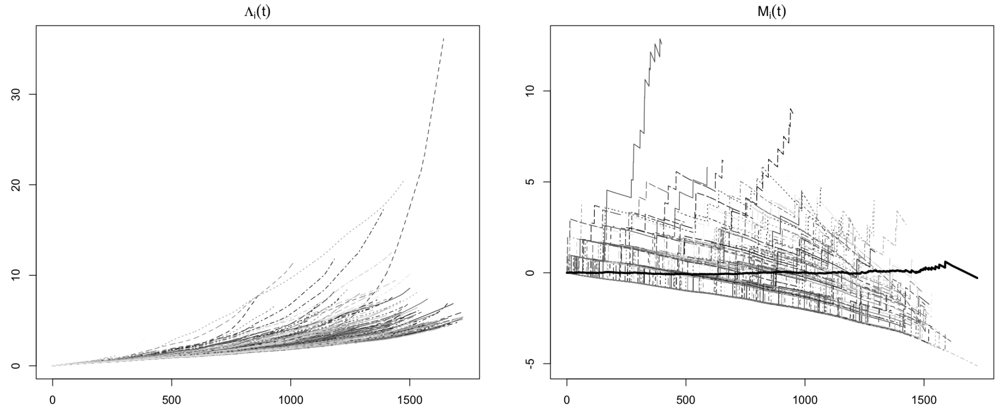
Figure 1: Results of baseline cumulative hazard function estimation: $\widehat{\Lambda}_0$ (dashed line) and its smoothed version $\widetilde{\Lambda}_0$ (solid line).

Once $\widetilde{\Lambda}_0$ has been computed, we can reconstruct individual cumulative hazard processes, letting $\boldsymbol{X}_i(t) = \boldsymbol{X}_i^d(t) = X_i^{age}(t)$, which represents the age of patient i. We can express age as a variable explicitly dependent on time (in days) writing $\boldsymbol{X}_i(t) = a_i + t/365$, where a_i represents the age of patient i at the beginning of the observation period. It is then possible to rewrite (9) as

$$\widetilde{\Lambda}_i(t) = \sum_{k=0}^{N_i(t)} e^{k \log \alpha + \beta a_i} \left[P_{\boldsymbol{X}}(t_{k+1}^{(i)}) - P_{\boldsymbol{X}}(t_k^{(i)}) \right],$$

with $P_{\boldsymbol{X}}(t) = \int_0^t \lambda_0(s) e^{\frac{\beta}{365} s} ds$.

The result of reconstruction of cumulative hazard processes for all the considered patients is shown in Figure 2(a). To verify that condition $\mathbb{E}[\Lambda_i(t)|\boldsymbol{X}_i(t)] = \mathbb{E}[N_i(t)|\boldsymbol{X}_i(t)]$ holds, it is possible to visualize average functions of point processes and cumulative hazard processes, using estimators (11). To address the problem of computing this conditional expectation, we can split the sample in classes of similar initial age $A_{c_1}, A_{c_2}, ...$, and assume that averaging on subjects from the same class produces a good approximation both to $\mathbb{E}[\Lambda_i(t)|\boldsymbol{X}_i(t)]$ and to $\mathbb{E}[N_i(t)|\boldsymbol{X}_i(t)]$. For example, the martingale residuals trajectories and their average for subjects belonging to the age class $A_{60} = \{i : a_i \in (55, 65]\}$ are shown in Figure 2(b); we can see that residuals $\widehat{M}_i(t) = N_i(t) - \widetilde{\Lambda}_i(t)$, $i \in A_{60}$, seem to have the expected behavior.

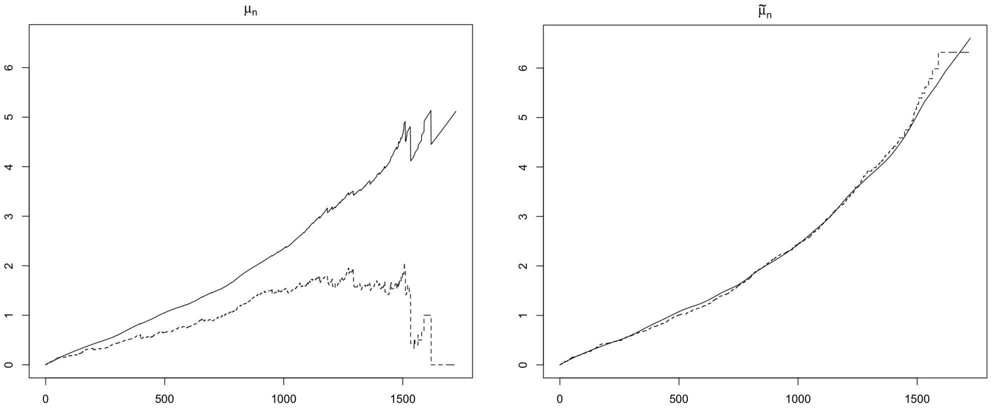


(a) Reconstructed realizations of cumulative hazard processes

(b) Trajectories of residuals $\widehat{M}_i(t) = N_i(t) - \widetilde{\Lambda}_i(t)$, $i \in A_{60}$, and their average (thick line).

Figure 2: Estimated trajectories and martingale residuals.

Figure 3 shows a comparison between average curves computed using pointwise estimator (10) and estimator (11) respectively; in the left panel we notice that the curves estimated with (10) are non monotone and heavily biased due to right censoring, while average curves estimated with (11), depicted in the right panel, suffer from censoring only at the right end of the domain, since available data become fewer with the progression of time.



(a) Curves obtained with the pointwise estimator (10).

(b) Curves obtained with estimator (11).

Figure 3: Average curves of counting process data (dashed lines) and of reconstructed cumulative hazard functions (solid lines) for age class A_{60}.

3.2 Generalized functional linear model estimation

Exploiting the methodology described in sections 2.3 and 2.4, estimated cumulative hazard functions are used to predict telemonitoring outcome, defined as a binary variable with value 1 if telemonitoring has regular conclusion and 0 if the period is terminated by an adverse event, i.e. hospitalization or surgical intervention. A dimensional reduction of functional data is operated via principal component analysis, then FPCA scores and other variables of clinical interest are used as covariates in a logistic regression model.

To avoid the problem of censoring, as previously mentioned, we choose patients for which at least 2 years of clinical history before telemonitoring are available in our records. Moreover, we restrict the time window for our analyses to exactly the 2 years preceding telemonitoring. Doing so, we obtain a dataset of n=747 curves, evaluated on a grid of length m=730 (hazard functions were computed on a vector for abscissa characterized by daily spacing).

Before proceeding to principal component analysis, curves are centered by subtracting their mean function $\tilde{\mu}_n(t)$ (which coincides with estimator $\mu_n(t)$ for the operated subselection of data); moreover, the noise term $\varepsilon_i(t)$ is discarded, since curves have already been estimated with smoothness.

We shall now select the components to consider in the subsequent analysis. A simple and effective criterion consists in choosing the first K components, such that their associated eigenvalues explain a proportion of variance $c > 95\%$. This criterion leads to the choice of the first $K = 2$ components, as detailed in Table 2.

Figure 4 shows in the top panels the 2 relevant functional principal components, and in the bottom panels $\tilde{\mu}_n(t) \pm \nu_k \phi_k(t)$, $k = 1, 2$. The first component is

	ν_1	ν_2
value	777.04	45.64
% variance	94.08	5.53
cum. % variance	94.08	99.60

Table 2: First $K = 2$ eigenvalues obtained with FPCA.

monotone increasing, and is highly dominant in the description of data curves, while the second one is decreasing, characterizing curves that do not grow very fast also on a long time period.

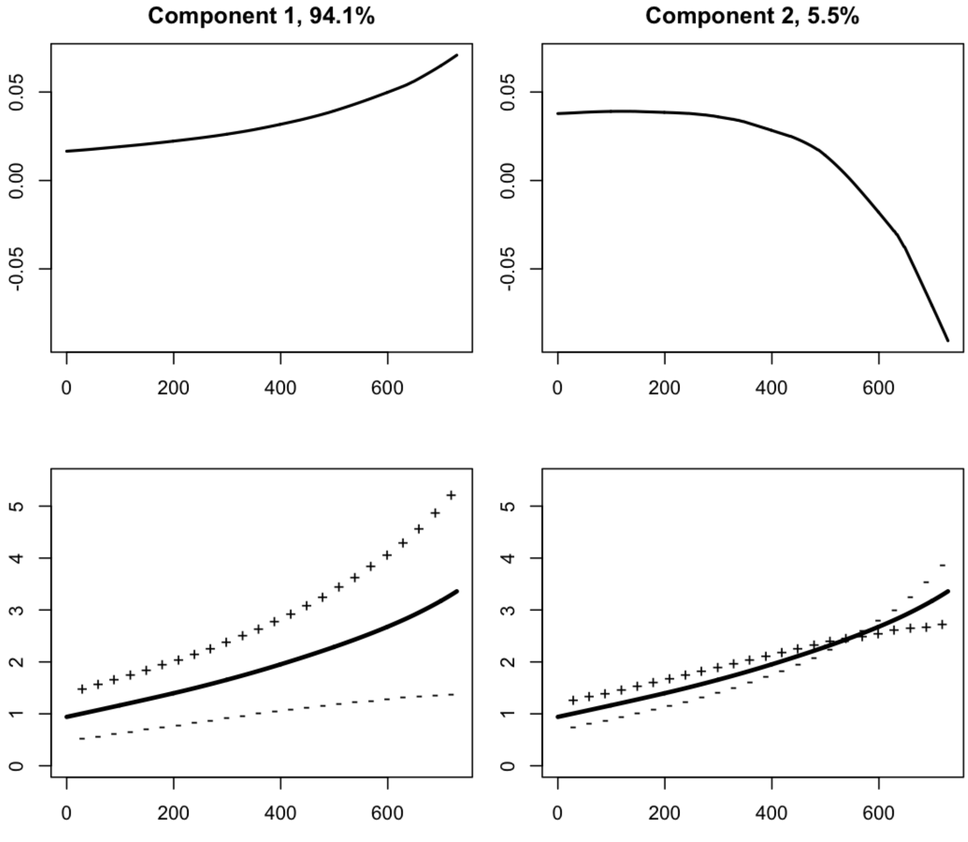


Figure 4: In the upper panels, first $K = 2$ eigenfunctions obtained with FPCA; in the lower panels, representation of $\mu_n(t)$ (solid line) and $\mu_n(t) \pm \nu_k \phi_k(t)$ ('+' or '$-$' respectively), $k = 1, 2$.

The scores of principal components 1 and 2 are then considered as variables that sum up the characteristics of data functions. They are used, together with the categorical variables sex, diagnosis and etiology, to predict telemonitoring outcome; in particular, variables diagnosis and etiology refer to the last hospitalization before telemonitoring. The former has 3 levels (*congestive*, *left* or *unspecified* heart failure), while the latter 5 levels (*hypertensive*, *hyschaemic*, *primary*, *valve*, *other*). We decided to fit a logistic regression model, which can be set in the framework exposed in section 2.4 with link function $g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$, being p_i the probability of normal outcome. Hence, model takes the following form

$$p_i = \frac{exp\Bigg(\sum_{k=1}^{2} \zeta_{ik}\delta_k + \mathbf{z}_i^T \boldsymbol{\gamma}\Bigg)}{1 + exp\Bigg(\sum_{k=1}^{2} \zeta_{ik}\delta_k + \mathbf{z}_i^T \boldsymbol{\gamma}\Bigg)}, \quad \text{for } i = 1, ..., n,$$

where $p_i = \mathbb{E}(Y_i|\mathbf{z}_i, \boldsymbol{\zeta}_{i1}, \boldsymbol{\zeta}_{i2})$ and matrix $\mathbf{Z} = [\mathbf{1}, \mathbf{z}_1, ..., \mathbf{z}_n]^T$ is composed by variables sex, diagnosis and etiology.

The model output of logistic regression is reported in Table 3. Scores 1 and 2 are both significant, and their signs are coherent with a possible interpretation: principal component 1 is an increasing function, so a larger score, which represents a steeper cumulative hazard process, implies a lower probability of regular conclusion; component 2, instead, is decreasing, and its estimated coefficient has opposite sign, indicating that patients who have lower cumulative hazard for longer times have higher probability of normal conclusion of the telemonitoring period. Also, we can notice a slight dependence on etiology; in particular, valvular etiology seems to increase the probability of early conclusion of telemonitoring caused by an adverse event. Instead, there is no significant difference in the probability of adverse events neither among men and women, nor among subjects with different types of diagnoses.

As a measure of goodness of fit we computed the Brier score, which is equal to 0.1614, and the AIC, which is equal to 688.6547.

4 Concluding remarks

In this work a novel approach to the analysis of telemonitoring data has been proposed, aimed at getting the precise insight of information on the patient's health status and clinical history from clinical and Public Health Databases. Database integration, counting process modeling of hospitalizations and generalized functional mixed models are methodologies that can be applied to the study of many different pathologies, thanks to their flexibility and capability of dealing with complex data.

The counting process model is a natural way of representing the occurrence of hospitalizations in time, and enables us to include in the proposed model a large piece of information contained in Public Health Databases to describe

Parameter	Estimate	Std. Error	p-value
γ_0 (Intercept)	14.8108	437.0832	0.9730
γ_1 (Sex)	0.1557	0.2167	0.4726
γ_2 (Etiology - Hypertensive)	0.0669	0.4469	0.8810
γ_3 (Etiology - Hyschaemic)	-0.0187	0.2506	0.9405
γ_4 (Etiology - Primary)	-0.0819	0.3199	0.7981
γ_5 (Etiology - Valve)	-0.8867	0.4673	0.0790
γ_6 (Etiology - Other)	-0.6599	0.3593	0.1248
γ_7 (Diagnosis - Congestive)	-13.8204	437.0832	0.9748
γ_8 (Diagnosis - Left)	-13.1587	437.0832	0.9760
γ_9 (Diagnosis - Unspecified)	-13.6343	437.0833	0.9751
δ_1 (FPCA score 1)	-0.0144	0.0039	0.0003
δ_2 (FPCA score 2)	0.0567	0.020490	0.0056

Table 3: Estimates, standard errors and p-values for parameters of logistic regression.

the clinical history of a patient. The model used is very general and allows to describe complex dynamics in an easily interpretable form.

Although it can seem contradictory to define functional data as ”synthetic”, it is clear that complex, heterogeneous data are easier to study if their effect is resumed with a process that represents their combined effect on instantaneous risk. The obtained trajectories are thus studied in the framework of generalized functional linear models, which offer a powerful tool to analyze dependencies and to perform classification and prediction in a wide range of applications, also in such complex practical contexts as the one considered. The use of FPCA offers the possibility to perform dimensional reduction of functional data, allowing to use well estabilished methods for GLM estimation and inference in a multivariate setting, and borrowing strength from both techniques.

Future improvements include the selection and use of various different time dependent and independent variables to study telemonitoring effectiveness, modifications in quality of life and mortality.

The application of the proposed methodology is a novelty in the study of home telemonitoring, representing the first example of use of information from Public Health Databases to reconstruct the patients’ clinical histories in a synthetic way. This methodology has led to interesting results that could have an impact on the planning of this care strategy. Further development of this framework in cooperation with medical staff could lead to the definition of a useful tool for telemonitoring outcome prediction, which could be used to support long term decisions and to perform health care assessment.

5 Aknowledgements

This work is part of the experimental home telemonitoring program for chronic heart failure “Nuove Reti Sanitarie - PTS”, promoted by Regione Lombardia. The authors wish to thank the staffs from Regione Lombardia, CEFRIEL and the Department of Management, Economics and Industrial Engineering of Politecnico di Milano and dr. Amerigo Giordano.

References

Aalen, O. O., Ø. Borgan, and H. K. Gjessing (2006). *Survival and event history analysis: a process point of view.* Springer-Verlag New York.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag New York.

Capomolla, S., G. Pinna, M. T. La Rovere, R. Maestri, M. Ceresa, M. Ferrari, O. Febo, A. Caporotondi, G. Guazzotti, F. Lenta, S. Baldin, A. Mortara, and F. Cobelli (2004). Heart failure case disease management program: a pilot study of home telemonitoring versus usual care. *European Heart Journal Supplements 6 (Supplement F)*, F91–F98.

CEFRIEL. `http://ftp.cefriel.it/nrs/`. Accessed 15th November 2010.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Crowell, J. I. (1992). Nonparametric estimation of a process mean from censored data. *Statistics & Probability Letters 15*, 253–257.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis.* Springer-Verlag New York.

Giordano, A., S. Scalvini, E. Zanelli, U. Corrá, G. L. Longobardi, V. A. Ricci, P. Baiardi, and F. Glisenti (2008). Multicenter randomised trial on home-based telemanagement to prevent hospital readmission of patients with chronic heart failure. *International Journal of Cardiology 131*(2), 192–199.

González, J. R., E. H. Slate, and E. A. Peña (2009). *gcmrec: General class of models for recurrent event data.* R package version 1.0-3.

He, X. and P. Ng (1999). Cobs: Qualitatively constrained smoothing via linear programming. *Computational Statistics 14*, 315–337.

Kalbfleisch, J. D. and R. L. Prentice (1980). *The Statistical Analysis of Failure Data.* John Wiley & Sons, New York.

Limnios, N. and M. Nikulin (Eds.) (2000). *Recent Advances in Reliability Theory: Methodology, Practice and Inference.* Birkhäuser Verlag AG.

Müller, H.-G. and U. Stadtmüller (2005). Generalized functional linear models. *The Annals of Statistics 33*(2), 774–805.

Ng, P. T. and M. Maechler (2009). *cobs: COBS – Constrained B-splines (Sparse matrix based).* R package version 1.2-0.

Peña, E. A., E. H. Slate, and J. R. González (2007). Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference 137*(6), 1727–1747.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis.* Springer Science+Business Media.

Scalvini, S., E. Zanelli, M. Volterrani, G. Martinelli, D. Baratti, O. Buscaya, P. Baiardi, F. Glisenti, and A. Giordano (2004). A pilot study of nurse-led, home-based telecardiology for patients with chronic heart failure. *Journal of Telemedicine and Telecare 10*, 113–117.

MOX Technical Reports, last issues

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

42/2010 Stefano Baraldo, Francesca Ieva, Anna Maria Paganoni, Valeria Vitelli:
Generalized functional linear models for recurrent events: an application to re-admission processes in heart failure patients

41/2010 Davide Ambrosi, Gianni Arioli, Fabio Nobile, Alfio Quarteroni:
Electromechanical coupling in cardiac dynamics: the active strain approach

40/2010 Carlo D'Angelo, Anna Scotti:
A Mixed Finite Element Method for Darcy Flow in Fractured Porous Media with non-matching Grids

39/2010 Carlo D'Angelo:
Finite Element Approximation of Elliptic Problems with Dirac Measure Terms in Weighted Spaces. Applications to 1D-3D Coupled Problems

38/2010 Nancy Flournoy, Caterina May, Piercesare Secchi:
Response-adaptive designs in clinical trials for targeting the best treatment: an overview

37/2010 Marco Discacciati, Alfio Quarteroni, Samuel Quinodoz:
Numerical approximation of internal discontinuity interface problems

36/2010 Gianni Arioli, Hans Koch:
Non-Symmetric low-index solutions for a symmetric boundary value problem

35/2010 Gianni Arioli, Hans Koch:
Integration of dissipative PDEs: a case study

34/2010 Antonella Abba', Luca Bonaventura:
A mimetic finite difference method for Large Eddy Simulation of incompressible flow

33/2010 Giovanni Migliorati, Alfio Quarteroni:
Multilevel Schwarz Methods for Elliptic Partial Differential Equations