Statistical modelling of time-varying covariates for survival data

Spreafico, M.

Chapter 3

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

This chapter has been published in *Biometrical Journal*, 63(5):948–967, 2021 as M. Spreafico and F. Ieva "Functional modeling of recurrent events on time-to-event processes" [189].

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 drug consumption, hospital admissions and many others. Examples include recurrences in breast cancer [174], asthma attacks [52], episodic relapses of follicular lymphoma [174], readmission after colorectal cancer [64, 36], epileptic seizures [215]. In patients with HF, two main types of events recur during treatment: (i) repeated consumption of multiple types of drugs and (ii) hospital readmissions [104, 21, 173]. 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, along with re-hospitalizations events which usually herald a substantial worsening of patient's survival prognosis. As discussed in Chapter 2, the natural and most appropriate way to look at these repeated events is to treat them as time-varying covariates, since their changing patterns over time could carry out information that may be related to patient's health status and disease progression.

In biostatistical, epidemiological and medical literature, several approaches to analyse recurrent event data have been proposed and compared [202, 103, 44, 95, 106, 11, 149]. 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 is the AG model by Andersen and Gill (1982) [12], which is an extension of the Cox proportional-hazard regression by Cox (1972) [46]. The AG model for recurrent events 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 [11]. Two alternative approaches are the stratified Cox-type conditional [156] and marginal [213] models, which can incorporate both overall and event-specific effects for each covariate. The stratified conditional Prentice-Williams-Peterson (PWP) model analyses repeated events ordered by stratification, based on the prior number of events during the follow-up period. However, it can give unreliable estimates for higher order of events [11]. The stratified marginal Wei-Lin-Weissfeld (WLW) model ignores the order of occurrence of the events. Therefore, an individual is at risk for every event as long as he/she is under observation, even if no previous events occurred, leading to a 'carry-over effect' as explained by [103] and [149]. As a further alternative, Cox model can be also extended using frailty models [85, 202, 175, 176, 44, 106, 60], 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 the random effects, and it is used to model individual patients' heterogeneity in the baseline hazards. Furthermore, approaches able to connect several event processes (recurrent and fatal/non-fatal ones) have been proposed. Among others, (copula-based) joint frailty models [177, 174, 54, 124, 125] allow the prediction of a terminal event time given recurrent event times. Alternatively, rate-based models [34, 197, 37, 224, 196] or multi-state models [13, 44] can be used in case of multiple types of recurrent events. The choice of the proper approach for the analysis of recurrent event data will therefore be determined by many factors, including among others, number and types of events, relationship between subsequent events and biological processes [11].

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. In this framework, Baraldo *et al.* (2013) [21] proposed a method to model the realized trajectories of the cumulative hazard functions underlying a recurrent event process of interest (i.e., hospital readmissions in time). Estimated trajectories were treated as functional data and included into a generalized linear model to predict a binary telemonitoring outcome. However, the authors focused only on a counting process formulation for recurrent events, without considering further information about them. Indeed, many situations and events are characterized by both a location (in time or space) and a weight or other distinguish attribute, called *mark* [47]. For example, in HF treatment a longer period in hospital could reflect the aggravation of patient's health condition, as well as a shorter drug coverage period could lead to nonadherence to therapy, commonly associated with adverse health conditions [102, 187, 188]. The development of models and methods able to deal with all these peculiar aspects is of statistical interest and of clinical relevance.

Motivated by the clinical question concerning the effect of re-hospitalizations and subsequent consumption of different drugs on survival in HF patients, in this chapter we proposed a new methodology that exploits recurrent events modelling [44], point processes

theory [113, 47] and Functional Data Analysis (FDA) [162] to represent time-varying events in terms of functions, plugging them into a suitable functional Cox model for overall survival. In order to take into account many aspects that could influence the events, our idea was to look at time-varying recurrent events as particular non-stationary stochastic counting processes which can depend on their marks, i.e., marked point processes [113, 47]. Starting from the idea by [21], we developed a *marked point process formulation for recurrent events* to compute the realized trajectories of the cumulative hazard functions (i.e., the *compensators*) underlying specific counting processes of interest, allowing the dependence on the marks. In particular, among the aforementioned methods to deal with recurrent events, we modelled the compensators through AG models [12], ending up with functional data that represent the dynamic evolution of the events risk. Then, we applied Functional Principal Component Analysis (FPCA) [162] in order to perform a dimensionality reduction and summarise information emerging from the functional compensators to a finite set of covariates, while losing a minimum part of the information. This information was finally included into a functional linear Cox regression model [109], extended to the case of multiple functional predictors.

The procedure presented in this chapter can hence be divided into two phases:

(i) the representation of time-varying functional compensators,

(ii) the modelling of such covariates in a time-to-event framework.

In doing so, we aimed to enrich 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' overall survival.

The remaining part of the chapter is organized as follows. In Section 3.1 we describe the real study design used in this work. In Section 3.2 we present the whole methodology. First, we focus on the main novelty introduced by the present work, i.e., the *marked point process formulation for recurrent events* to represent the compensators (Section 3.2.1). Then we introduce the functional linear Cox regression model for overall survival in case of multiple functional predictors (Section 3.2.2). In Section 3.3 we apply the proposed methodology to HF administrative database provided by *Regione Lombardia - Healthcare Division* [164] . Finally, Section 3.4 contains some concluding remarks, discussion of strengths and limitations of the proposed approach and opportunities for future work. Statistical analyses were performed in the R software environment [161]. Source code is available as Supporting Information of [189].

3.1. Materials and Administrative data

3.1.1. Administrative data sources

As in Chaper 2, in this work we focused on a representative sample of the real administrative *HFData* database [136] provided by *Regione Lombardia - Healthcare Division* [164]

related to non-paediatric patients living in Lombardy with their first HF discharge (*index event*) between January 2006 to December 2012. As explained in Section 2.2.2, patients' clinical history of hospitalizations or drug consumption could be reconstructed using secondary registry data related to (i) patient admission to hospital (i.e., date of discharge from hospital, length of stay in hospital) and (ii) pharmaceutical purchases (i.e., ATC code, date of purchase, number of treatment days covered by the prescription). Among the disease-modifying drugs for HF patients [138, 139, 154], we focused on polypharmacy treatment as a combination of Angiotensin-Converting Enzyme (ACE) inhibitors, Beta-Blocking (BB) agents and Anti-Aldosterone (AA) agents.

3.1.2. Study design and outcome measure

Figure 3.1 shows the study design. A 5-years *pre-study period* from 2000 to 2005 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 (time T_0 in Figure 3.1) and was divided into the *observation period* (365 days from the index date), used for the compensators reconstruction, and the *follow-up period*, used for the survival analysis, whose starting time was $T_0^* = T_0 + 365$. The modelling of the compensators related to the stochastic processes of interest regarded the time interval $[T_0; T_0^*]$ in Figure 3.1. 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). Overall survival was measured from the end of

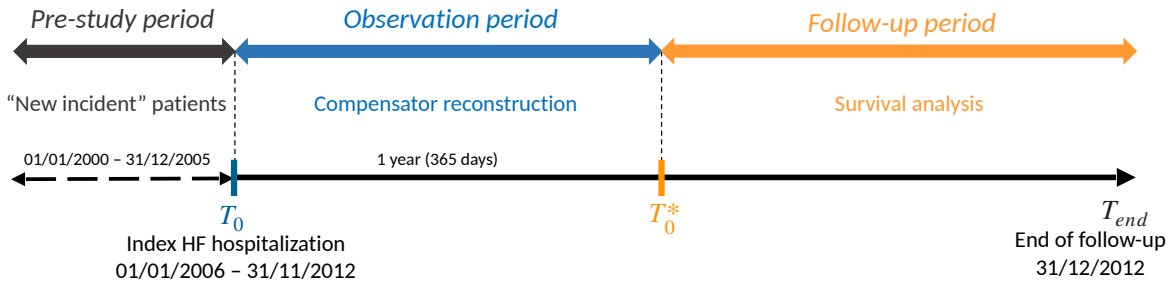

Figure 3.1. Study design for a HF patient of the study cohort. The *pre-study period* is used to define "incident" HF patients. The *observation period* is used for the selection of patient's clinical history and the compensators reconstruction. The *follow-up period* is used for survival analysis. T_0 is the time instant the patient is discharged by her/his first hospitalization and enrolled into the current study. $T_0^* = T_0 + 365$ is the starting time of the follow-up. T_{end} is the minimum between the death or the administrative censoring (December 31^{st}, 2012).

the *observation period* (T_0^* in Figure 3.1) to the date of death or to the administrative censoring date (December 31st, 2012). Outcome (T_i, D_i) denotes the observed time-to-death data of patient $i \in \{1, ..., n\}$, where $T_i = \min(T_i^*, C_i)$ is the observed event time, T_i^* is the true event time, C_i is the censoring time and $D_i = \mathbb{1}(T_i^* \leq C_i)$ is the event indicator, with $I(\cdot)$ being the indicator function that takes the value 1 when $T_i^* \leq C_i$, and 0 otherwise. Independent censoring between true death and censoring times was assumed.

3.2. Statistical Methodologies

We now introduce the methodology developed and then applied on the case study of interest in Section 3.3. In Section 3.2.1 we focus on the main novelty introduced by the present work, i.e., the *marked point process formulation for recurrent events*. In Section 3.2.2 we introduce the functional linear Cox regression model for overall survival in case of multiple functional compensators.

3.2.1. Marked point process formulation for recurrent events

A recurrent event process is characterized by an increasing sequence of *events times*, where each element denotes the time of the corresponding event [44]. To this sequence of times could be associated (i) a *counting process* that at time t records the cumulative number of events occurred up to t [44] and (ii) other random elements, called *marks*, containing further information about the events [113, 47]. Marks can also be thought of as the size, weight or magnitude related to the jumps of the counting process. Extending the approach by [21], we now introduce the *marked point process formulation for recurrent events* to compute the realized trajectories of the compensators underlying a specific counting process of interest, allowing the dependence on the marks.

Let us consider a set $\mathcal{M}$ of recurrent events for a set of n individuals as stochastic processes. For each patient $i \in \{1, ..., n\}$, let $\left\{t_{i,j}^{(m)}, j = 0, 1, ..., n_i^{(m)}\right\}$ be the increasing sequence of event times related to recurrent event process m, where $n_i^{(m)}$ is the total number of events of type m experienced by the i-th subject, $t_{i,j}^{(m)}$ denotes the time of the j-th event and $t_{i,0}^{(m)} = 0 \, \forall i, m$. Let $\mathbf{w}_i^{(m)}$ be the vector of marks elements, where each *jump mark* $\mathrm{w}_{i,j}^{(m)}$ is the magnitude of the information associated to each *jump time* $t_{i,j}^{(m)}$. The observations (possibly censored) may be considered as the realisation of $N_1^{(m)}, ..., N_n^{(m)}$ processes, where $N_i^{(m)}$ is the stochastic process which counts the observed events (or jumps) of the process m in the *observation period* related to the i-th individual. According to the Doob–Meyer (D-M) decomposition theorem [142], each counting process $N_i^{(m)}(t)$, adapted to the filtration $\{\mathcal{F}_{t,i}^{(m)}, t \geq 0\}$ representing the history of realisations of the process itself, can be seen as:

$$N_i^{(m)}(t) = M_i^{(m)}(t) + \Lambda_i^{(m)}(t) = M_i^{(m)}(t) + \int_0^t \lambda_i^{(m)}(s) ds \tag{3.1}$$

where $M_i^{(m)}(t)$ is a zero-mean uniformly integrable martingale which represents the residual of the process, and $\Lambda_i^{(m)}(t) = \int_0^t \lambda_i^{(m)}(s)ds$ is a unique predictable, non-decreasing, *cadlag* (right-continuous with left limits) and integrable process, i.e., the *compensator* (or *cumulative hazard*). Process $\lambda_i^{(m)}(t)$ is the *conditional intensity function*, in which we omitted the conditioning with respect to the history $\mathcal{F}_{t,i}^{(m)}$ for ease of notation, and represents the infinitesimal risk of occurrence of an event m at time t, given the history, i.e., $\lambda_i^{(m)}(t) = \lim_{\Delta t \to 0} \mathbb{E}\left[N_i^{(m)}(t+\Delta t) - N_i^{(m)}(t) | \mathcal{F}_{t,i}^{(m)}\right] / \Delta t$. The compensator $\Lambda_i^{(m)}(t)$ may be thought of 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 marked point process, assuming that a given distribution regulates the size of the jumps. A marked point process is the couple of processes describing the behaviour of jump times and marks modelled through the *conditional intensity function* $\lambda_i^{(m)}\left(t, \mathbf{w}_i^{(m)}\right)$, i.e., the infinitesimal risk of occurrence of event m at time t with marks $\mathbf{w}_i^{(m)}$ given the history:

$$\lambda_i^{(m)}\left(t, \mathbf{w}_i^{(m)}\right) = \lambda_{ig}^{(m)}(t) f_i^{(m)}\left(\mathbf{w}_i^{(m)}\right) \tag{3.2}$$

where $\lambda_{ig}^{(m)}$ is the intensity process of the counting process, also called *ground intensity*, and $f_i^{(m)}$ is the multivariate density of the marks $\mathbf{w}_i^{(m)}$. Using this formulation, conditional independence of jump times and marks is assumed. Note that, if $\lambda_i^{(m)}\left(t, \mathbf{w}_i^{(m)}\right)$ is properly modelled, the D-M decomposition in (3.1) is still valid in the marked point process framework considering Equation (3.2) as conditional intensity process.

To handle recurrent events and allow predictors to change over time, we use the counting process formulation for recurrent events introduced by [12], also called AG model for recurrent events, assuming a particular distribution for the marks in order to ease computations. In particular, we assume that the density $f_i^{(m)}$ depends on some time-dependent features related to the marks $\mathbf{w}_i^{(m)}$. Under these hypotheses, for each event m the conditional intensity function $\lambda_i^{(m)}\left(t, \mathbf{w}_i^{(m)}\right)$ in Equation (3.2) related to patient i takes the form:

$$\begin{aligned} \lambda_i^{(m)}\left(t, \mathbf{w}_i^{(m)}\right) &= Y_i^{(m)}(t)\lambda_0^{(m)}(t) \exp\left\{\boldsymbol{\beta}^{(m)^T} \mathbf{x}_i^{(m)}(t)\right\} \exp\left\{\boldsymbol{\gamma}^{(m)^T} \mathbf{z}_i^{(m)}(t)\right\} \\ &= Y_i^{(m)}(t)\lambda_0^{(m)}(t) \exp\left\{\boldsymbol{\beta}^{(m)^T} \mathbf{x}_i^{(m)}(t) + \boldsymbol{\gamma}^{(m)^T} \mathbf{z}_i^{(m)}(t)\right\} = \lambda_i^{(m)}(t) \end{aligned} \tag{3.3}$$

where $\mathbf{x}_i^{(m)}(t)$ and $\mathbf{z}_i^{(m)}(t)$ are the possibly time-dependent vectors of covariates of the i-th individual, the latter related to the marks $\mathbf{w}_i^{(m)}$. Parameters $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\gamma}^{(m)}$ are fixed vectors of coefficients, $\lambda_0^{(m)}$ is the baseline hazard function shared across patients and $Y_i^{(m)}$ is a predictable process taking values in $\{0, 1\}$. Whenever $Y_i^{(m)} = 1$, the i-th individual is under observations, i.e., $Y_i^{(m)}$ takes the role of the censoring variable.
Parameters $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\gamma}^{(m)}$ are estimated maximizing the partial likelihood function constructed given the history, using a counting process approach [12]. The baseline cumulative

hazard $\Lambda_0^{(m)}(t) = \int_0^t \lambda_0^{(m)}(s)ds$ can be estimated $\forall m \in \mathcal{M}$ using the *Breslow estimator* [32] $\hat{\Lambda}_0^{(m)}(t)$, which returns a step-function. However, since true underlying functions $\Lambda_0^{(m)}(t)$ are absolutely continuous, we smooth the estimates using the approach adopted in [21], obtaining regularised version of $\Lambda_0^{(m)}(t)$, namely $\tilde{\Lambda}_0^{(m)}(t)$.

Let us now consider the sequence $0 = t_{i,0}^{(m)} < t_{i,1}^{(m)} < ... < t_{i,N_i^{(m)}(\tau)}^{(m)}$ of realised jump times related to process $N_i^{(m)}(t)$, with τ equal to the censoring time (possibly equal for all individuals or not) and $n_i^{(m)} = N_i^{(m)}(\tau)\,\forall m, i$. In our case, τ is the censoring time of the *observation period*, i.e., T_0 in Figure 3.1. We can express the realisations of each compensator $\Lambda_i^{(m)}(t)$ for the process m of the i-th patient as a function of $\Lambda_0^{(m)}(t)$, $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\gamma}^{(m)}$:

$$\begin{aligned}
\Lambda_i^{(m)}(t) &= \int_0^t \lambda_i^{(m)}(s)ds = \int_0^t Y_i^{(m)}(s)\lambda_0^{(m)}(s)\exp\left\{\boldsymbol{\beta}^{(m)^T}\mathbf{x}_i^{(m)}(s) + \boldsymbol{\gamma}^{(m)^T}\mathbf{z}_i^{(m)}(s)\right\}ds \\
&= \sum_{j=1}^{N_i^{(m)}(t)} \int_{t_{i,j-1}^{(m)}}^{min\left(t_{i,j}^{(m)},t\right)} \lambda_0(s)\exp\left\{\boldsymbol{\beta}^{(m)^T}\mathbf{x}_i^{(m)}(t_{i,j-1}) + \boldsymbol{\gamma}^{(m)^T}\mathbf{z}_i^{(m)}(t_{i,j-1})\right\}ds \\
&= \sum_{j=1}^{N_i^{(m)}(t)} \exp\left\{\boldsymbol{\beta}^{(m)^T}\mathbf{x}_i^{(m)}(t_{i,j-1}) + \boldsymbol{\gamma}^{(m)^T}\mathbf{z}_i^{(m)}(t_{i,j-1})\right\}\left[\Lambda_0^{(m)}\left(\min\left(t_{i,j}^{(m)},t\right)\right) - \Lambda_0^{(m)}\left(t_{i,j-1}^{(m)}\right)\right].
\end{aligned} \tag{3.4}$$

An estimate of the compensator in Equation (3.4) can be then obtained as:

$$\hat{\Lambda}_i^{(m)}(t) = \sum_{j=1}^{N_i^{(m)}(t)} \exp\left\{\hat{\boldsymbol{\beta}}^{(m)^T}\mathbf{x}_i^{(m)}(t_{i,j-1}) + \hat{\boldsymbol{\gamma}}^{(m)^T}\mathbf{z}_i^{(m)}(t_{i,j-1})\right\}\left[\tilde{\Lambda}_0^{(m)}\left(min\left(t_{i,j}^{(m)},t\right)\right) - \tilde{\Lambda}_0^{(m)}\left(t_{i,j-1}^{(m)}\right)\right] \tag{3.5}$$

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

To check the fitting of $\hat{\Lambda}_i^{(m)}(t)$, we have to verify whether the estimates of martingale residuals $M_i^{(m)}(t)$ involved in the D-M decomposition (3.1), i.e., the residuals [203] given by

$$\hat{M}_i^{(m)}(t) = \hat{\Lambda}_i^{(m)}(t) - N_i^{(m)}(t), \tag{3.6}$$

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

This formulation extends the one proposed in [21], 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 $\forall m \in \mathcal{M}$, we end up with a multivariate time-dependent data $\left\{\Lambda_i^{(m)}\right\}_{m\in\mathcal{M}}$ for each patient i, characterizing her/his recurrent events dynamics during the *observation period* $[T_0; T_0^*]$. These compensator trajectories may be thought of as patient-specific time-varying covariates and, mathematically speaking, as positive non-decreasing L^2-functions over the temporal domain $[T_0; T_0^*]$.

3.2.2. Functional linear Cox regression model with multiple functional compensators

To include the functional compensators into a survival model, the functional linear Cox regression model introduced by Kong *et al.* (2018) [109] can be extended to the case of multiple functional predictors, i.e., Multivariate Functional Linear Cox Regression Model (MFLCRM). For each patient i, let $\left\{\Lambda_i^{(m)}\right\}_{m\in\mathcal{M}}$ be the realizations of the $|\mathcal{M}|$-variate compensators related to a set $\mathcal{M}$ of recurrent events. The functional compensators are included in the hazard function of Cox model [46] as:

$$h_i\left(t|\boldsymbol{\omega}_i, \left\{\Lambda_i^{(m)}\right\}_{m\in\mathcal{M}}\right) = h_0(t)\exp\left\{\boldsymbol{\theta}^T\boldsymbol{\omega}_i + \sum_{m\in\mathcal{M}}\int_{S_m}\Lambda_i^{(m)}(s)\alpha^{(m)}(s)ds\right\} \tag{3.7}$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{\omega}_i$ is the vector of scalar (non functional) covariates with regression parameters $\boldsymbol{\theta}$. The realizations $\left\{\Lambda_i^{(m)}\right\}_{m\in\mathcal{M}}$ are defined over the temporal domains $S_m = [T_0; T_0^*]\ \forall m$. Parameters $\alpha^{(m)}(s)$ denote the functional regression coefficients.

By applying Functional Principal Component Analysis (FPCA) [162], each functional compensator $\Lambda_i^{(m)}(s)$ can be approximated with a finite sum of K_m orthonormal basis $\left\{\xi_1^{(m)}, ..., \xi_{K_m}^{(m)}\right\}$:

$$\Lambda_i^{(m)}(s) \approx \mu^{(m)}(s) + \sum_{k=1}^{K_m} f_{ik}^{(m)}\xi_k^{(m)}(s) \tag{3.8}$$

where $\mu^{(m)}(s)$ is the functional compensator mean and $f_{ik}^{(m)}$ is the FPC score of individual i related to the k-th orthonormal base $\xi_k^{(m)}$ and K_m is the truncation parameter, representing the number of FPCs to be considered. In particular, the score $f_{ik}^{(m)}$ represents the projection of the i-th functional observation $\Lambda_i^{(m)}(t)$ related to event m along the direction of the k-th principal component $\xi_k^{(m)}(t)$. From (3.8) the integrals in (3.7) can be approximated considering:

$$\begin{aligned}\int_{S_m}\left[\Lambda_i^{(m)}(s) - \mu^{(m)}(s)\right]\alpha^{(m)}(s)ds &\approx \int_{S_m}\sum_{k=1}^{K_m} f_{ik}^{(m)}\xi_k^{(m)}(s)\alpha^{(m)}(s)ds \\ &= \sum_{k=1}^{K_m} f_{ik}^{(m)}\int_{S_m}\xi_k^{(m)}(s)\alpha^{(m)}(s)ds = \sum_{k=1}^{K_m} f_{ik}^{(m)}\alpha_k^{(m)}\end{aligned} \tag{3.9}$$

where $\alpha_k^{(m)}$ is the scalar representing the quantity $\int_{S_m}\xi_k^{(m)}(s)\alpha^{(m)}(s)ds$. Introducing approximation (3.9) in Equation (3.7), the hazard function becomes:

$$\begin{aligned}h_i\left(t|\boldsymbol{\omega}_i, \left\{\Lambda_i^{(m)}\right\}_{m\in\mathcal{M}}\right) &= h_0(t)\exp\left\{\boldsymbol{\theta}^T\boldsymbol{\omega}_i + \sum_{m\in\mathcal{M}}\left[\int_{S_m}\mu^{(m)}(s)\alpha^{(m)}(s)ds + \sum_{k=1}^{K_m} f_{ik}^{(m)}\alpha_k^{(m)}\right]\right\} \\ &= h_0^*(t)\exp\left\{\boldsymbol{\theta}^T\boldsymbol{\omega}_i + \sum_{m\in\mathcal{M}}\sum_{k=1}^{K_m} f_{ik}^{(m)}\alpha_k^{(m)}\right\}\end{aligned} \tag{3.10}$$

where $h_0^*(t) = h_0(t) \exp\left\{\sum_{m\in\mathcal{M}} \int_{S_m} \mu^{(m)}(s)\alpha^{(m)}(s)ds\right\}$ is the baseline hazard function and $\alpha_k^{(m)} = \int_{S_m} \xi_k^{(m)}(s)\alpha^{(m)}(s)ds$ is the regression parameter related to the k-th FPC score of the functional compensator of event m. Therefore, defining the following quantities:

$$\widetilde{\boldsymbol{\theta}} = \left[\boldsymbol{\theta}^T, \left\{\left(\alpha_1^{(m)}, ..., \alpha_{K_m}^{(m)}\right)\right\}_{m\in\mathcal{M}}\right]^T$$
$$\widetilde{\boldsymbol{\omega}}_i = \left[\boldsymbol{\omega}_i^T, \left\{\left(f_{i1}^{(m)}, ..., f_{iK_m}^{(m)}\right)\right\}_{m\in\mathcal{M}}\right]^T$$

and substituting them in Equation (3.10), through FPCA the MFLCRM can be expressed as Cox model with hazard function

$$h_i(t|\widetilde{\boldsymbol{\omega}}_i) = h_0(t) \exp\left\{\widetilde{\boldsymbol{\theta}}^T \widetilde{\boldsymbol{\omega}}_i\right\}.$$

All the properties of the Cox model still hold in this framework and the vector of coefficients $\tilde{\boldsymbol{\theta}}$ can be estimated by maximising the partial likelihood function [46]. In R software [161] the MFLCRM can be fitted through coxph function of package survival by [201].

In this analysis, the truncation parameters K_m, representing the number of FPCs to be considered for each event m, are chosen through a 10-fold cross validation procedure to select the best set of covariates among patients' baseline characteristics $\boldsymbol{\omega}_i$ and scores $f_{ik}^{(m)}$, according to the highest Concordance Index [151].

The entire procedure may be resumed in four steps, as shown in Figure 3.2:

- 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.

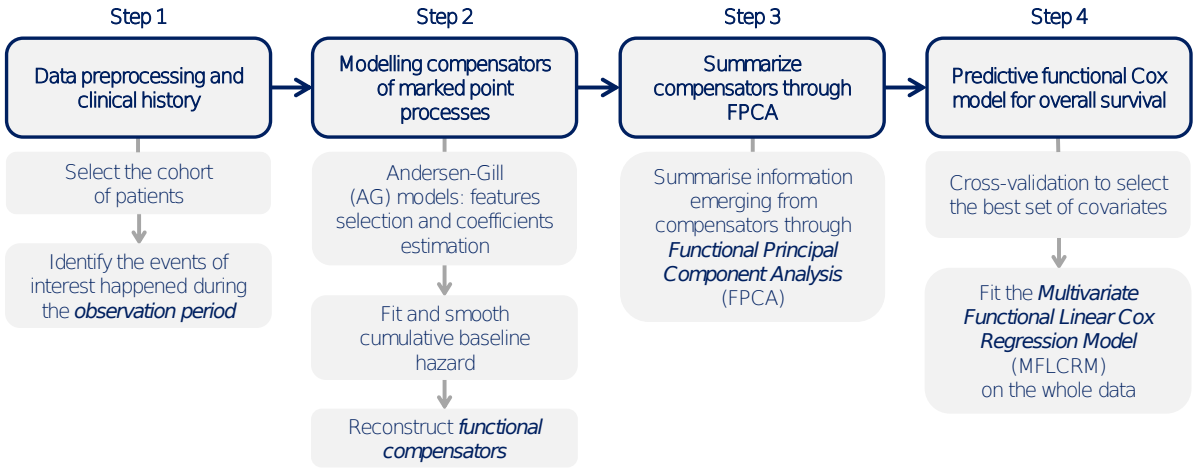

Figure 3.2. Summary of the entire methodological procedure presented in Section 3.2.

3.3. Data application

We now proceed with the application of the methodology described in Section 3.2/Figure 3.2 to the administrative database of *Lombardy Region*, in order to study how re-hospitalizations and multiple drugs consumption processes affect overall survival in HF patients. R source code is available as Supporting Information of [189].

3.3.1. 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 overall 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 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 $\mathcal{M} = \{m : ACE, BB, AA, HF\,hosp\}$. In particular, we selected only events within the 1-year *observation period* (censoring time $\tau = T_0^*$). For each patient $i \in \{1, \ldots, n = 4,541\}$, repeated events of process m were modelled as a marked point process $N_i^{(m)}(t)$, with *jump times* $t_{i,j}^{(m)}$ equal to event times (i.e., date of j-th admission in hospital or date of j-th drug purchase) and *jump marks* $\mathrm{w}_{i,j}^{(m)}$ equal to the length of stay in hospital or the duration of drug coverage respectively, where $j \in \left\{0, 1, \ldots, N_i^{(m)}(\tau)\right\}$. Figure 3.3 shows the counting processes $N_i^{(m)}(t)$ 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. 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 4 purchases (IQR = [0;8]), and 2,890 patients (63.6%) purchased BB at least once with a median of 4 purchases (IQR = [0;7]), where the median number of events m at time τ is given by $median_{i\in\{1,\ldots,n\}} N_i^{(m)}(\tau)$. 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 no purchases (IQR = [0;4]) and 2,699 patient (59.4%) were re-hospitalized due to HF, with a median of 1 HF hospitalization (IQR = [0;2]).

In order to proceed with the analyses, we reformatted the administrative data as explained in Appendix A.1.

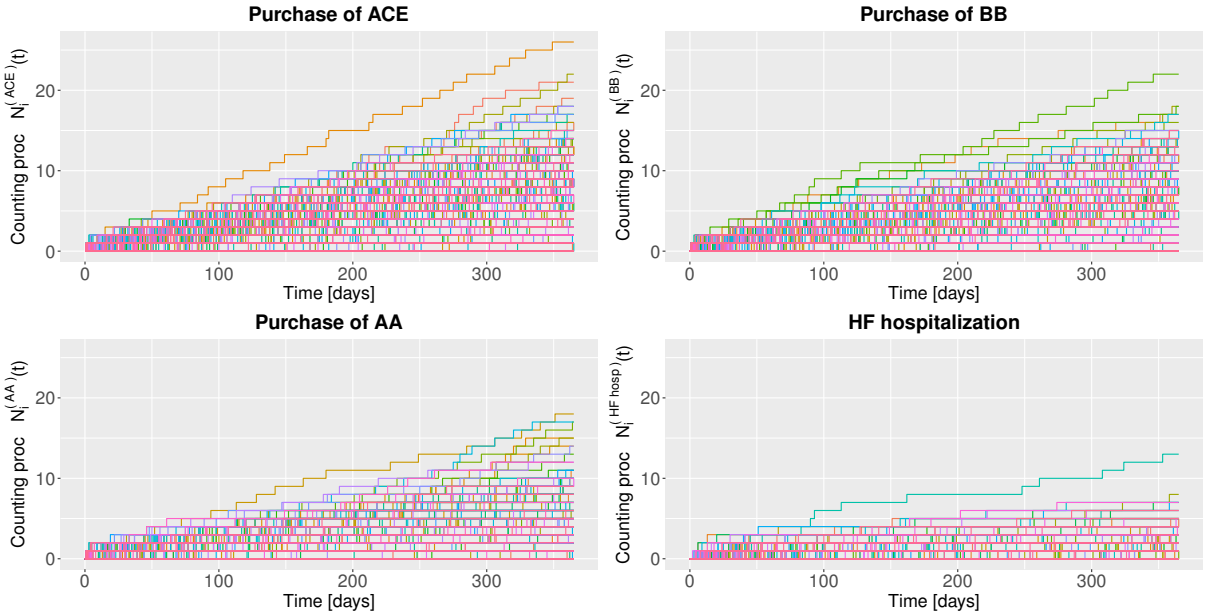

Figure 3.3. Representation of counting processes $N_i^{(m)}(t)$ related to purchases of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and of HF hospitalizations (bottom-right panel) during the observation period for a sample of 500 HF patients belonging to the administrative database. Each non-decreasing step function is related to a different patient.

3.3.2. 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 3.2.1. For each process $m \in \{ACE, BB, AA, HF\,hosp\}$, we first select the best set of features for the AG model for recurrent events in Equation (3.3) 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 [73]. Finally, we reconstruct the compensator trajectories as functions of the estimated coefficients and of the smoothed estimate of the cumulative baseline hazard through Equation (3.5).

Features selection and coefficients estimation

For each process $m \in \{ACE, BB, AA, HF\,hosp\}$, we used as covariates $\mathbf{z}_i^{(m)}(t)$ of patient i: the time-dependent variable *enum* which indicates the number of events related to process m occurred in the past and the time-dependent variable *marks* representing the sum of the corresponding marks. Also the logarithmic transformations (shifted away from 0) of the same variables, i.e., $\log(enum{+}1)$ and $\log(marks{+}1)$, 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 $\mathbf{x}_i^{(m)}(t)$. In particular, for each process m 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 A.2 for details). Once covariates were selected, we fitted four AG models in Equation (3.3),

one for each process m, using the selected features on the entire dataset to estimate coefficients $\hat{\boldsymbol{\beta}}^{(m)}$ and $\hat{\boldsymbol{\gamma}}^{(m)}$.

In Table 3.1 selected features, hazard ratios and corresponding 95% CI are reported. Among all the models tested through the cross-validation procedure, features related to *enum*, *marks* and their interaction were selected and their coefficients were always significantly different from 0. In particular, the procedure selected the original features for HF hospitalization process ($m = HF\,hosp$) and their logarithmic transformations for drug purchases ($m \in \{ACE, BB, AA\}$). 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. 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. HRs related to the number of past events *enum* and to the sum of the past marks *marks* were greater than 1. This could be interpreted as a "self-exciting" behaviour: many events (drug purchases or hospitalizations) in the past and higher marks (the purchase of big quantities of drug or having spent longer periods in hospital) both increase the risk of a new event. HR related to the interaction terms were lower than 1, meaning that (i) in case of the same number of events, the increase in the risk of experiencing a new event is softened by higher *marks*, or (ii) in case of the same cumulative marks, it is softened by an higher number of events *enum*. Furthermore, males [HR > 1] were more likely to buy medications or being re-hospitalized than females, except for AA purchases [HR < 1], and elder patients were more likely to be re-hospitalized than younger ones [HR > 1].

Table 3.1. Selected features, Hazard Ratios (HRs) and corresponding 95% Confidence Intervals (CIs) of the AG models for recurrent events for the stochastic processes describing the purchase of ACE inhibitors, BB agents, AA agents and the HF hospitalizations.

Process m	**Selected features**	**HR**	**[2.5; 97.5]% CI**
ACE	*gender (Male)*	1.0586	[1.0309; 1.0871]
	$\log(enum + 1)$	4.5271	[4.1674; 4.9178]
	$\log(marks + 1)$	1.1026	[1.0862; 1.1192]
	$\log(enum + 1) \times \log(marks + 1)$	0.9148	[0.9033; 0.9265]
BB	*gender (Male)*	1.0612	[1.0333; 1.0898]
	$\log(enum + 1)$	5.4270	[5.1195; 5.7529]
	$\log(marks + 1)$	1.1404	[1.1206; 1.1606]
	$\log(enum + 1) \times \log(marks + 1)$	0.8332	[0.8213; 0.8454]
AA	*gender (Male)*	0.9435	[0.9073; 0.9811]
	$\log(enum + 1)$	9.8781	[8.6116; 11.3310]
	$\log(marks + 1)$	1.2023	[1.1722; 1.2332]
	$\log(enum + 1) \times \log(marks + 1)$	0.7780	[0.7561; 0.8005]
$HF\ hosp$	*age*	0.9957	[0.9934; 0.9979]
	gender (Male)	1.1510	[1.0854; 1.2207]
	enum	1.4319	[1.3809; 1.4848]
	marks	1.0083	[1.0051; 1.0116]
	$enum \times marks$	0.9976	[0.9968; 0.9985]

Fit and smooth cumulative baseline hazard

Once we estimated the coefficients $\hat{\boldsymbol{\beta}}^{(m)}$ and $\hat{\boldsymbol{\gamma}}^{(m)}$ of each AG model for recurrent events of type m, we computed the estimated cumulative baseline hazards $\hat{\Lambda}_0^{(m)}(t)$ using the Breslow estimator. We smoothed them through the use of constrained B-splines [73] with increasing monotone constraints and no roughness penalties. In particular, we used 20 knots for the B-spline basis and we assumed that they took value 0 at time $t = 0$.

Figure 3.4 shows both the estimates obtained with the Breslow estimator $\hat{\Lambda}_0^{(m)}(t)$ (dashed blue lines) and the corresponding smoothed estimates $\tilde{\Lambda}_0^{(m)}(t)$ (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 $\forall m \in \mathcal{M}$ we obtained monotonically increasing estimates $\tilde{\Lambda}_0^{(m)}(t)$ of the cumulative baseline hazards with $\tilde{\Lambda}_0^{(m)}(0) = 0$.

Reconstruct compensators

At this point, we could reconstruct the trajectories of the compensators $\hat{\Lambda}_i^{(m)}(t)$ of the four considered stochastic processes for all the patients, exploiting Equation (3.5). The trajectories of compensators $\hat{\Lambda}_i^{(m)}(t)$ constitute our functional data. Figure 3.5 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 $\hat{\Lambda}_i^{(m)}(t)$ are monotonically non-decreasing and take value 0 at time $t = 0$, as did the smoothed baseline cumulative hazards $\tilde{\Lambda}_0^{(m)}(t)$. For each patient

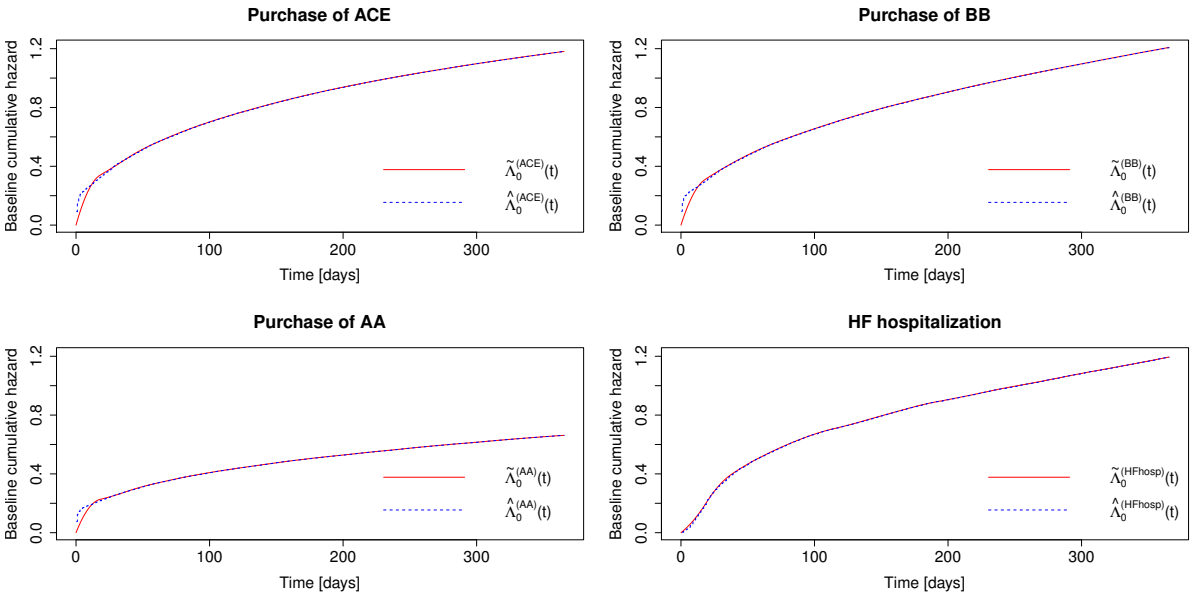

Figure 3.4. Cumulative baseline hazards of the Cox models for recurrent events describing the stochastic processes of purchases of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and of HF hospitalizations (bottom-right panel), fitted with the Breslow estimator $\hat{\Lambda}_0^{(m)}(t)$ (dashed blue lines) and smoothed $\tilde{\Lambda}_0^{(m)}(t)$ according to the procedure described in [21] (solid red lines).

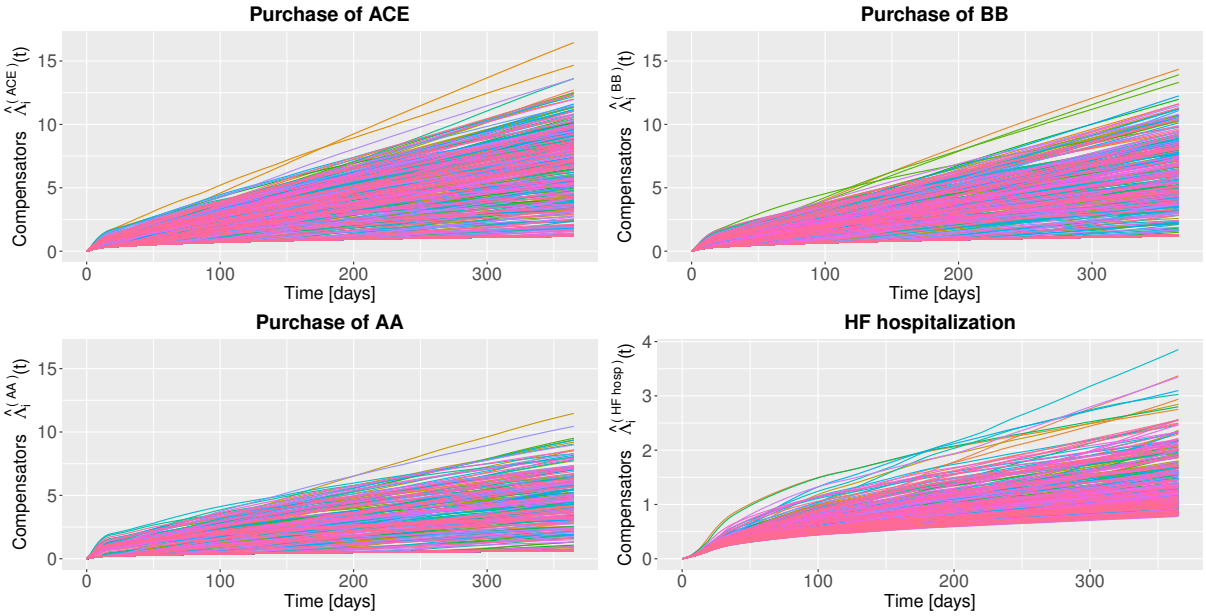

Figure 3.5. Compensators $\hat{\Lambda}_i^{(m)}(t)$ of the marked counting processes of purchases of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and of HF hospitalizations (bottom-right panel) fitted using Equation (3.5) for a sample of 500 HF patients belonging 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.

i, the compensator curve $\hat{\Lambda}_i^{(m)}(t)$ represents the expected number of events by time t given the covariates, i.e., the dynamic evolution of the events risk. This means that for a patient with a higher curve the cumulative risk of a new event (i.e., drug purchases or re-hospitalizations) is higher over time compared to a patient with a less steep curve. The large variability of the compensators across different patients reflects the variability of the realizations of their recurrent events times and marks.

Finally, we had to check for adequate fitting of the procedure. In order to do so, for each process of interest, we plotted the residuals evaluated in the whole observation period and we checked graphically that their means $\bar{M}^{(m)}(t)$ were approximately equal to 0. Figure 3.6 show the fitted residuals $\hat{M}_i^{(m)}(t)$ 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 $\bar{M}^{(m)}(t)$, computed using all the $n = 4,541$ patients. From the figure we observed that the time-varying means were approximately constant lines equal to zero for all the considered processes. Hence, we might conclude that we succeeded in fitting the compensators of the stochastic processes.

For each patient $i \in \{1, \ldots, 4,541\}$, we ended up with a four-variate time-varying data given by the compensator trajectories $\left\{\hat{\Lambda}_i^{(m)}(t)\right\}_{h \in \mathcal{M}}$ with $\mathcal{M} = \{ACE, BB, AA, HF\,hosp\}$, which could be though of as positive non-decreasing L^2-functions over the temporal domain $[T_0; T_0^*]$.

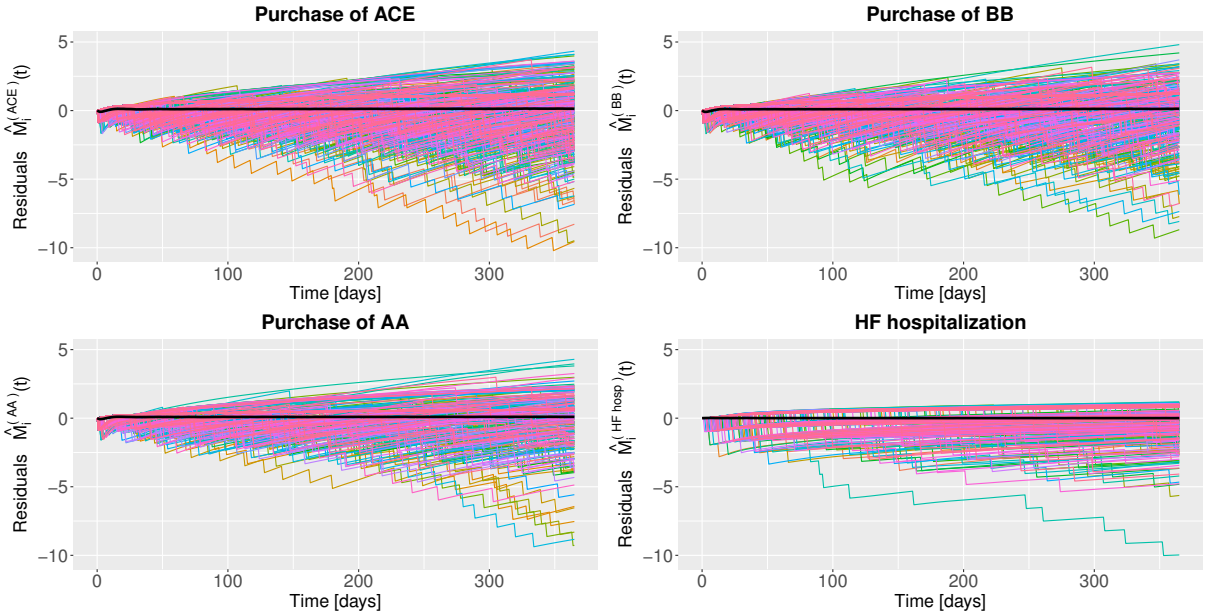

Figure 3.6. Residuals $\hat{M}_i^{(m)}(t)$ of the compensators of the stochastic process describing the purchase of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and of HF hospitalizations (bottom-right panel) for a sample of 500 HF patients belonging to the administrative database, computed according to Equation (3.6). Each line is related to a different patient. Solid black lines represent the temporal average residual curve $\bar{M}^{(m)}(t)$ computed using all the $n = 4,541$ patients.

3.3.3. Step 3: Summarize compensators through Functional Principal Component Analysis

Once we computed the functional trajectories of the compensators $\hat{\Lambda}_i^{(m)}(t)$, we performed Functional Principal Component Analysis (FPCA) [162] in order to summarise information emerging from the time-varying compensators to a finite set of covariates while losing 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 m we observed that two Principal Components (PCs) were enough to have a L^2-reconstruction error lower than 1%.

Figure 3.7 and Figure 3.8 show results of FPCA on functional compensators and are related to first and second PCs, respectively. 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). Top panels show that first and second PCs, i.e., $\xi_1^{(m)}(t)$ and $\xi_2^{(m)}(t)$, across the four processes types $m \in \{ACE, BB, AA, HF\ hosp\}$ have similar shapes. Bottom panels report the plots of compensators as perturbation of the mean [162]. In particular, the black lines constitute the average compensators curves $\mu^{(m)}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{\Lambda}_i^{(m)}(t)$, also denoting subjects with null FPC scores. Red plus and blue minus curves represent the perturbations $\mu^{(m)}(t) \pm c_k \sqrt{\nu_k^{(m)}}\xi_k^{(m)}(t)$ (red '+' and blue '−' respectively), where $\nu_k^{(m)}$ is the eigenvalues related to the k-th component and c_k are constants chosen in order to let the values lie within one ($c_k = 1$) or three ($c_k = 3$) standard deviations (i.e., square roots of $\nu_k^{(m)}$).

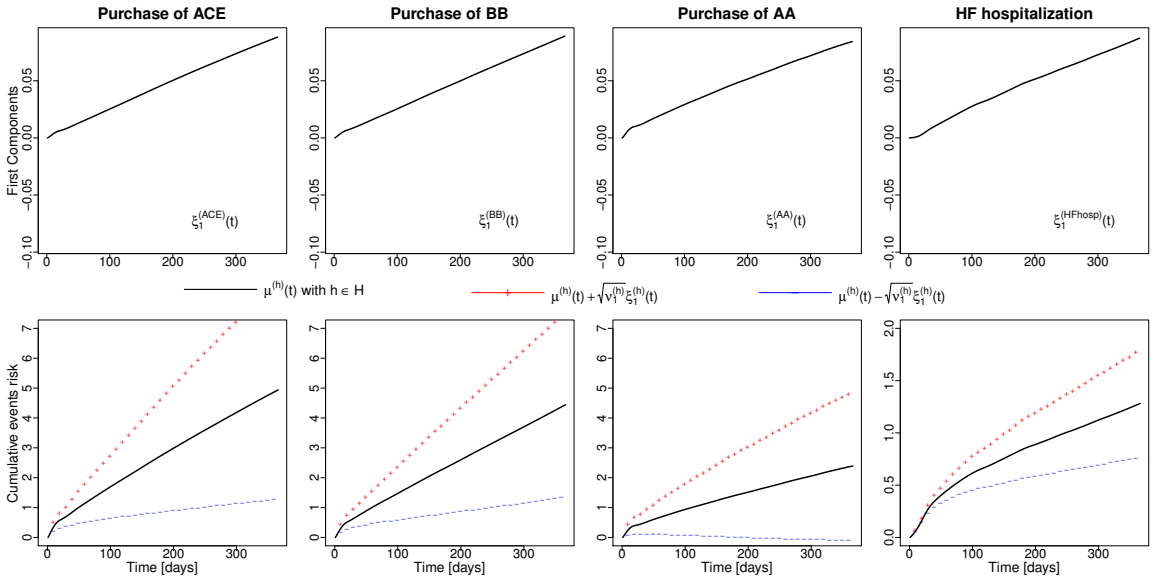

Figure 3.7. First functional Principal Components (PCs) of the compensators of the stochastic processes describing the purchase of ACE (first column), BB (second column), AA (third column) and HF hospitalization (fourth column). Upper panels show the first PCs $\xi_1^{(m)}(t)$ with $m \in \mathcal{M} = \{ACE, BB, AA, HFhosp\}$. Lower panels report the average compensators curves $\mu^{(m)}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i^{(m)}(t)$ (black lines) and $\mu^{(m)}(t) \pm \sqrt{\nu_1^{(m)}}\xi_1^{(m)}(t)$ (red '+' and blue '−' respectively) where $\nu_1^{(m)}$ are the eigenvalues related to the first components. 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.

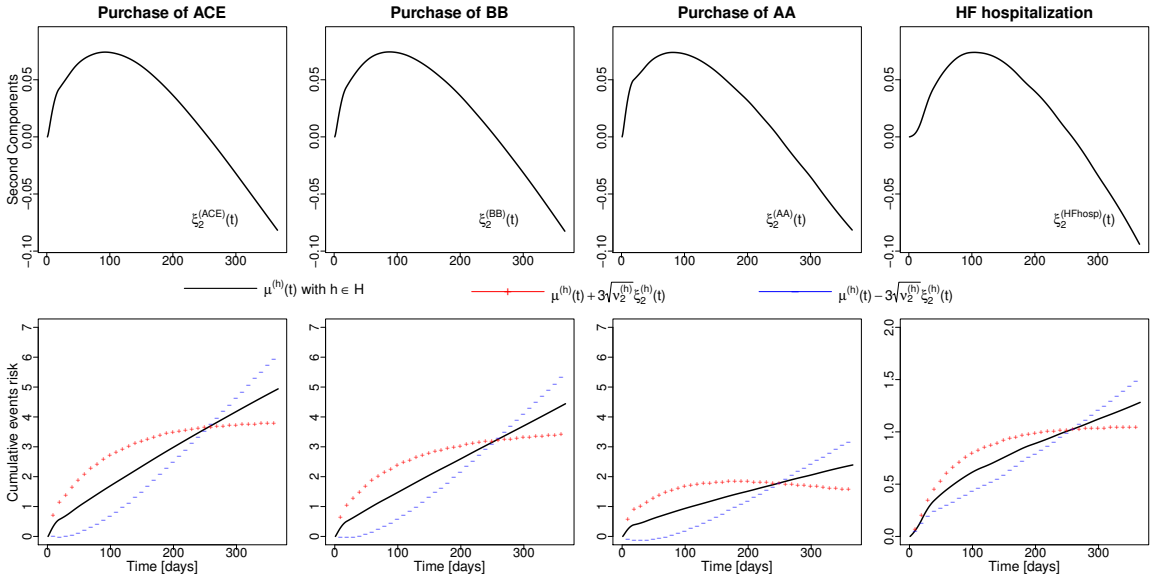

Figure 3.8. Second functional Principal Components (PCs) of the compensators of the stochastic processes describing the purchase of ACE (first column), BB (second column), AA (third column) and HF hospitalization (fourth column). Upper panels show the second PCs $\xi_2^{(m)}(t)$ with $m \in \mathcal{M} = \{ACE, BB, AA, HFhosp\}$. Lower panels report the average compensators curves $\mu^{(m)}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i^{(m)}(t)$ (black lines) and $\mu^{(m)}(t) \pm 3\sqrt{\nu_2^{(m)}}\xi_2^{(m)}(t)$ (red '+' and blue '−' respectively) where $\nu_2^{(m)}$ are the eigenvalues related to the second components. 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.

In Figure 3.7 we observe that the first components $\xi_1^{(m)}(t)$ distinguish patients with different events risks. In particular, positive scores related to the first PC (red plus curve) reflect higher curves with respect to negative ones (blue minus curve), indicating that a patient with a high score on the first component is likely to experience more events than a patient with a low score. Figure 3.8 shows that the second components $\xi_2^{(m)}(t)$ distinguish patients with different time distribution of the events. In particular, a patient with a high score (red plus curve) 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 than a patient with a low score (blue minus curve), indicating different events timing.

3.3.4. Step 4: Predictive functional Cox model for overall survival

At this point we wanted to quantify the association between time-varying processes and patients' overall survival through a Multivariate Functional Linear Cox Regression Model (MFLCRM) in Equation (3.10). First, we applied 10-fold cross validation to select the best set of covariates among possible combinations of patients' baseline characteristics *age*, *gender* and truncation parameters K_m of FPCA with $m \in \{ACE, BB, AA, HF\,hosp\}$, according to the highest median Concordance Index [151]. The selected MFLCRM, given by

$$\begin{aligned} h_i\left(t|\boldsymbol{\omega}_i, \left\{\Lambda_i^{(m)}\right\}_{m\in\mathcal{M}}\right) = h_0^*(t)\exp\big\{&\theta_1 age_i + \theta_2 gender_i + \\ &\alpha_1^{(ACE)} f_{i1}^{(ACE)} + \alpha_1^{(BB)} f_{i1}^{(BB)} + \alpha_1^{(AA)} f_{i1}^{(AA)} + \\ &\alpha_1^{(HF\,hosp)} f_{i1}^{(HF\,hosp)} + \alpha_2^{(HF\,hosp)} f_{i2}^{(HF\,hosp)}\big\}, \end{aligned} \tag{3.11}$$

was then fitted on the whole data to quantify the association between functional compensators and overall survival.

Table 3.2 reports the summary of fitted model (3.11). All the covariates resulted statistically significant at confidence level 5%, except for $f_1^{(AA)}$. Elder patients coherently have a higher risk of dying [HR = 1.067] and being a male corresponds to 1.25-times faster experience of the event. The HR relative to the scores of the first PCs for ACE and BB processes, i.e., $f_1^{(ACE)}$ and $f_1^{(BB)}$, 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 $f_1^{(HF\,hosp)}$ 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, $f_2^{(HF\,hosp)}$ is strongly significant with HR = 0.773 < 1 (95% CI = [0.725; 0.825]). 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.

Table 3.2. Hazard ratios (HRs) along with 95% Confidence Intervals (CIs) of the final multivariate functional linear Cox regression model (MFLCRM) for overall survival fitted on the whole cohort using the covariates selected through 10-fold cross-validation.

Covariates	**HR**	**[2.5; 97.5]% CI**	p-value
gender (Male)	1.2540	[1.1080; 1.4194]	< 0.001
age	1.0670	[1.0592; 1.0748]	< 0.001
$f_1^{(ACE)}$	0.9977	[0.9963; 0.9992]	0.003
$f_1^{(BB)}$	0.9964	[0.9945; 0.9982]	< 0.001
$f_1^{(AA)}$	1.0006	[0.9986; 1.0026]	0.550
$f_1^{(HF\,hosp)}$	1.0157	[1.0049; 1.0266]	0.004
$f_2^{(HF\,hosp)}$	0.7733	[0.7251; 0.8247]	< 0.001

3.4. Final remarks

In this chapter, a novel approach to reconstruct the compensators of suitable marked point processes of interest as time-varying covariates has been proposed. This approach was exploited to enrich information to be included 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 [21], allowing the counting processes to depend on their marks and moving towards the multivariate setting.

From the study on the administrative database of *Regione Lombardia*, 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 consume 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. To the best of our knowledge, our approach represents the first attempt in literature of merging potential of FDA, recurrent events theory and survival analysis.

Thanks to its flexibility, the proposed methodology could be extended and generalized to many different settings, adapting the procedure to the different biological and clinical aspects of the specific application. In particular, alternative ways to get the trajectories

related to the L^2 functional compensators could be considered. The AG model for recurrent events in Equation (3.3) represents only one of the possible approaches to express the conditional intensity function. Alternative methods or distributions for the marks could be considered according to many factors, among others number of events, relationship between subsequent events and intrinsic characteristics of the processes. For example, our case study was also analysed considering a shared gamma-frailty model [175], in which the intensity function in Equation (3.3) was assumed to partly depend on an unobservable random variable that acted multiplicatively on it. In that case, the compensator trajectories were expressed as functions of estimated coefficients, smoothed cumulative baseline hazard and estimated frailties. Obtained results were comparable to the ones shown in the paper in terms of both estimated effects on patients' overall survival and clinical implications. In case of a limited number of events, stratified Cox models for recurrent events, such as the Prentice-Williams-Peterson [156] or the Wei-Lin-Weissfeld [213] model, could be used modifying Equation (3.4) in order to consider the proper strata of the cumulative baseline hazards. As a further alternative, in case of multiple events with cyclical occurrence, the best choice would be to account for seasonality in the model through cyclic functions, such as in the rate model with multiple event types by [196]. In that case, the L^2 functions could be obtained by smoothing the cumulative rate functions. Therefore, thanks to its adaptability, the presented methodology can be generalized and applied to the study of many different pathologies characterized by complex data sources.

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. In the present work, the final choice for a pre-defined *observation period* of 1 year after the index hospitalization was made under clinical indication, once performed a sensitivity analysis to evaluate the robustness of our method using two different clinically acceptable periods of 6 months and 1 year whose results led to common conclusions. From a modelling point of view, the assumption of independence between jump times and marks in Equation (3.2) could in general be relaxed, but this could lead to several issues [132]. In fact, considering re-hospitalization process, it is difficult to conjecture a mathematical relation of length of stay in hospital with time of hospitalization. The same is valid for drug purchases. Moreover, there could be computational limitations in terms of modelling a temporary dependence. Since dependence is harder to be dealt with due both to computational and modelling issues, we limited our analysis to the independence case, which was considered a clinically acceptable assumption. The development of proper statistical tools to test this hypothesis can be of great help for our topic, since existing techniques for testing independence are rather complex to apply and customize to the current context. Furthermore, FPCA was performed in $L^2[T_0; T_0^*]$ 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, as in the previous chapters. First, not being able to ascertain whether the patient was currently consuming the dispensed drug remains the major limitation of using drug purchases as a proxy for drug intake, which is the only possible way through administrative data. Second, the use of theoretical Defined Daily Dose (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 [187, 220]. In future analysis, 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. Since administrative data are collected with no clinical question in mind and mainly for managerial and economic purposes [89], the validity of using these kind of data is critically dependent on the reliability of the data [115, 180, 90]. Nevertheless, in the last decade significant improvements have been gained through administrative data sources, and their use in clinical biostatistics has become an accepted practice, representing a great challenge for statistics and related modelling [90].

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. For example, it could be of clinical interest to further extend the study of polypharmacy by considering also drug-drug interaction terms, which could be included in the model through compensator-compensator interaction terms. However, a compensator-compensator interaction term involves the modelling of bivariate (or more in general multivariate) marked point processes, which represents a non-trivial task beyond the scope of the present work.

In summary, the presented methodology, involving database integration, marked point process modelling of critical events and FDA techniques, enabled a manageable and relatively simple analysis of the results, describing complex dynamics in an easily interpretable form. Both parts of the procedure represent flexible approaches that can be used to quantify personal behaviours and to investigate their effect of on survival. On one hand, the developed marked point processes formulation could be applied in many different clinical contexts characterized by recurrent occasions. On the other, the use of FPCA to extract additional information contained in the functions and to include them into a MFCLRM can be easily applied to all settings where the time-varying characteristics of interest are adequately reconstructed by FDA, as we will see in Chapter 4 for the case of biomarkers and chemotherapy dose in osteosarcoma patients. Its possible generalization to many different contexts, combined with cooperation with medical staff, could therefore lead to improvements in the definition of useful tools for health care assessment and treatment planning.

A. Appendix to Chapter 3

A.1. 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 3.3 – Step 1 of the procedure), we had to reformat the administrative data building four different datasets, one for each process $m \in \mathcal{M} = \{m : ACE, BB, AA, HF\,hosp\}$, in the form required by coxph function for recurrent events of survival R package by [201]. Table 3.3 shows an example of reformatted dataset related to ACE purchases process for a hypothetical patient with four ACE events during the observation period. In the Table, *start* indicates the time of the patient's previous event (equal to 0 for the index date), *stop* is the time of the current event (equal to 365.5 if it is the censoring event), *status* is the event indicator (0 if censored, 1 otherwise), *enum* is the number of events related to process m occurred in the past and *marks* is the sum of the corresponding marks. In particular, the choice to consider the time limit at 365.5 was made in order to not have events at censoring time $t = 365$. Moreover, it could also happen that a patient i experienced the first event of type m during the index day. In that specific case, we considered jump time equal to 0.5, i.e., $t_{i,1}^{(m)} = 0.5$, in order to not have events at time $t = 0$. Hence, for each process m we ended up with a long-format dataset with multiple rows for each patient (specifically the number of patient's events of type m during the observation period plus one). In particular, in the first row of each patient *enum* and *marks* are always 0 and in the last one *status* is always equal to 0.

Table 3.3. Example of reformatted dataset related to ACE purchases process for a hypothetical patient with four ACE events during the 1-year *observation period*.

ID	*start*	*stop*	*status*	*gender*	*age*	*enum*	*marks*
id	0	0.5	1	*Female*	87	0	0
id	0.5	83	1	*Female*	87	1	56
id	83	91	1	*Female*	87	2	70
id	91	215	1	*Female*	87	3	98
id	215	365.5	0	*Female*	87	4	112

A.2. Mean Absolute Martingale Residual

Given two or more Andersen-Gill (AG) models for recurrent events in Equation (3.3) 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 (3.6): smaller residuals correspond to a greater predictive power of the model. Therefore, to compare models fitted with different features, for each process m we would like to use the

Mean Absolute Martingale Residual (MAMR):

$$MAMR^{(m)} = \sum_{i=1}^{n} \frac{\int_0^T |\hat{M}_i^{(m)}(s)|ds}{T} \tag{3.12}$$

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

To correctly compute the MAMR, we should first compute the compensators using Equation (3.5) 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:

$$\widehat{MAMR}^{(m)} = \frac{1}{\sum_{i=1}^{n} n_i^{(m)}} \sum_{i=1}^{n} \sum_{j=1}^{n_i^{(m)}} |\hat{\hat{M}}_i^{(m)}(t_{i,j}^{(m)})| \tag{3.13}$$

where i and m are respectively the patient and event indices, $\hat{\hat{M}}_i^{(m)}(t)$ is the residual obtained by fitting the compensator without smoothing the baseline hazard, i.e., using $\hat{\Lambda}_0^{(m)}(t)$ instead of $\tilde{\Lambda}_0^{(m)}(t)$ in Equation (3.5), $n_i^{(m)}$ is the total number of events of type m experienced by the i-th patient and $t_{i,j}^{(m)}$ is the time instant in which patient i experienced the j-th event of type m.
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.