MOX-Report No. 23/2020

Functional modelling of recurrent events on time-to-event processes

Spreafico, M.; Ieva, F.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Functional modelling of recurrent events on time-to-event processes

Marta Spreafico[1,2] marta.spreafico@polimi.it

Francesca Ieva[1,2,3] francesca.ieva@polimi.it

[1]MOX – Department of Mathemathics, Politecnico di Milano, Milan 20133, Italy
[2]CHRP, National Center for Healthcare Research and Pharmacoepidemiology, University of Milano-Bicocca, Milan 20126, Italy
[3]CADS, Center for Analysis Decisions and Society, Human Technopole, Milan 20157, Italy

April 2, 2020

Abstract

In clinical practice many situations can be modelled in the framework of recurrent events. It is often the case where the association between the occurrence of events and time-to-event outcomes is of interest. The purpose of our study is to enrich the information available for modelling survival with relevant dynamic features, properly taking into account their possibly time-varying nature, as well as to provide a new setting for quantifying the association between time-varying processes and time-to-event outcomes. We propose an innovative methodology to model information carried out by time-varying processes by means of functional data. The main novelty we introduce consists in modelling each time-varying variable as the compensator of marked point process the recurrent events are supposed to derive from. By means of Functional Principal Component Analysis (FPCA), a suitable dimensional reduction of these objects is carried out in order to plug them into a survival Cox regression model. We applied our methodology to data retrieved from the administrative databases of Lombardy Region (Italy), related to patients hospitalized for Heart Failure (HF) between 2000 and 2012. We focused on time-varying processes of HF hospitalizations and multiple drugs consumption and we studied how they influence patients' long-term survival. The introduction of this novel way to account for time-varying variables allowed for modelling self-exciting behaviours, for which the occurrence of events in the past increases the probability of a new event, and to make personalized predictions, quantifying the effect of personal behaviours and therapeutic patterns on survival.

Key-words: Functional data analysis; Survival models; Marked point processes; Recurrent event processes; Administrative databases; Heart failure

1 Introduction

In clinical practice many situations can be modelled in the framework of recurrent events, i.e. the repeated occurrence of the same type of events for the same patient over time. Chronic patients are usually involved in long-term therapies, that are often characterized by repeated situations like office visits, subsequent drugs assumption, hospital admissions and many others. Examples include recurrences in breast cancer,[1] asthma attacks in asthma,[2] episodic relapses of follicular lymphoma,[1] readmission after colorectal cancer,[3] epileptic seizures in epilepsy[4] or re-hospitalizations in Heart Failure (HF),[5–7] which will be the context considered for the present work. Heart Failure (HF) is a major and growing public health issue, characterized by high costs, steep morbidity and mortality rates.[8] Despite the advances in the understanding the pathophysiology of chronic HF and the improvement of therapy, HF mortality and morbidity rates remain high.[9,10] However, different studies proved that a proper and monitored drug intake in HF patients could improve their clinical status, functional capacity and quality of life, prevent hospital admission and reduce mortality.[11] HF patients are usually in a polytherapy, i.e. they usually take multiple drugs at the same time. Since models capable of simultaneously treating multiple drugs have not been well developed in pharmacotherapy, it could be interesting to concomitantly analyse more than one medication at the same time. Repeated clinical events of any type constitute the patient's clinical history and carry out information that may be related to patient's health status, disease progression and prognosis. Representing the whole evolution of patients' clinical history through dynamic processes, properly taking into account their time-varying nature, is then the natural and most appropriate way to look at these events. Moreover, it is often the case where the association between the occurrence of clinical events and time-to-event outcomes (e.g. time to treatment failure or death) is of interest. Interests lie both in the dynamics of time-varying recurrent processes themselves and in the final outcome. Studying their relationship could also offer new insights into the direction of personalised treatment, representing a challenging task both for clinical and statistical research.

In biostatistical, epidemiological and medical literature, several approaches to analyse recurrent event data have been proposed.[12–14] Different methods differ in the assumptions and in the interpretation of the results, but they all take into account the correlation between repeated events regarding the same individual. The most frequently applied method for recurrent event data is the model by Andersen-Gill (AG),[15] which is an extension of the Cox proportional-hazard model.[16] The AG model introduces the counting process formulation in terms of increments in the number of events along time. It assumes that the correlation between event times for an individual can be explained by past events, which share a common baseline hazard. In this way, the dependence could be captured by appropriate specification of time-varying covariates which are functions of the realisation of past events, such as the number of previous occurrences. This model is usually indicated for analysing data when correlations among events for each individual are induced by measured covariate and the interest lies in the overall effect on the intensity of the occurrence of the event.[12] Another approach is the Prentice-Williams-Peterson (PWP) model,[17] which incorporates the order of events. The PWP model[17] analyses multiple events ordered by stratification, based on the prior number of events during the follow-up period. It can incorporate both overall and event-specific effects for each covariate. However, it can give unreliable estimates for higher order of events.[12] As a further alternative, Cox's model can be extended using frailty models,[18–21] in which a random covariate that induces dependence among the event times is introduced. This approach assumes that recurrent event times are independent conditional on the covariates and random effects, and it is used to model individual patients' heterogeneity in the baseline hazards. In addition, models able to connect several event processes (recurrent and

fatal/non-fatal ones) have been proposed.[13] Among others, multi-state models[21,22] can be used to model several event time processes and are fully characterised through estimation of transition probabilities between states. Therefore, the choice of the proper approach for the analysis of recurrent event data will be determined by many factors, including among others, number of the events, relationship between subsequent events and biological processes.[12]

Aforementioned methods are used to analyse single or several event processes, possibly connecting them to another event of interest. However, none of these approaches has been used to extrapolate information from repeated events in the form of dynamic functional covariates, and then study how these covariates affect other specific events, such as patient's death. Information related to the dynamic history of patient's recurrent events could be obtained exploiting various sources, including administrative databases. In fact, in recent years the use of computers, mobile devices, wearables and other biosensors allowed to gather and store huge amounts of health-related data, which are collected in administrative healthcare databases. Administrative data address 'operational' goals, since they are collected mainly for managerial and economic purposes, but are increasingly used also for clinical and epidemiological purposes.[23] In particular, they allow to reconstruct patients' clinical history, leading to a new kind of epidemiological research based on real-time availability and low-cost data,[24] often referred as "Real-World Data". However, the validity of using administrative databases is critically dependent on the reliability of the data, the accuracy of disease coding in the administrative records and the possible occurrence of mismatches or incoherences during linkage strategies,[24–26] Nevertheless, in the last decade significant improvements have been gained in administrative databases, and their use in clinical biostatistics has become an accepted practice, representing a great challenge for statistics and related modelling.[24]

For all these reasons, within this study we aim to analyse the impact of re-hospitalizations, which usually herald a substantial worsening of the long-term prognosis, and subsequent consumption of different drugs regarded as time-varying covariates on the time to death of the patients, that is the time-to-event outcome of interest. Data are retrieved from the administrative databases of *Regione Lombardia - Healthcare Division*,[27] within the research project described in Mazzali et al.[28] A key characteristic of HF is that it is a pathology that alternates phases of stability to sudden worsening of the patient's condition. For this reason, it is not possible to assume a stationary pattern for critical events, as also underlined by Baraldo et al.[6] Our idea is to look at time-varying recurrent events for a set of individuals as particular non-stationary stochastic counting processes which can depend on their past marks, i.e., *marked point processes*. Exploiting aforementioned models for recurrent events, we take into account many aspects that influence the event risk and we compute the realized trajectories of the cumulative hazard functions underlying the event processes. Cumulative hazard functions, also called *compensators* of the stochastic processes, can be then though as positive non-decreasing L^2-functions over the temporal domain. Therefore, we can make use of Functional Data Analysis[29] to extract and summarize their main features with the aim of enriching the information available for modelling survival with relevant dynamic features, as well as to provide a new setting for quantifying the association between time-varying processes and patients' long-term survival.

In this work we then propose a new methodology able to effectively extract and resume information from (possibly multivariate) functional data, intended as trajectories of compensators representing recurrent events, which could influence the time-to-event outcome of interest, plugging them into a suitable functional Cox's regression model.[30] Our procedure can be divided into two phases, the first concerning the representation of time-varying covariates and the second related to the modelling of such covariates in a time-to-event framework. Firstly, patients'

clinical history data retrieved from administrative database are used to model the cumulative hazard functions related to the processes of HF hospitalization and drugs purchase as a proxy of drugs intake. In particular, among the aforementioned methods to deal with recurrent events, we model the cumulative hazard functions through AG Cox models for counting processes.[15] The main novelty we introduce in this work consists in modelling each longitudinal process as the compensator of a marked point process the recurrent events are supposed to derive from. In this first part we end up with time-varying covariates representing the dynamic evolution of the events risk, which can be though as positive non-decreasing L^2-functions over the temporal domain and are therefore smoothed accordingly. The second phase is motivated by the purpose of enriching the information available for modelling survival with these dynamic features. Therefore, we apply Functional Principal Component Analysis (FPCA)[29] in order to perform a dimensionality reduction and summarise information emerging from the functional compensators to a finite set of covariates, while loosing a minimum part of the information. Scores resulting from FPCA are then included into a predictive functional Cox's model for long-term survival, adjusting for patients' baseline characteristics. In doing so, we also provide a new setting for quantifying the association between time-varying processes and patients' long-term survival.

The remaining part of the paper is organized as follows. Section 2 describes the real administrative HF database analysed within this work. Section 3 presents the methodology, with a detailed description of the reconstruction of compensators related to suitable marked point processes for recurrent events in Section 3.1. Section 4 reports the application of the methodology proposed in Section 3 to HF administrative database. Finally, Section 5 contains some concluding remarks, discussion of strengths and limitations of the proposed approach and opportunities for future works. All the analyses are carried out using the software `R`.[31]

2 Dataset

In this section we describe the real case-study analysed within this work. In Section 2.1 we introduce the administrative data sources. In Section 2.2 we describe the study design and the outcome measure.

2.1 Administrative data sources

The project database was built for non-paediatric (age $\geq$ 18 years) patients living in Lombardy (one of the biggest and most populated Italian region, accounting for 10 million residents) hospitalized with a principal diagnostic code of HF between January 2000 and December 2012.[28] Data were provided by *Regione Lombardia - Healthcare Division*,[27] within the research project described in Mazzali et al.[28] Enrolment occurred from the date of discharge of the first HF hospitalization (i.e. the index date). In order to protect privacy, information retrieved from the different databases were linked via a single anonymous ID (identification) code. Details regarding data extraction and selection are discussed in Mazzali et al.[28]

Patients' clinical history of hospitalizations or drugs assumption could be reconstructed using data related to i) patient admission to hospital (Hospital Discharge Charts - HSC), which contain data related to hospital admissions and time to death (or administrative censoring), ii) pharmaceutical purchases, which provide information on the number and times of drug purchases. Since data on drugs prescriptions were nor publicly available neither accessible, the approximation of drug consumption with drug purchase was the only viable option. Examples and limitations of using this approach into a pharmacoepidemiological setting are discussed by Spreafico et al.[32]

Each record in the dataset was therefore related to an hospitalization or a drug purchase of a given patient. With regard to ordinary hospital admission, the date of discharge from hospital and the length of stay in hospital were retrieved. For drug purchases, identified by their Anatomical Therapeutic Chemical (ATC) codes (WHO Collaborating Centre for Drug Statistics Methodology website: `https://www.whocc.no.`), the date of purchase and the number of days of treatment covered by the prescription, based on the number of boxes and the Defined Daily Dose (DDD)[33] for that specific medicinal product, were retrieved. Among the disease-modifying drugs for HF patients,[11,34,35] we focused on Angiotensin-Converting Enzyme (ACE) inhibitors, Beta-Blocking (BB) agents and Anti-Aldosterone (AA) agents.

2.2 Study design and outcome measure

In this work we focused on a representative sample of the administrative database of Lombardy Region related to patients with their first HF discharge between January 2006 to December 2012, excluding subjects who died during the index hospitalization. A 5-years *pre-study period* from 2000 to 2005 (Figure 1) was used in order to consider only "incident" HF patients, i.e. patients with no contacts with healthcare system in the previous five years due to HF. The study-period started from the first discharge for HF (index time T_{start} in Figure 1) and was divided into the *observation period* (365 days from the index discharge date), used for the compensators reconstruction, and the *follow-up period*, used for the survival analysis, whose starting time was $T_0 = T_{start} + 365$. The modelling of the compensators related to the stochastic processes of interest regarded the time interval $[T_{start}; T_0]$ in Figure 1. Therefore, only patients alive at the end of the *observation period* were selected in the study cohort and followed up to observe survival outcomes. We underline that this choice, necessary for the reconstruction of compensator trajectories, could imply a survival bias in case of the exclusion of too many early dying patients (that is not our case since only 6.8% of patients died during the observation period).

Study outcome of interest was patient's death for any cause. Deaths were collected from the Hospital Discharge Forms Database (for in-hospital deaths) or Vital Statistics Regional Dataset (for out-hospital deaths). For the survival analysis, each patient was followed from the end of the observation period until the end of the study or the date of death (see Figure 1). Hence, the long term survival of each patient was measured on the time interval $[T_0; T_{end}]$. The administrative censoring date was December 31^{st}, 2012.

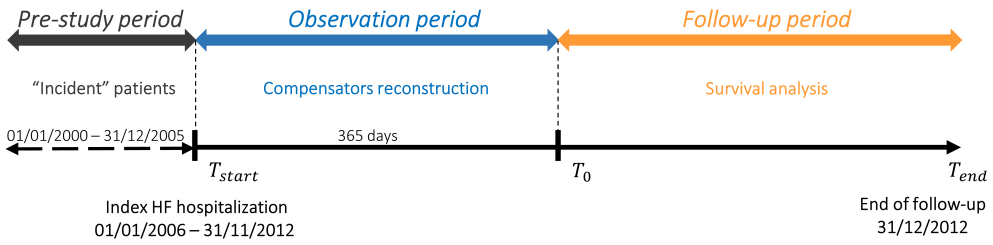

Figure 1: Study design for a HF patient of the study cohort. The *pre-study period* is used to defined "incident" HF patients. The *observation period* is used for the selection of patient's clinical history and the compensators reconstruction. The *follow-up period* is used for survival analysis. T_{start} is the time instant the patient is discharged by her/his first hospitalization and enrolled into the current study. $T_0 = T_{start} + 365$ is the starting time of the follow-up. T_{end} is the minimum between the death or the administrative censoring (December 31^{st}, 2012).

3 Statistical Methodologies

We now introduce and describe the methodology of our work declined on the case study of interest. In Section 3.1 we focus on the main novelty introduced by the present work, i.e., the idea of representing the compensators of suitable marked point processes for recurrent events as functional covariates possibly affecting the outcome process of interest. In Section 3.2 we resume the entire procedure of our analysis.

3.1 Marked point process formulation for recurrent events

Let's consider a set $\mathcal{H}$ recurrent events for a set of n individuals as stochastic processes. Let's use *marked point process for recurrent events*,[36] where a *jump mark* $m_{i,j}^{(h)}$ is associated to each *jump time* $t_{i,j}^{(h)}$, where $h \in \mathcal{H}$, $i \in \{1, ..., n\}$ and $j \in \left\{0, 1, ..., n_i^{(h)}\right\}$ are respectively the process, the subject and the jump indices and $n_i^{(h)}$ is the total number of events of type h experienced by the i-th subject. The observations (possibly censored) may be considered as the realisation of $N_1^{(h)}, ..., N_n^{(h)}$ processes, where $N_i^{(h)}$ is the stochastic process which counts the observed events (or jumps) of the process h in the life of the i-th individual (in our case during the observation period). According to the Doob-Meyer (D-M) decomposition theorem,[37] each counting process $N_i^{(h)}(t)$ can be seen as:

$$N_i^{(h)}(t) = M_i^{(h)}(t) + \Lambda_i^{(h)}(t) = M_i^{(h)}(t) + \int_0^t \lambda_i^{(h)}(s)ds \tag{1}$$

where $M_i^{(h)}(t)$ is a zero-mean uniformly integrable martingale which represents the residual of the process, and $\Lambda_i^{(h)}(t) = \int_0^t \lambda_i^{(h)}(s)ds$ is a unique predictable, non-decreasing, *cadlag* (right-continuous with left limits) and integrable process, i.e. the *compensator* (or *cumulative hazard*). This compensator may be thought as a positive non-decreasing L^2-function over the temporal domain and will be the core of our modelling effort.

A counting process where jumps may have different size can be modelled as a point process, assuming that a given distribution regulates the size of the jumps. A marked point process is then the couple of processes describing the behaviour of jumps and marks, and it is usually modelled through the *conditional intensity function* $\lambda^{(h)}\left(t, \mathbf{m}^{(h)} \middle| \mathcal{F}_t^{(h)}\right)$, i.e. the expected rate of event h at time t with marks $\mathbf{m}^{(h)}$:

$$\lambda^{(h)}\left(t, \mathbf{m}^{(h)} \middle| \mathcal{F}_t^{(h)}\right) = \lambda_g^{(h)}\left(t \middle| \mathcal{F}_t^{(h)}\right) f^{(h)}\left(\mathbf{m}^{(h)} \middle| \mathcal{F}_t^{(h)}\right) \tag{2}$$

where h is the process of interest, $\mathcal{F}_t^{(h)}$ is the filtration of the process itself and it is interpreted as the history of realisations of the process itself, $\lambda_g^{(h)}$ is the intensity process of the counting process, also called *ground intensity*, and $f^{(h)}$ is the multivariate density of the marks $\mathbf{m}^{(h)}$. Using this formulation, conditional independence of jump times and marks is assumed.

To handle recurrent events and allow predictors to change over time, we use the counting process formulation of the Cox model for recurrent events introduced by Andersen and Gill[15] and we assume a particular distribution for the marks in order to ease computations. Under these hypotheses, for each event h the conditional intensity function $\lambda_i^{(h)}(t)$ of patient i in Equation

(2) takes the form:

$$\begin{aligned}\lambda_i^{(h)}(t) &= Y_i^{(h)}(t)\lambda_0^{(h)}(t)\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(t)\right\}\exp\left\{\boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(t)\right\} \\ &= Y_i^{(h)}(t)\lambda_0^{(h)}(t)\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(t)+\boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(t)\right\}\end{aligned} \tag{3}$$

where $\mathbf{x}_i^{(h)}(t)$ is the possibly time-dependent vector of covariates of the i-th individual, $\mathbf{z}_i^{(h)}(t)$ is the time-dependent vector of covariates related to the marks $\mathbf{m}_i^{(h)}$ of the i-th individual, $\boldsymbol{\beta}^{(h)}$ and $\boldsymbol{\gamma}^{(h)}$ are fixed vectors of coefficients, $\lambda_0^{(h)}$ is the underlying hazard function shared across patients, and $Y_i^{(h)}$ is a predictable process taking values in $\{0,1\}$. Whenever $Y_i^{(h)}=1$, the i-th individual is under observations, i.e. $Y_i^{(h)}$ takes the role of the censoring variable.
The estimation of the parameters $\boldsymbol{\beta}^{(h)}$ and $\boldsymbol{\gamma}^{(h)}$ is based on a partial likelihood function,[16] and maximised by applying the Newton-Raphson iterative procedure.[38] $\forall h \in \mathcal{H}$ the baseline cumulative hazard $\Lambda_0^{(h)}(t)=\int_0^t \lambda_0^{(h)}(s)ds$ can be estimated using the *Breslow estimator*[39] $\hat{\Lambda}_0^{(h)}(t)$, which returns step-function. However, since true underlying functions $\Lambda_0^{(h)}(t)$ are absolutely continuous, we smooth the estimates using the approach adopted in Baraldo et al,[6] obtaining regularised version of $\Lambda_0^{(h)}(t)$, namely $\tilde{\Lambda}_0^{(h)}(t)$.

Let's now consider $t_{i,0}^{(h)} < t_{i,1}^{(h)} < ... < t_{i,N_i^{(h)}(\tau)}^{(h)}$ the realised jump times of process $N_i^{(h)}(t)$, with τ equal to the censoring time (possibly equal for all individuals or not), $t_{i,0}^{(h)}=0\ \forall h,i$ and $n_i^{(h)}=N_i^{(h)}(\tau)\,\forall h,i$. In our case, τ is the censoring time of the observation period, i.e. T_0 in Figure 1. We can express the realisations of each compensator $\Lambda_i^{(h)}(t)$ for the process h of the i-th patient as a function of $\Lambda_0^{(h)}(t)$, $\boldsymbol{\beta}^{(h)}$ and $\boldsymbol{\gamma}^{(h)}$:

$$\begin{aligned}\Lambda_i^{(h)}(t) &= \int_0^t \lambda_i^{(h)}(s)ds = \int_0^t Y_i^{(h)}(s)\lambda_0^{(h)}(s)\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(s)+\boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(s)\right\}ds \\ &= \sum_{j=1}^{N_i^{(h)}(t)}\int_{t_{i,j-1}^{(h)}}^{min\left(t_{i,j}^{(h)},t\right)}\lambda_0(s)\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(t_{j-1})+\boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(t_{j-1})\right\}ds \\ &= \sum_{j=1}^{N_i^{(h)}(t)}\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(t_{j-1})+\boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(t_{j-1})\right\}\left[\Lambda_0^{(h)}\left(min\left(t_{i,j}^{(h)},t\right)\right)-\Lambda_0^{(h)}\left(t_{i,j-1}^{(h)}\right)\right]\end{aligned} \tag{4}$$

An estimate of the compensator in Equation (4) can be then obtained as:

$$\hat{\Lambda}_i^{(h)}(t)=\sum_{j=1}^{N_i^{(h)}(t)}\exp\left\{\hat{\boldsymbol{\beta}}^{(h)^T}\mathbf{x}_i^{(h)}(t_{j-1})+\hat{\boldsymbol{\gamma}}^{(h)^T}\mathbf{z}_i^{(h)}(t_{j-1})\right\}\left[\tilde{\Lambda}_0^{(h)}\left(min\left(t_{i,j}^{(h)},t\right)\right)-\tilde{\Lambda}_0^{(h)}\left(t_{i,j-1}^{(h)}\right)\right] \tag{5}$$

where $\hat{\boldsymbol{\beta}}^{(h)}$ and $\hat{\boldsymbol{\gamma}}^{(h)}$ are the estimated vectors of coefficients and $\tilde{\Lambda}_0^{(h)}(t)$ is the smoothed estimate of the cumulative baseline hazard.

To check the fitting of $\hat{\Lambda}_i^{(h)}(t)$, we have to verify whether the estimates of martingale residuals $M_i^{(h)}(t)$ involved in the D-M decomposition (1), i.e. the residuals[40] given by

$$\hat{M}_i^{(h)}(t)=\hat{\Lambda}_i^{(h)}(t)-N_i^{(h)}(t), \tag{6}$$

may be effectively considered as realisations of zero-mean martingales. In order to do so, we can plot the residuals evaluated in the whole observation period and check if the average residual curve $\bar{M}^{(h)}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{M}_i{}^{(h)}(t)$ is approximately close to 0 over time.

This formulation extends the one proposed in Baraldo et al,[6] allowing the counting processes to depend on their marks and setting up a framework for multiple processes to be considered. In fact, applying this procedure $\forall h \in \mathcal{H}$, we end up with a multivariate time-dependent data for each patient, characterizing her/his recurrent events dynamics during the the observation period $T_0 - T_{start}$. These compensator trajectories may be thought as patient-specific time-varying covariates and, mathematically speaking, as positive non-decreasing L^2-functions over the temporal domain $[T_{start}; T_0]$.

3.2 Methodological procedure

The entire procedure may be resumed in four steps. Steps 1 and 2 are devoted to reconstruct the compensators of suitable marked point processes as time-varying (functional) covariates. Steps 3 and 4 set up a suitable framework for including such time-varying covariates in a time-to-event model.

1. **Data preprocessing & clinical history**
   We first select the cohort of patients being part of the analysis, i.e. incident patients survived at least for one year. Then we identify the set $\mathcal{H}$ of longitudinal events of interest to be modelled as marked point processes for recurrent events. In particular, we select only events happened during *observation period*, i.e. the events related to the clinical history of each patient.

2. **Modelling compensators of marked point processes**
   For each event $h \in \mathcal{H}$, we reconstruct the compensator trajectories of the marked point processes for recurrent events through Equation (5), applying the theoretical and mathematical formulation introduced in Section 3.1.

3. **Summarize compensators through FPCA**
   The compensator trajectories computed at Step 2 may be thought as patient-specific time-varying covariates and, mathematically speaking, as positive non-decreasing L^2-functions over the temporal domain $[T_{start}; T_0]$. Therefore, we apply Functional Principal Component Analysis (FPCA)[29] in order to perform a dimensionality reduction and summarise information emerging from the functional compensators to a finite set of covariates (scores) to be plugged into a model for patients' survival.

4. **Predictive survival Cox's model**
   We apply 10-fold cross validation to select the best set of covariates among patients' baseline characteristics and scores resulting from the FPCA on compensators, according to the highest Concordance Index.[41] Finally, we fit a functional Cox regression model[30] in order to quantify the association between time-varying processes and patients' long-term survival.

4 Application and Results

We now proceed with the application of the four steps to the administrative database of Lombardy Region, in order to study how processes like re-hospitalizations and multiple drugs consumption affect long-term survival in HF patients.

4.1 Step 1: Data preprocessing & clinical history

We focused on a representative sample of the administrative database of Lombardy Region related to 4,872 patients with their first HF discharge between January 2006 to December 2012. Excluding patients who died during the *observation period*, a final cohort of $n = 4,541$ (93.2%) patients was selected. Overall, at index hospitalization, mean age of the study cohort was 73.98 years ($s.d. = 11.37$) with a percentage of male patients equal to 54.4% (2,466 patients). The median value of long-term survival was 37.32 (IQR 20.53-54.93) months. At administrative censoring date 1,200 patients (26.4%) were dead and 3,341 (73.6%) were censored.

We identified four types of stochastic processes of interest: hospitalizations due to HF, purchases of ACE, BB and AA drugs, identified by their ATC codes. Hence, the set of recurrent events of interest was $\mathcal{H} = \{h : ACE, BB, AA, HF\,hosp\}$. In particular, we selected only events within the observation period (censoring time $\tau = T_0$). For each patient $i \in \{1, \ldots, n = 4,541\}$, repeated events of process h were modelled as a marked point process $N_i^{(h)}(t)$, with *jump times* $t_{i,j}^{(h)}$ equal to event times (i.e. date of j-th admission in hospital or date of j-th drug purchase) and *jump marks* $m_{i,j}^{(h)}$ equal to the length of stay in hospital or the duration of drug coverage respectively, where $j \in \left\{0, 1, \ldots, N_i^{(h)}(\tau)\right\}$. Figure 2 shows the counting processes $N_i^{(h)}(t)$ describing ACE purchase (top-left panel), BB purchase (top-right panel), AA purchase (bottom-left panel) and HF hospitalization (bottom-right panel) for a sample of 500 HF patients belonging to the administrative database. As expected, they are non-decreasing step functions over the observation period. Overall, at the end of the observation period (time $t = \tau = T_0$), the most frequent events were ACE and BB purchases: 2,916 patients (64.2%) purchased ACE at least once with a median of 7 purchases (IQR = [4;10]), and 2,890 patients (63.6%) purchased BB at least once with a median of 7 purchases (IQR = [4;9]), where the median number of events h at time τ is given by $median_{i \in \{1,\ldots,n\}} N_i^{(h)}(\tau)$. Purchase of AA and hospitalization due to HF were less frequent: 2,007 patients (44.2%) purchased AA at least once with a median of 5 purchases (IQR = [3;7]) and 2,699 patient (59.4%) were re-hospitalized due to HF, with a median of 2 HF hospitalizations (IQR = [1;3]).

In order to proceed with the analyses and model the compensators of the longitudinal processes, we finally had to reformat the administrative data in four different dataset, one for each process h, as explained in Appendix A. Once the four datasets have been prepared, we can now reconstruct the compensators of these longitudinal processes.

4.2 Step 2: Modelling compensators of marked point processes

We can now reconstruct the compensators of the marked point processes for recurrent events, as explained in Section 3.1. For each process $h \in \{ACE, BB, AA, HF\,hosp\}$, we first select the best set of features for the Cox's model of recurrent events using 10-fold cross validation and we estimate the selected coefficients on the whole dataset. Then, we fit and smooth cumulative baseline hazard using the constrained B-spline smoothing algorithm introduced by He and Ng.[42] Finally, we reconstruct the compensator trajectories as functions of the estimated coefficients and of the smoothed estimate of the cumulative baseline hazard using Equation (5).

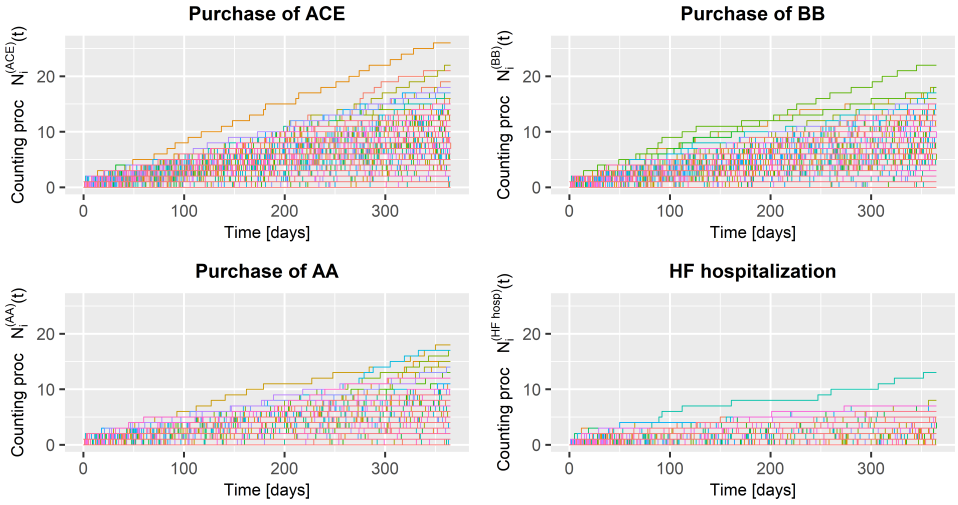

Figure 2: Representation of counting processes $N_i^{(h)}(t)$ related to purchases of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and HF hospitalizations (bottom-right panel) during the observation period for a sample of 500 HF patients belonging to the administrative database. Each non-decreasing step function is related to a different patient.

4.2.1 Features selection and coefficients estimation

For each process $h \in \{ACE, BB, AA, HF\,hosp\}$, we used as covariates $\mathbf{z}_i^{(h)}(t)$ of patient i: the number of events related to process h occurred in the past $Nm_i^{(h)}(t)$ and the sum of the corresponding marks $y_i^{(h)}(t)$. Also the logarithmic transformations (shifted away from 0) of the same variables, i.e., $\log(Nm_i^{(h)}(t)+1)$ and $\log(y_i^{(h)}(t)+1)$, and respective interactions, were considered. Adjustments for *age* and *gender* at baseline were performed. The vector of all the covariates considered for the model is indicated by $\mathbf{x}_i^{(h)}(t)$. In particular, for each process h we performed a 10-fold cross-validation to determine the best sets of features according to the lowest Mean Absolute Martingale Residual (MAMR) (see Appendix B for details regarding MAMR and its estimation). Once covariates were selected, we fitted four Cox models, one for each process h, using the selected features on the entire dataset to estimate coefficients $\hat{\boldsymbol{\beta}}^{(h)}$ and $\hat{\boldsymbol{\gamma}}^{(h)}$.

In Table 1 selected features, hazard ratios and corresponding 95% CI are reported. Among all the models tested through the cross-validation procedure, features related to $Nm^{(h)}(t)$, $y^{(h)}(t)$ and their interaction were selected and their coefficients were always significantly different from 0. Furthermore, the signs of the fitted coefficients relative to these three types of features were consistent throughout the four processes, allowing us to give similar interpretations. On one hand, considering processes related to drug purchases $h \in \{ACE, BB, AA\}$, we observed that the HRs related to the number of past events $\log(Nm^{(h)}(t)+1)$ and to the sum of the past marks $\log(y^{(h)}(t)+1)$ were greater than 1. This could be interpreted as a "self-exciting" behaviour: many drug purchases in the past and the purchase of big quantities of drug both increase the risk of a new purchase. Moreover, HR related to the interaction terms $\log(Nm^{(h)}(t)+1) \times \log(y^{(h)}(t)+1)$ were lower than 1, meaning that the increase in risk is softened in case of several drug purchases and/or a great quantities of drug purchased. For ACE and BB purchases, younger patients were most likely to buy medication than older ones [HRs$<$1], and gender was not selected as

predictive feature. Differently, for AA purchases females were most likely to buy AA agents than males [HR<1] and age variable was not selected through cross-validation. On the other hand, considering HF hospitalization process $h = HF\,hosp$, we observed that the procedure selected the original features $Nm^{(h)}(t)$, $y^{(h)}(t)$ and $Nm^{(h)}(t) \times y^{(h)}(t)$ instead of their logarithmic transformations. This was probably due to the fact that hospitalizations were rarer than drug purchases, so they might have a greater effect in increasing the risk of experiencing a new event. We found that the HRs related to $Nm^{(h)}(t)$ and $y^{(h)}(t)$ were greater than 1, indicating again a "self-exciting" behaviour: being hospitalised often in the past, and having spent long periods of time at the hospital both increase the risk of a new hospitalization. Moreover, HR for the interaction term $Nm^{(h)}(t) \times y^{(h)}(t)$ was lower than 1, meaning that the increase in risk was softened in case of many hospitalizations and/or in the case of a long time spent at the hospital in the past. Finally, we observed that variables related to both age [HR<1] and gender [HR>1] were selected and statistically significant: younger (or male) patients were most likely to be re-hospitalized than older ones (or females).

Process h	**Selected features**	**HR**	**CI (2.5%)**	**CI (97.5%)**
ACE	*age*	0.9967	0.9957	0.9978
	$\log(Nm^{(ACE)}(t)+1)$	4.5216	4.1633	4.9107
	$\log(y^{(ACE)}(t)+1)$	1.1036	1.0872	1.1202
	$\log(Nm^{(ACE)}(t)+1) \times \log(y^{(ACE)}(t)+1)$	0.9141	0.9025	0.9258
BB	*age*	0.9928	0.9917	0.9939
	$\log(Nm^{(BB)}(t)+1)$	5.5360	5.2147	5.8770
	$\log(y^{(BB)}(t)+1)$	1.1340	1.1144	1.1540
	$\log(Nm^{(BB)}(t)+1) \times \log(y^{(BB)}(t)+1)$	0.8283	0.8161	0.8406
AA	*gender (Male)*	0.9435	0.9073	0.9811
	$\log(Nm^{(AA)}(t)+1)$	9.8781	8.6116	11.3310
	$\log(y^{(AA)}(t)+1)$	1.2023	1.1722	1.2332
	$\log(Nm^{(AA)}(t)+1) \times \log(y^{(AA)}(t)+1)$	0.7780	0.7561	0.8005
$HF\ hosp$	*age*	0.9957	0.9934	0.9979
	gender (Male)	1.1510	1.0854	1.2207
	$Nm^{(HF\,hosp)}(t)$	1.4319	1.3809	1.4848
	$y^{(HF\,hosp)}(t)$	1.0083	1.0051	1.0116
	$Nm^{(HF\,hosp)}(t) \times y^{(HF\,hosp)}(t)$	0.9976	0.9968	0.9985

Table 1: Selected features, Hazard Ratios and corresponding 95% CI of the Cox models for recurrent events for the stochastic processes describing the purchase of ACE inhibitors, BB agents, AA agents and the HF hospitalizations.

4.2.2 Fit and smooth cumulative baseline hazard

Once we estimated the coefficients $\hat{\boldsymbol{\beta}}^{(h)}$ and $\hat{\boldsymbol{\gamma}}^{(h)}$ of each Cox model for recurrent events of type h, we computed the estimated cumulative baseline hazards $\hat{\Lambda}_0^{(h)}(t)$ using the Breslow estimator. Since this procedure provide a step function $\left(\hat{\Lambda}_0^{(h)}(t)\right)$, we smoothed them using the constrained B-spline smoothing algorithm introduced by He and Ng[42] with increasing monotone constraints and no roughness penalties. In particular, we used 20 knots for the B-spline basis and we assumed that $\tilde{\Lambda}_0^{(h)}(t_{start}) = 0$.

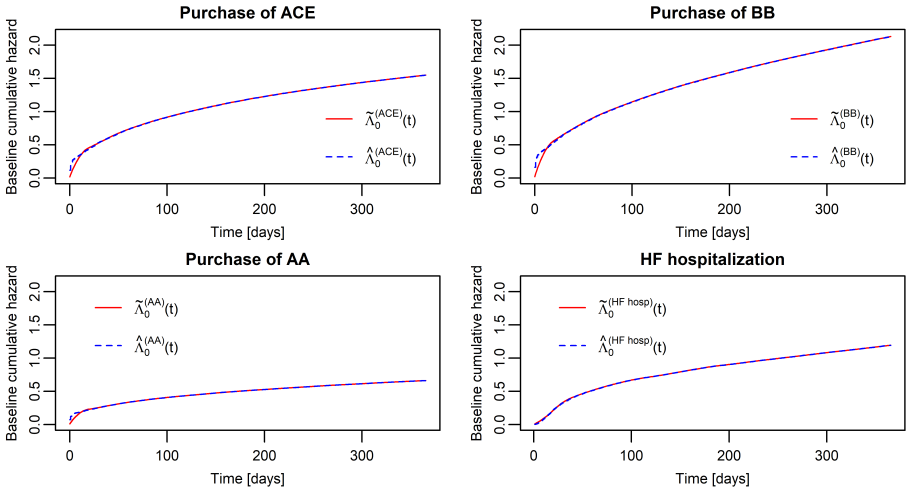

Figure 3: Cumulative baseline hazards of the Cox models for recurrent events describing the stochastic processes of purchases of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and of HF hospitalizations (bottom-right panel), fitted with the Breslow estimator $\hat{\Lambda}_0^{(h)}(t)$ (dashed blue lines) and smoothed $\tilde{\Lambda}_0^{(h)}(t)(t)$ according to the procedure described in Baraldo et al[6] (solid red lines).

Figure 3 shows both the estimates obtained with the Breslow estimator $\hat{\Lambda}_0^{(h)}(t)$ (dashed blue lines) and the corresponding smoothed estimates $\tilde{\Lambda}_0^{(h)}(t)$ (solid red lines) for the four stochastic processes describing ACE purchase (top-left panel), BB purchase (top-right panel), AA purchase (bottom-left panel) and HF hospitalization (bottom-right panel). We observed that $\forall h \in \mathcal{H}$ we obtained monotonically increasing estimates $\tilde{\Lambda}_0^{(h)}(t)$ of the cumulative baseline hazards with $\tilde{\Lambda}_0^{(h)}(t_{start}) = 0$.

4.2.3 Reconstruct compensators

At this point, we could reconstruct the trajectories of the compensators $\hat{\Lambda}_i^{(h)}(t)$ of the four considered stochastic processes for all the patients, exploiting Equation (5). The trajectories of compensators $\hat{\Lambda}_i^{(h)}(t)$ constitute our functional data. Figure 4 shows the compensators of the stochastic processes describing ACE purchase (top-left panel), BB purchase (top-right panel), AA purchase (bottom-left panel) and HF hospitalization (bottom-right panel) of the same sample of 500 HF patients mentioned above. We observed that the trajectories $\hat{\Lambda}_i^{(h)}(t)$ are monotonically non-decreasing and take value 0 at time t_{start}, as did the smoothed baseline cumulative hazards $\tilde{\Lambda}_0^{(h)}(t)$. The large variability of the compensators across different patients reflects the variability of the realizations of their recurrent events.

Finally, we had to check for adequate fitting of the procedure. In order to do so, we controlled if the fitted residuals $\hat{M}_i^{(h)}(t)$ in Equation (6) may be effectively considered as realisations of martingales. For each process of interest, we plotted the residuals evaluated in the whole observation period and we checked graphically that their means $\bar{M}^{(h)}(t)$ were approximately equal to 0. Figure 5 show the fitted residuals $\hat{M}_i^{(h)}(t)$ for each process for the sample of the 500 patients mentioned above (*ACE*: top-left; *BB*: top-right; *AA*: bottom-left; *HF hosp*: bottom-

right). The black line in each panel corresponds to the temporal average residual curve $\bar{M}^{(h)}(t)$, computed using all the $n = 4,541$ patients. From the Figure we observed that the time-varying means were approximately a constant equal to zero for all the considered processes. Hence, we might conclude that we succeeded in fitting the compensators of the stochastic processes.

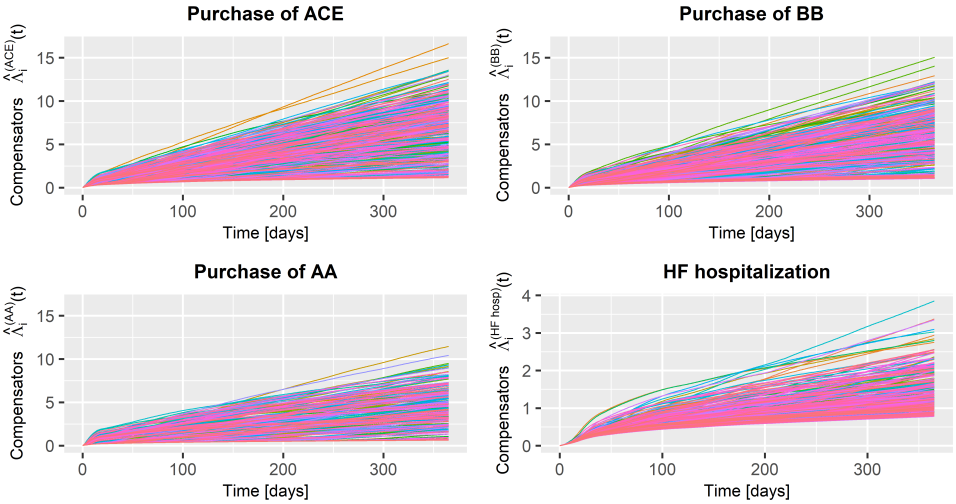

Figure 4: Compensators $\hat{\Lambda}_i^{(h)}(t)$ of the marked counting processes of purchases of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and of HF hospitalizations (bottom-right panel) fitted using Equation (5) for a sample of 500 HF patients belonging to the administrative database. Each line is related to a different patient. Note that in HF hospitalizations the ordinate axis range is smaller than the other ones due to less number of hospitalization events with respect to drugs purchases.

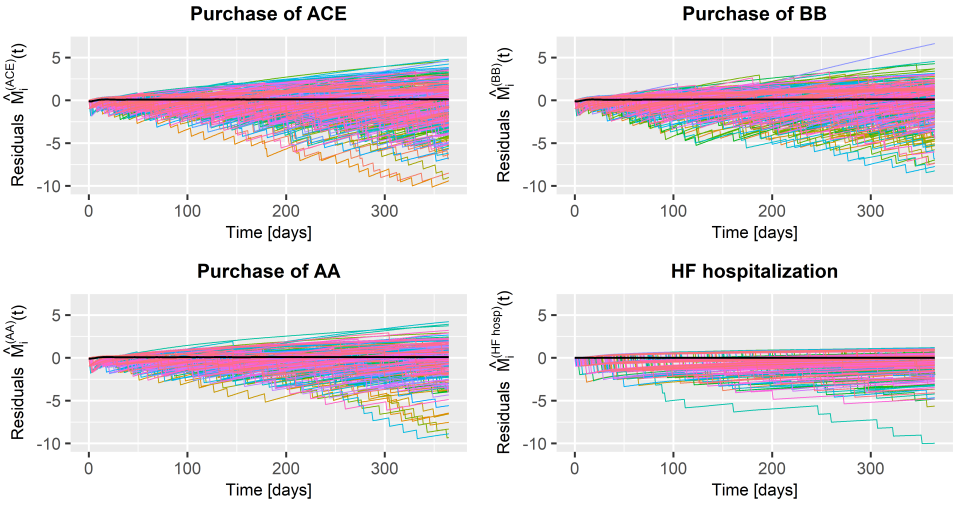

Figure 5: Residuals $\hat{M}_i^{(h)}(t)$ of the compensators of the stochastic process describing the purchase of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and the HF hospitalization (bottom-right panel) for a sample of 500 HF patients belonging to the administrative database, computed according to Equation (6). Each line is related to a different patient. Solid black lines represent the temporal average residual curve $\bar{M}^{(h)}(t)$ computed using all the $n = 4,541$ patients.

For each patient $i \in \{1, \ldots, 4,541\}$, we ended up with a four-variate time-varying data given by the compensator trajectories $\left(\hat{\Lambda}_i^{(h)}(t),\, h \in \mathcal{H} = \{ACE, BB, AA, HF\,hosp\}\right)$, which could be though as positive non-decreasing L^2-functions over the temporal domain $[T_{start}; T_0]$. We can now work on functional compensators $\hat{\Lambda}_i^{(h)}(t)$, applying methods Steps 3 and 4 of our methodology as described in Section 3.

4.3 Step 3: Summarize compensators through FPCA

Once computed the functional trajectories of the compensators $\hat{\Lambda}_i^{(h)}(t)$, we performed FPCA in $L^2\,[T_{start}; T_0]$ in order to summarise information emerging from the time-varying compensators to a finite set of covariates while loosing a minimum part of the information. Although it was no longer guaranteed that the functions reconstructed through FPCA were positive and non-decreasing, for each process h we observed that two Principal Components (PCs) were enough to have a L^2-reconstruction error lower than 1%.

Figure 6 and Figure 7 show results of FPCA on functional compensators and are related to first PCs $\phi_1^{(h)}(t)$ and second PCs $\phi_2^{(h)}(t)$, respectively (top panels). In both figures, each column is related to a different type of process (ACE: first column; BB: second column; AA: third column; *HF hosp*: fourth column). Bottom panels report the average compensators curves $\mu^{(h)}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{\Lambda}_i^{(h)}(t)$ (black lines) and $\mu^{(h)}(t) \pm c_k \sqrt{\nu_k^{(h)}}\phi_k^{(h)}(t)$ (red '+' and blue '−' respectively) where $\nu_k^{(h)}$ is the eigenvalues related to the k-th component, c_k are constants and $k = 1, 2$. We observe that PCs across the four processes types $h \in \{ACE, BB, AA, HF\,hosp\}$ have similar shapes (see top panels). The first components $\phi_1^{(h)}(t)$ distinguish patients with different risks. In particular, a patient with a high score on the first PC is likely to experience more events than a patient with a low score. The second components $\phi_2^{(h)}(t)$ distinguish patients with different time distribution of the events. In particular, a patient with a high score on the second PC is likely to experience more events in the first months of the observation period and less events in the last months of the observation period than a patient with a low score.

Once summarised the information emerging from the functional compensators, we can now use the FPC scores to enrich information for modelling patients' long-term survival through a predictive Cox's regression model.

4.4 Step 4: Predictive survival Cox's model

At this point we wanted to quantify the association between time-varying processes and patients' long-term survival. In order to do so, firstly we applied 10-fold cross validation to select the best set of covariates among all the possible combinations of patients' baseline characteristics *age*, *gender* and the scores resulting from the FPCA on compensators, i.e. $score_k^{(h)}$ with $k = 1, 2$ and $h \in \{ACE, BB, AA, HF\,hosp\}$. According to the highest median Concordance Index,[41] we selected all the covariates except the score related to the second PC of ACE process and the score related to the first PC of AA process. Then, we fitted the Cox's regression model with that choice of covariates on the whole data:

$$\begin{aligned}\lambda_i(t|\mathbf{x}_i) = \lambda_0(t) \exp\big\{&\beta_1 age_i + \beta_2 gender_i + \beta_3 score_{1,i}^{(ACE)} + \beta_4 score_{2,i}^{(AA)} + \\ &\beta_5 score_{1,i}^{(BB)} + \beta_6 score_{2,i}^{(BB)} + \beta_7 score_{1,i}^{(HF\,hosp)} + \beta_8 score_{2,i}^{(HF\,hosp)}\big\}.\end{aligned} \tag{7}$$

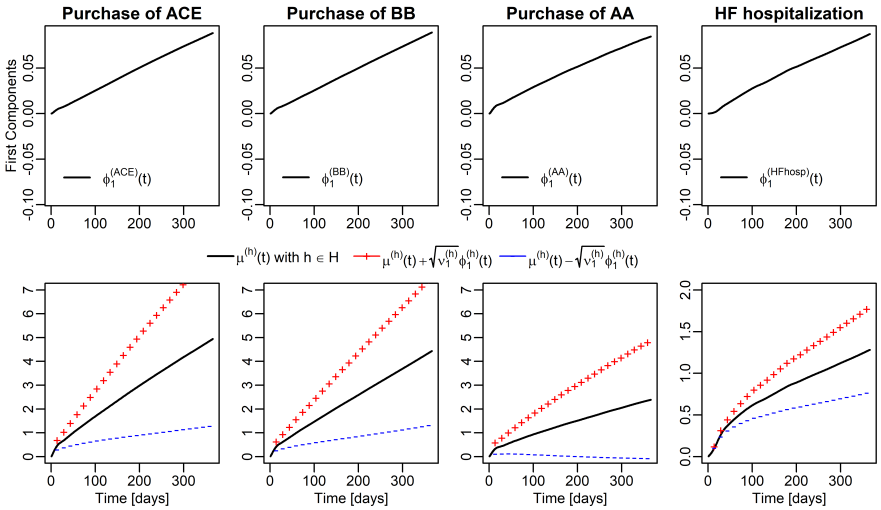

Figure 6: First functional Principal Components (PCs) of the compensators of the stochastic processes describing the purchase of ACE (first column), BB (second column), AA (third column) and HF hospitalization (fourth column). Upper panels show the first PCs $\phi_1^{(h)}(t)$ with $h \in \mathcal{H} = \{ACE, BB, AA, HFhosp\}$. Lower panels report the average compensators curves $\mu^{(h)}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{\Lambda}_i^{(h)}(t)$ (black lines) and $\mu^{(h)}(t) \pm \sqrt{\nu_1^{(h)}}\phi_1^{(h)}(t)$ (red '+' and blue '−' respectively) where $\nu_1^{(h)}$ are the eigenvalues related to the first PCs. Note that in HF hospitalizations the ordinate axis range is smaller than the other ones due to less number of hospitalization events with respect to drugs purchases.

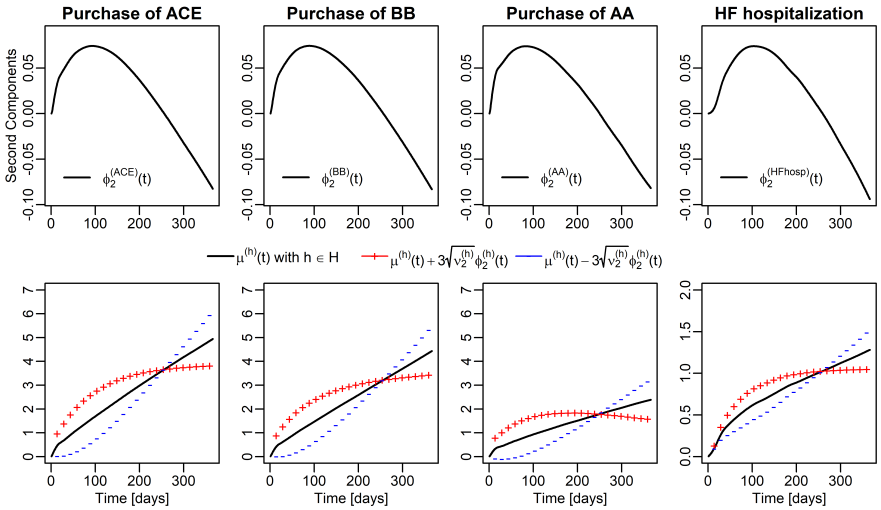

Figure 7: Second functional Principal Components (PCs) of the compensators of the stochastic processes describing the purchase of ACE (first column), BB (second column), AA (third column) and HF hospitalization (fourth column). Upper panels show the second PCs $\phi_2^{(h)}(t)$ with $h \in \mathcal{H} = \{ACE, BB, AA, HFhosp\}$. Lower panels report the average compensators curves $\mu^{(h)}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{\Lambda}_i^{(h)}(t)$ (black lines) and $\mu^{(h)}(t) \pm 3\sqrt{\nu_2^{(h)}}\phi_2^{(h)}(t)$ (red '+' and blue '−' respectively) where $\nu_2^{(h)}$ are the eigenvalues related to the second PCs. Note that in HF hospitalizations the ordinate axis range is smaller than the other ones due to less number of hospitalization events with respect to drugs purchases.

Table 2 reports the summary of the fitted model (7). All the covariates resulted statistically significant at confidence level 5%, except for $score_2^{(AA)}$ and $score_2^{(BB)}$. Elder patients coherently have a higher risk of dying [HR = 1.0656] and being a male corresponds to 1.2-times faster experience of the event [HR = 1.2317]. The HR relative to the scores of the first PCs for ACE and BB processes, i.e. $score_1^{(ACE)}$ and $score_2^{(ACE)}$, are lower than 1, indicating that a proper ACE/BB drug intake is correlated to longer life expectancy. On the contrary, the HR related to $score_1^{(HF\,hosp)}$ is greater than 1, standing as a proxy of patients' critical conditions: patients experiencing many hospitalizations in the past present a higher risk of dying. Interestingly, even if the second PC of compensators related to $HF\,hosp$ process concerned only the 2% of the total explained variance of the original data, $score_2^{(HF\,hosp)}$ is strongly significant with HR $= 0.7539 < 1$ (95% $CI = [0.7256; 0.8253]$). This means that patients with many hospitalizations at the beginning of the observation period and few hospitalizations in the end have higher survival probability, since they probably correspond to the ones who had already experienced a critical phase of the disease and survived from it.

Covariates	**HR**	**CI (2.5%)**	**CI (97.5%)**	p-values
age	1.0656	1.0577	1.0736	< 0.001
gender (Male)	1.2317	1.0892	1.3928	< 0.001
$score_1^{(ACE)}$	0.9977	0.9962	0.9992	0.003
$score_2^{(AA)}$	0.9949	0.9781	1.0121	0.561
$score_1^{(BB)}$	0.9965	0.9945	0.9984	< 0.001
$score_2^{(BB)}$	1.0071	0.9911	1.0234	0.385
$score_1^{(HF\,hosp)}$	1.0158	1.0047	1.0269	0.005
$score_2^{(HF\,hosp)}$	0.7739	0.7256	0.8253	< 0.001

Table 2: Hazard ratios (HRs) with 95% Confidence Intervals (CI) and p-values of the final Cox's model for overall long-term survival fitted on the whole cohort using the covariates selected through 10-fold cross-validation.

5 Discussion and Conclusions

In this work, a novel approach to reconstruct the compensators of suitable marked point processes of interest as time-varying covariates has been proposed in order to exploit this approach for enriching information to be plugged into a survival model. The development of this procedure is due to the need of effectively describing and resuming information from dynamic processes affecting an outcome of interest, with the purpose of obtaining deeper insight on the patient's health status using administrative databases. This methodology extends the one proposed in Baraldo et al,[6] allowing the counting processes to depend on their marks. From the study on the administrative database of Lombardy Region, we observed that modelling patient's clinical history in terms of compensators of suitable stochastic processes as time-varying covariates and plug them into a survival model represents an effective, interpretative and forecasting approach for exploring the effects of these processes on patients' survival. The marked point process formulation is a natural way of representing the occurrence of hospitalizations or drugs purchases over time. The use of FPCA allowed to extract additional information contained in the functions, representing a powerful exploratory and modelling technique for highlighting trends and variations in the shape of the processes over time. The introduction of this novel way to account for time-varying variables by means of compensators allowed for modelling self-exciting behaviours,

for which the occurrence of events in the past increases the probability of a new event. This enabled us to include a large piece of information contained in the administrative data to describe the patient's clinical history. Furthermore, our approach was able to take into account the fact that HF patients usually assumes different types of drugs at the same time, representing a novelty for clinical and pharmacological research in the direction of properly treating multimorbidity patients and polypharmacy. The presented methodology, involving database integration, marked point process modelling of critical events and FDA techniques, can be applied to the study of many different pathologies characterized by complex data sources, thanks to its flexibility. The procedure is very general and allows for a handleable and relatively simple analysis of the results, describing complex dynamics in an easily interpretable form. To the best of our knowledge, our approach represents the first attempt in literature of merging potential of FDA and survival analysis.

Some limitations of the present study have to be mentioned. Firstly, the use of a pre-defined observation period could lead to survival bias due to cohort selection. Indeed, it is necessary that patients survived for a period at least equal to the length of the period used to compute the functional compensators trajectories. This could imply a survival bias in case of the exclusion of too many early dying patients. This is softened if low-rate short-term mortality diseases are considered. Moreover, FPCA was performed in $L^2\left[T_{start}; T_0\right]$ and not in the subspace of positive non-decreasing L^2-functions. In this way, we obtained a good reconstruction of compensators approximated using PCs but it was no longer guaranteed that these functions were positive and non-decreasing. Other limitations are mainly due to the use of secondary databases in the real case-study. In fact, in the administrative database the number of days of drug coverage, that represent the jump marks in case of processes related to ACE, BB and AA purchases, was based on the number of boxes and the Defined Daily Dose (DDD). The use of theoretical DDD instead of Prescribed Daily Doses (PDD) could reflect a bias in the computation of coverage days, i.e. of jump marks, if the underlying PDD/DDD ratio is different from 1.[32,33] Therefore, it could be interesting to explore, whenever the linkage is possible, databases with information about dosages prescribed by doctors, in order to obtain a more realistic analysis of coverage periods. Furthermore, the use of administrative databases allowed to use drugs purchases as proxy of drugs intake with a big limitation: we were not able to assert if the patient was currently consuming the dispensed drug. These issues are related to the nature of administrative data: they address 'operational' goals, i.e. they are collected with no clinical question in mind and mainly for managerial and economic purposes, and the validity of using these kind of data is critically dependent on the reliability of the data. Nevertheless, they are population based, comprehensive, capture real health system use, longitudinal and can be linked to other data, representing a valuable clinical research resource.

Despite the aforementioned limitations, our approach opens doors for many further developments, both in the fields of statistical methods and clinical research. The proposed predictive models could be enriched by considering other relevant clinical information as covariates, and enlarging the cohort of patients. In the end, our approach constitutes a really flexible methodology that can be used to make personalized predictions, quantifying the effect of personal behaviours and therapeutic patterns on survival. Its possible generalization to many different settings, added to a cooperation with medical staff, could lead to improvements in the definition of a useful tool for health care assessment and treatment planning.

Acknowledgments

The authors wish to thank Dr. Davide Burba for the seminal analyses he carried out in his MD thesis, which represented a starting point for the current dissertation, and Prof. Anna Maria Paganoni from Politecnico di Milano for her stimulating suggestions.

References

1. Rondeau, V. Statistical models for recurrent events and death: Application to cancer events. *Math Comput Model* **52,** 949–955 (2010).
2. Duchateau, L., Janssen, P., Kezic, I. & Fortpied, C. Evolution of recurrent asthma event rate over time in frailty models. *J R Stat Soc C* **52,** 355–363 (2003).
3. Charles-Nelson, A., Katsahian, S. & Schramm, C. How to analyze and interpret recurrent events data in the presence of a terminal event: An application on readmission after colorectal cancer surgery. *Stat Med* **38,** 3476–3502 (2019).
4. WHO, International League Against Epilepsy and International Bureau for Epilepsy. *Atlas: Epilepsy Care in the World 2005* (Geneva: World Health Organization, 2005).
5. Kennedy, B. Repeated Hospitalizations and Self-rated Health among the Elderly: A Multivariate Failure Time Analysis. *Am J Epidemiol* **153,** 232–241 (Feb. 2001).
6. Baraldo, S., Ieva, F., Paganoni, A. M. & Vitelli, V. Outcome Prediction for Heart Failure Telemonitoring Via Generalized Linear Models with Functional Covariates. *Scand J Stat* **40,** 403–416 (2013).
7. Rogers, J. K., Yaroshinsky, A., Pocock, S. J., Stokar, D. & Pogoda, J. Analysis of recurrent events with an associated informative dropout time: Application of the joint frailty model. *Stat Med* **35,** 2195–2205 (2016).
8. Lloyd-Jones, D. *et al.* Executive summary: heart disease and stroke statistics–2010 update: a report from the American Heart Association. *Circulation* **121,** 948–954 (2010).
9. MERIT-HF Study Group. Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomised Intervention Trial in Congestive Heart Failure (MERIT-HF). *Lancet* **353,** 2001–2007 (1999).
10. Kalogeropoulos, A. *et al.* Progression to Stage D Heart Failure Among Outpatients With Stage C Heart Failure and Reduced Ejection Fraction. *JACC Heart Fail* **5,** 528–537 (2017).
11. Ponikowski, P. *et al.* 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J* **37,** 2129–2200 (2016).
12. Amorim, L. D. A. F. & Cai, J. Modelling recurrent events: a tutorial for analysis in epidemiology. *Int J Epidemiol* **44,** 324–333 (Dec. 2014).
13. Ozga, A., Kieser, M. & Rauch, G. A systematic comparison of recurrent event models for application to composite endpoints. *BMC Med Res Methodol* **18** (2018).
14. Yadav, C. P., Sreenivas, V., Khan, M. A. & Pandey, R. M. An Overview of Statistical Models for Recurrent Events Analysis: A Review. *OMICS* **8,** 1–5 (2018).
15. Andersen, P. K. & Gill, R. D. Cox's Regression Model for Counting Processes: A Large Sample Study. *Ann Stat* **10,** 1100–1120 (1982).

16. Cox, D. Regression models and life tables (with discussion). *J R Stat Soc* **34,** 187–220 (1972).

17. Prentice, R. L., Williams, B. J. & Peterson, A. V. On the regression analysis of multivariate failure time data. *Biometrika* **68,** 373–379 (Aug. 1981).

18. Kleinbaum, D. G. & Klein, M. *Survival Analysis: A Self-Learning Text* (Springer New York, 2013).

19. Kelly, P. J. & Lim, L. L.-Y. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Stat Med* **19,** 13–33 (2000).

20. Gasperoni, F., Ieva, F., Paganoni, A., Jackson, C. & Sharples, L. Non-parametric frailty Cox models for hierarchical time-to-event data. *Biostatistics* (Dec. 2018).

21. Cook, R. J. & Lawless, J. F. *The Statistical Analysis of Recurrent Events* (New York: Springer, 2007).

22. Andersen, P. K. & Keiding, N. Multi-state models for event history analysis. *Stat Methods Med Res* **11,** 91–115 (2002).

23. Ieva, F. & Gasperoni, F. Discussion of the paper "Statistical challenges of administrative and transaction data" by David J. Han. *J R Stat Soc A* **181,** 591–592 (2018).

24. Ieva, F., Jackson, C. & Sharples, L. Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology. *Stat Methods Med Res* **26,** 1350–1372 (2017).

25. Lee, D. *et al.* Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. *Med Care* **43,** 182–188 (2005).

26. Saczynski, J. *et al.* A systematic review of validated methods for identifying heart failure using administrative data. *Pharmacoepidem Dr S* **21,** 129–140 (2012).

27. Regione Lombardia. HFData project: Utilization of Regional Health Source Databases for Evaluating Epidemiology, short- and medium-term outcome and process indicators in patients hospitalized for heart failure. *Progetto di Ricerca Finalizzata di Regione Lombardia - HFData-RF-2009-1483329* (2012).

28. Mazzali, C. *et al.* Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012. *BMC Health Serv Res* **16** (2016).

29. Ramsay, J. & Silverman, B. W. *Functional Data Analysis* (Springer New York, 2005).

30. Kong, D., Ibrahim, J. G., Lee, E. & Zhu, H. FLCRM: Functional linear cox regression model. *Biometrics* **74,** 109–117 (2018).

31. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2018). `https://www.R-project.org/`.

32. Spreafico, M. *et al.* Adherence to disease-modifying therapy in Heart Failure patients at 1-year from HF hospitalization: findings from a community-based study. *Ame J Cardiovasc Drug* (2019).

33. World Health Organization. *Introduction to Drug Utilization Research* (WHO Library Cataloguing-in-Publication Data, 2003).

34. McMurray, J., Cohen-Solal, A., Dietz, R., *et al.* Practical recommendations for the use of ACE inhibitors, beta-blockers, aldosterone antagonists and angiotensin receptor blockers in heart failure: Putting guidelines into practice. *Eur J Heart Fail* **7,** 710–721 (2005).

35. McMurray, J. J. *et al.* ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treat-ment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. *Eur Heart J* **14,** 803–869 (2012).

36. Ross, S. *Stochastic Processes* (Wiley, 1995).

37. Meyer, P. A. A decomposition theorem for supermartingales. *Illinois J Math* **6,** 193–205 (1962).

38. Lee, E. T. & Wang, J. *Statistical Methods for Survival Data Analysis* (Wiley, 2003).

39. Breslow, N. E. Analysis of survival data under the proportional hazards model. *Int Stat Rev,* 45–57 (1975).

40. Therneau, T. M., Grambsch, P. M. & Fleming, T. R. Martingale-based residuals for survival models. *Biometrika* **77,** 147–160 (Mar. 1990).

41. Pencina, M. J. & D'Agostino, R. B. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* **23,** 2109–2123 (2004).

42. He, X. & Ng, P. COBS: Qualitatively constrained smoothing via linear programming. *Comput Stat* **14,** 315–338 (1999).

43. Therneau, T. M. & Lumley, T. Package 'survival'. *Survival analysis Published on CRAN* (2014).

A Data Preparation

Once selected the cohort of patients being part of the analysis and identified the events related to each patient's clinical history (Section 4.1 - Step 1 of the procedure), we had to reformat the administrative data building four different datasets, one for each process $h \in \mathcal{H} = \{h : ACE, BB, AA, HF\,hosp\}$, in the form required by coxph function for recurrent events of survival R package.[43] Table 3 shows an example of reformatted dataset related to HF hospitalization process for a random patient with three hospitalizations due to HF during the observation period. In the Table, *status* is the event indicator (0 if censored, 1 otherwise), *start* indicates the time of the patient's previous event (equal to -0.5 if it is the first event), *stop* is the time of the current event (equal to 365.5 if it is the censoring event), $Nm^{(h)}(t)$ is the number of events related to process h occurred in the past and $y^{(h)}(t)$ is the sum of the corresponding marks. In particular, the choice to consider the time limits at -0.5 and 365.5 was made in order to not have events at time $t = 0$ or at censoring time $t = 365$. Hence, for each process h we ended up with a long-format dataset with multiple rows for each patient (specifically the number of patient's events of type h during the observation period plus one). In particular, in the first row of each patient $Nm^{(h)}(t)$ and $y^{(h)}(t)$ are always 0 and in the last one *status* is always equal to 0.

ID	*status*	*start*	*stop*	*gender*	*age*	$Nm^{(h)}(t)$	$y^{(h)}(t)$
id	1	-0.5	30	*Female*	79	0	0
id	1	30	52	*Female*	79	1	21
id	1	52	52	*Female*	79	2	35
id	0	76	365.5	*Female*	79	3	59

Table 3: Example of reformatted dataset related to HF hospitalization process for a random patient with three hospitalizations due to HF during the observation period.

B Mean Absolute Martingale Residual

Given two or more Cox models for recurrent events in Equation (3) fitted using different sets of covariates, we need a metric to evaluate the goodness of fit of each model and select the best set of features. Since we are dealing with stochastic processes and recurrent events, we cannot rely on standard regression metrics, like mean squared error. A possible way is given by functions of the residuals in Equation (6): smaller residuals correspond to a greater predictive power of the model. Therefore, to compare models fitted with different features, for each process h we would like to use the Mean Absolute Martingale Residual (MAMR):

$$MAMR^{(h)} = \sum_{i=1}^{n} \frac{\int_0^T |\hat{M}_i^{(h)}(s)|ds}{T} \tag{8}$$

where T represents the length of the observation period. Using this indicator, smaller the MAMR, better the model.

To correctly compute the MAMR, we should first compute the compensators using Equation (5) and then evaluate the residuals on a grid of points. Since we want to use this quantity only to rank models fitted with different sets of predictors, to avoid high computational costs we decided to rely on the following estimate:

$$\widehat{MAMR}^{(h)} = \frac{1}{\sum_{i=1}^{n} n_i^{(h)}} \sum_{i=1}^{n} \sum_{j=1}^{n_i^{(h)}} |\hat{M}_i^{(h)}(t_{i,j}^{(h)})| \tag{9}$$

where i and h are respectively the patient and event indices, $\hat{\hat{M}}_i^{(h)}$ is the residual obtained by fitting the compensator without smoothing the baseline hazard, i.e. using $\hat{\Lambda}_0^{(h)}$ instead of $\tilde{\Lambda}_0^{(h)}$ in Equation (5), $n_i^{(h)}$ is the total number of events of type h experienced by the i-th patient and $t_{i,j}^{(h)}$ is the time instant in which patient i experienced the j-th event of type h.
This estimate is not accurate since the residuals are evaluated only when events happen (rather than on the continuous interval corresponding to the one year observation period) and because the estimate is done by reconstructing the compensators without the smoothing of the baseline hazard. However, it allows to rank models while limiting computational needs (refer to residuals function of survival R package[43]).

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

21/2020 Benacchio, T.; Bonaventura, L.; Altenbernd, M.; Cantwell, C.D.; Düben, P.D.; Gillard, M.; Gir
Resilience and fault-tolerance in high-performance computing for numerical weather and climate prediction

22/2020 Zeni, G.; Fontana, M.; Vantini, F.
Conformal Prediction: a Unified Review of Theory and New Challenges

20/2020 Almi, S.; Belz, S.; Micheletti, S.; Perotto, S.
A DIMENSION-REDUCTION MODEL FOR BRITTLE FRACTURES ON THIN SHELLS WITH MESH ADAPTIVITY

19/2020 Stella, S.; Vergara, C.; Maines, M.; Catanzariti, D.; Africa, P.; Dematté, C.; Centonze, M.; Nob
Integration of maps of activation times in computational cardiac electrophysiology

16/2020 Paolucci, R.; Mazzieri, I.; Piunno, G.; Smerzini, C.; Vanini, M.; Ozcebe, A.G.
Earthquake ground motion modelling of induced seismicity in the Groningen gas field

17/2020 Cerroni, D.; Formaggia, L.; Scotti, A.
A control problem approach to Coulomb's friction

18/2020 Fumagalli, A.; Scotti, A.; Formaggia, L.
Performances of the mixed virtual element method on complex grids for underground flow

15/2020 Fumagalli, I.; Fedele, M.; Vergara, C.; Dede', L.; Ippolito, S.; Nicolò, F.; Antona, C.; Scrofani,
An Image-based Computational Hemodynamics Study of the Systolic Anterior Motion of the Mitral Valve

13/2020 Pozzi S.; Domanin M.; Forzenigo L.; Votta E.; Zunino P.; Redaelli A.; Vergara C.
A data-driven surrogate model for fluid-structure interaction in carotid arteries with plaque

14/2020 Calissano, A.; Feragen, A; Vantini, S.
Populations of Unlabeled Networks: Graph Space Geometry and Geodesic Principal Components