Dynamic clustering of hazard functions: an application to disease progression in chronic heart failure

Francesca Ieva[1] · Anna Maria Paganoni[2] · Teresa Pietrabissa[2]

Received: 20 July 2015 / Accepted: 25 January 2016 / Published online: 4 February 2016

Abstract We analyse data collected from the administrative datawarehouse of an Italian regional district (Lombardia) concerning patients affected by Chronic Heart Failure. The longitudinal data gathering for each patient hospital readmissions in time, as well as patient-specific covariates, is studied as a realization of non homogeneous Poisson process. Since the aim behind this study is to identify groups of patients behaving similarly in terms of disease progression and then healthcare consumption, we conjectured the time segments between two consecutive hospitalizations to be Weibull distributed in each hidden cluster. Adding a frailty term to take into account the within subjects unknown variability, the corresponding patient-specific hazard functions are reconstructed. Therefore, the comprehensive distribution for each time to event variable is modelled as a Weibull Mixture. We are then able to easily interpret the related hidden groups as healthy, sick, and terminally ill subjects.

Keywords Heart failure · Survival analysis · Proportional hazards model · Frailty models

✉ Anna Maria Paganoni
anna.paganoni@polimi.it

Francesca Ieva
francesca.ieva@unimi.it

Teresa Pietrabissa
teresa.pietrabissa@mail.polimi.it

[1] ADAMSS Center & Department of Mathematics
"F. Enriques", Università degli Studi di Milano, via Saldini 50, 20133, Milan, Italy

[2] MOX - Department of Mathematics, Politecnico di Milano, via Bonardi 9, 20133, Milan, Italy

1 Introduction

Heart Failure (HF) is a physiological state in which the result is a lack of blood flow to the body. Often clinicians refer to heart failure as Chronic Heart Failure (CHF), as to identify patients symptomatic of a long duration disease. Chronic heart failure can be caused by multiple factors: rheumatic heart disease, valve disorder, diastolic/systolic dysfunction, cardiomyopathy, hypertension; moreover, heart failure is diagnosed through a variety of signs, like increased rate of breathing, pulmonary edema, pleural effusion, nocturia, peripheral edema and more [9].

HF is the leading cause of hospitalisation in people older than 65 years. A 2010 update from the American Heart Association (AHA) estimated that there were 5.8 million people with HF in the United States in 2006 (see [17, 20], among others). There are an estimated 23 million people with HF worldwide, often accounting for a total medical expenditure that is much greater than any other disease. Despite dramatic improvement in outcomes with medical therapy, admission rates following HF hospitalization remain high [24], with around 50 % of patients readmitted to hospital within 6 months of discharge, see [6, 14, 16]. A reduction in readmission rates might simultaneously reduce costs and improve quality of care. Anyway, whenever reduction of admissions is not a proper target to aim for, accurate modelling and prediction of the disease dynamics may enable a more efficient planning and management of resources.

In the application of interest, we deal with data coming from the administrative database of Northern Italy regional district (Regione Lombardia). In the Lombardia district, the HF incidence over the last decade ranged between 25,000 and 30,000 cases per year in a population of 9.7 million inhabitants (ISTAT [13]). This unavoidably leads to a huge number of hospitalizations, with consequent problems related to management and organizational issues and, last but not least, considerable costs. Within the Italian healthcare regulation system, every hospital admission is recorded in an administrative datawarehouse called SDO (Scheda di Dimissione Ospedaliera, i.e., hospital discharge paper) database, in order to enable hospitals to be refunded for the services they provide to the patients.

The growing interest in the study of administrative data with epidemiology and health-care management purposes is testified by the high number of works recently published both in statistical and epidemiological journals. Our paper is closely connected to the study of patients hospital readmission process (see [1, 15] among others). In fact readmission rate is broadly accepted as a performance indicator and cost driver in HF and not only. So readmission predictive models are of great interest both to support programs aimed at reducing avoidable readmissions and to assess hospitals quality of care. For example, in [25] multivariate generalized regression models are extensively used to identify and target high-risk patients; in [26] the problem of avoidable readmissions is faced with tree based clustering algorithms and [4] studies hospitals performances exploiting multiple outcomes multivariate models. In this wide stream of research problems the main novelty of our paper is in modeling administrative data (i.e. patients histories) in terms of hospital admissions as trajectories of a non-homogeneous counting process. In particular we aim at moving from modeling hospital readmissions of patients to predicting evolution of disease progression in different subgroups of subjects. This prediction tool has twofold advantages: helping healthcare decision makers optimizing resources allocation in the context of a pathology (like HF) with a high impact in the consumption of healthcare resources on one hand, and, on the other hand, doing so according to the predicted risk profile of the patients. Statistically speaking, there are several methodological approaches to the modelling of times to multiple events per subject. For example, if the interest lies in both the time to event and in the nature of the event, the occurrence of subsequent events may be investigated by multi-state modelling approach (see, among others, [2, 10] and references therein) or carrying out a hazard-based analysis, focusing the modelling effort on the waiting/gap times between subsequent events (see [7] for an appraisal of modelling approaches to counting processes).

For the problem of interest, both the approaches mentioned before can be considered. The choice depends on the final aim of the analysis. In [11] an example of a multi-state modelling strategy for the joint analysis of outcomes and hospital admissions in CHF patients is proposed. In that case the aim was to show a flexible approach, able to capture important features of admission-discharge dynamics such as multiple ordered events and the competing risks of death and hospitalisation, in a novel application based on data arising from the administrative database of Regione Lombardia. In the present case we still focus on patients hospitalizations, modelling them like trajectories of a non-homogeneous counting process. The inter-times between hospitalizations are modelled as independent, parametric, not necessarily identically distributed random variables. This leads to the estimation of hazard functions not coming from i.i.d. distributions of interval times. To catch the heterogeneity of the observed population we assume the presence of K latent groups of patients behaving differently in terms of disease progression. This leads to a mixture model for each inter-hospitalization time. Moreover, the paper proposes a simulation strategy to construct a sample of trajectories of the counting process associated to each different group, together with the corresponding K hazard function templates. Then hazard function trajectories of the cumulative process underlying the observed hospitalizations counting process are computed using non parametric techniques and accounting for overdispersion due to subject-specific frailty and covariates. Finally, the modelling effort focuses on prediction. In fact, for a new patient we can compute the probability of belonging to the k-th cluster in an empirical-Bayesian way and consequently we can estimate the subject-specific hazard function. Once the evolution patterns are identified, the model becomes a tool for prediction of patient disease progression and then the corresponding healthcare consumption, providing a predictive tool to people in charge with healthcare governance.

The paper is organized as follows: we describe the data selection and inclusion criteria in Section 2, and we explain the details of the modelling approach in Section 3. Key results from applying these methods to the Lombardia HF admissions data are presented in Section 4. In Section 5 we end with a discussion of the main results and of challenges of using administrative data. All the analyses are carried out using R codes and environment [21]. Codes are available upon request to the authors.

2 Data and extraction criteria

Nowadays administrative databases play a central role in the evaluation of healthcare systems, because of their widespread diffusion and the real-time, low-cost information they provide (see [8, 27], among others). There is an increasing agreement among epidemiologists on the validity of diseases and intervention registries based on administrative databases (see, for example, [3, 5, 12, 28], and references therein). Therefore more and more frequently administrative data are used to address epidemiological issues in observational studies. The most critical issue when using administrative databases for observational studies is represented by the selection criteria of the population. In fact several different criteria may be used, and they will result in different images of prevalence or incidence of diseases. In the case of interest, we focus on patients affected by Heart Failure (HF). Concerning this pathology, since every hospital admission ends in a record collected in the administrative datawarehouse, the database of SDO (*Scheda di Dimissione Ospedaliera*, i.e., hospital discharge paper) has been used in order to identify HF episodes and related subsequent hospitalizations. In fact, the SDO database contains data for each hospitalization that a patient experiences along time, providing information both on patient features (in terms of sex, age) and on her/his hospitalization details (date of admission and discharge, diagnoses and procedures, type of admission, type of discharge, vital status at discharge, hospital of admission/discharge). The case study presented here concerns data arising from a project named "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction", funded by Ministry of Health and Regione Lombardia. We consider hospital discharges from 2000 up to 2012.[1] Possible date of death for each patient was linked from death registry by the institution that hosts the databases so that it is possible to estimate both in-hospital and long term survival time for each patient. Survival times are right censored to the end of the study (December, 31$^{\text{st}}$ 2012). The total number of patients included in the study is 251,451, corresponding to 482,701 events.

To conduct our analysis, we select a specific cohort. Patients that, during the observation period, experienced any cardiogenic shock (a life-threatening medical condition characterised by low blood pressure, rapid heartbeat and poor end-organ perfusion) have been discarded since they are considered by clinicians as a different subpopulation. Only patients older than 18 years at their first admission and whose Length Of Stay (LOS) was not null were included. Finally we selected patients with a maximum of five hospitalisations and whose first discharge happened either in 2006 or 2007, in order to have almost 5 years follow up. As a result, the dataset reduces to 56,505 events, related to 34,298 patients.

2.1 Data characteristics

In this section , we give a brief description of the selected cohort, in order to understand and later to interpret the results of our analyses.

Patients' age ranges between 18 and 104 years, with a mean of 77.25 years and a standard deviation of 11.06. In this cohort there is a majority of women (53.04 % vs 46.96 %). Women tend to be older than men, having a mean age of 80.09 years in contrast with men's mean age, equal to 74.04 years (Wilcoxon test, alternative: true location shift not equal to 0, p-value $< 2.2e - 16$).

The percentage of dying patients throughout the study time period is equal to 60.61 %, with a strong evidence of higher mortality rate for women than men (2-sample test for equality of proportions, alternative: men's less than women's death, p-value $< 2.2e - 16$).

3 Model specification

3.1 Patients clustering

Let T_i, $i = 1, ..., H$ be the random variables modelling the time between the i-th hospitalization and the following event, that could be the $(i+1)$-th hospitalization, the decease or the end of the study. We set $H = 5$, due to sparisty of data from patients with more than 6 admissions. We model T_i and the corresponding $f_{T_i}(t)$, $i = 1, ..., H$ as a Weibull mixture:

$$f_{T_i}(t) = \sum_{k=1}^{K} \pi_k f(t; \eta_{ki}, \gamma_{ki}), \tag{1}$$

where

$$f(t; \eta_{ki}, \gamma_{ki}) = \frac{\gamma_{ki}}{\eta_{ki}^{\gamma_{ki}}} \, t^{(\gamma_{ki}-1)} \, \exp\{-(t/\eta_{ki})^{\gamma_{ki}}\} \tag{2}$$

is the Weibull density, whose parameters change according to the number of the considered hospitalization i and the hidden group k. Hence this model has $2(K \times H) + K$ parameters. We set $K = 3$. This choice is reasonable, since in any cohort of patients there are three macro-groups: we can consistently name them healthy, sick and terminally ill, and we will later prove that this choice has a clear clinical meaning. The diagnosis of HF can be done at very early stages or at final ones, indeed. Moreover the choice of a Weibull mixture model seemed reasonable, also looking at the empirical

[1] Data have been requested and extracted by Lombardia Informatica (the company that manages the data warehouse of Regione Lombardia) within the research Strategic Program "Utilization of Regional Health Service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalized for heart failure". As one of the operative units working on this research project we can access to hospitalization data concerning only the time window 2000-2012.

distributions of inter-event times T_i (see Fig. 1). In the first panel of Fig. 1, also the presence of censored patients can be easily detected.

The hazard function in each hidden group k and each hospitalization i, is of the following form:

$$h_i(t; \eta_{ki}, \gamma_{ki}) = \frac{\gamma_{ki}}{\eta_{ki}^{\gamma_{ki}}}\, t^{(\gamma_{ki}-1)}. \tag{3}$$

When introducing the Proportional Hazards Model (PHM) into this framework, we can set the scale parameter η_{ki} to be equal to the exponential term in the PHM (see [18, 19]), obtaining:

$$h_i(t; \gamma_{ki}, \beta_{ki}) = \gamma_{ki}\, t^{(\gamma_{ki}-1)} \exp(\beta_{ki}). \tag{4}$$

From the hazard function (4) we are able to obtain the survival function, hence the resulting density of the Weibull proportional hazard model mixtures, recalling that $h(t|k,i) = \frac{f(t|k,i)}{S(t|k,i)}$.

The parameters π_k, β_{ki} and γ_{ki} that characterise the Weibull mixture (see Eqs. 1–4) are estimated through the EM algorithm proposed in [19]. So we obtain, for each group $k = 1, ..., K$ a baseline hazard function, hereafter $\lambda_0(t|k)$. Each patient $j = 1, \ldots, J$ is then assigned to one of the K clusters according to the mechanism detailed in [18].

3.2 Patient-specific hazard reconstruction

Once the analysed cohort has been divided into three subgroups, we can attempt a reconstruction of patient-specific hazard functions. We set the hazard function for patient j in cluster k as:

$$\lambda_j(t|k) = \lambda_0(t|k) v_j \exp(\boldsymbol{\beta}_k' \mathbf{Z}_j), \tag{5}$$

$$v_j \overset{\text{i.i.d.}}{\sim} log - Normal(0, \sigma_k^2). \tag{6}$$

This is known as the shared frailty model, where v_j is the frailty term for patient j and it is constant (*shared*) for all the events related to the j-th subject. The term frailty comes from medicine, referring to feeble people which are characterised by having an increased risk for morbidity and mortality. As a matter of fact, in frailty models the frailty term is introduced as a multiplicative random effect to estimate the mortality risk of an individual into a population. We will deal, among all possibilities, only with shared frailty models, as these are best suitable for our dataset. Shared frailty models depend on the idea that unities in the same cluster share the same frailty term: as we are dealing with longitudinal data, for us it will be that the events concerning the same patient will share the same frailty term. Moreover, we set the frailty term to have a log-Normal density law common to all patients in the same group k. Every v_j is then a realisation from the distribution in Eq. 6.

As mentioned before, the $\lambda_0(t|k)$ term in Eq. 5 is the baseline hazard function built upon the $h_i(t|k,i)$ of Eq. 4 for the k−th group. This is estimated non-parametrically according to [22, 23]. Finally we have introduced in the model some covariates $\mathbf{Z}_j$: *Age* (measured in years) of patient j at the beginning of i-th hospitalisation; *Intensive Therapy*, a boolean variable indicating whether patient j was recovered in the intensive therapy unit during the i-th hospitalisation; *Procedures* a boolean variable indicating whether at least one procedure among ICD (Implantable Cardioverter Defibrillator), CABG (Coronary Artery Bypass Grafting) and PTCA (Percutaneous Transluminal Coronary Angioplasty) was performed on patient j during the i-th hospitalisation; *Comorbidities*, a boolean variable indicating whether patient j has at least two comorbidities among renal disorder, tumors, anemia, liver disorder and others during the i-th hospitalisation. Note that all the included covariates are time-dependent. In particular, we standardised the age in order to avoid a computational overflow (see also [23]). Corresponding coefficients $\boldsymbol{\beta}_k$ are specific for the considered group k.

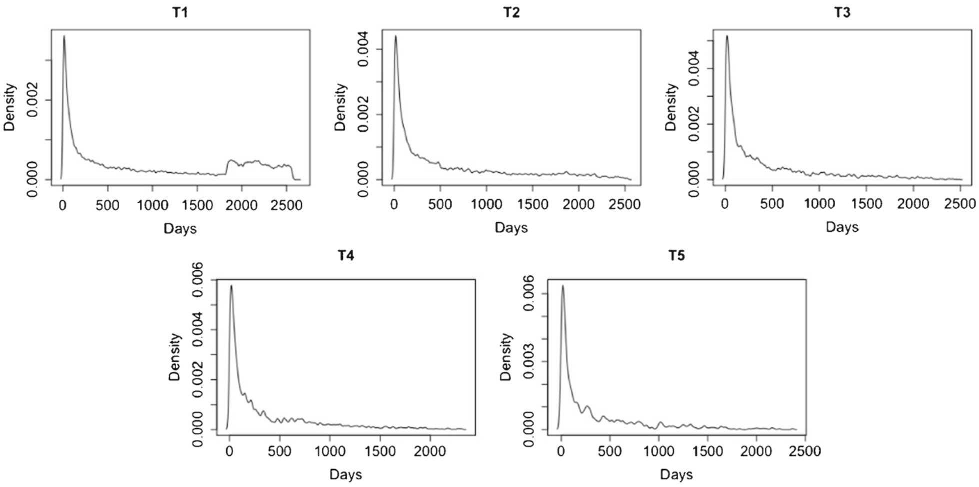


Fig. 1 Empirical densities of inter-event times T_i, $i = 1, \ldots, H$

Within this category of models we are especially interested in estimating the frailty variances σ_k^2, $k = 1, \ldots, K$, which represent the unknown variability that exists among patients in the same group. This may bring important information, together with patient-specific covariates, for predicting the evolution of patients' disease in each group.

3.3 Prediction of disease evolution

One of the most interesting aspects of the current work is the prediction of the behaviour of a new patient (j_{new}), diagnosed with heart failure.

We implemented the same algorithm presented above in order to assign the new patient to one of the three already determined clusters. Based only on information regarding the observed inter-event times of new patient j_{new}, we can compute the probability of belonging to one of the three clusters. Indicating this probability as $\nu(j_{new}|k)$, we evaluate it in a empirical-Bayesian way:

$$\nu(j_{new}|k) = \frac{p_k f_{j_{new}}(k)}{\sum_{k=1}^{K} p_k f_{j_{new}}(k)} \tag{7}$$

where p_k is the empirical probability of belonging to the k-th cluster ($p_k = n_k/N$, n_k being the cardinality of k-th cluster, and $N = n_1 + n_2 + n_3$), and where $f_{j_{new}}(k) = \prod_{i=1}^{H} \phi_i(j_{new}|k)$ is the full likelihood function for the new patient, supposing she/he belongs to the k-th cluster.

The contribution of the i-th hospitalisation of patient j_{new}, within the k-th cluster, to the full likelihood consists of the product between $\lambda_{j_{new}}(t_i|k)$ and the corresponding survival function $S_{j_{new}}(t_i|k)$:

$$\begin{aligned} \phi_i(t_{j_{new},i}|k) &= \lambda_{j_{new}}(t_{j_{new},i}|k) S_{j_{new}}(t_{j_{new},i}|k) \\ &= \frac{\gamma_{ki}}{\eta_{ki}^{\gamma_{ki}}} t_{j_{new},i}^{(\gamma_{ki}-1)} \exp(-(\frac{t_{j_{new},i}}{\eta_{ki}})^{\gamma_{ki}}). \end{aligned} \tag{8}$$

To account for those hospitalisations i such that $i > i_{max}(j_{new})$, where $i_{max}(j_{new})$ is the last one experienced by the considered patient, we use the model's parameter $\tau_i(k)$, i.e., the probability of having at least i hospitalisations when belonging to the k-th cluster. In so doing, we can compute the contribution $\phi_i(j_{new}|k)$ to the full likelihood over all possible hospitalisations, for $i = 1, \ldots, H$:

$$\phi_i(j_{new}|k) = \begin{cases} \phi_i(t_{j_{new},i}|k)\tau_i(k) \\ \text{if } j_{new} \text{ has at least } i \text{ hospitalisations,} \\ 1 - \tau_i(k) \\ \text{if } j_{new} \text{ has less than } i \text{ hospitalisations.} \end{cases} \tag{9}$$

At this point it is easy to compute $f_{j_{new}}(k)$, by multiplying $\phi_i(j_{new}|k)$ over i, hence obtaining the posterior probability $\nu(j_{new}|k)$ of belonging to the k-th cluster (see Eq. 7).

To evaluate the full likelihood function, we use the parameters obtained from the clustering of the original dataset, and estimated through the EM algorithm, i.e. the probability of having at least i hospitalisations when patient j belongs to the k-th cluster, and the parameters of the Weibull Mixture Model, η_{ki} and γ_{ki}.

4 Analysis of results

The great majority of patients in our cohort die before the end of the study due to very serious conditions. These patients, for the nature of their disease, have a relevant impact when analysing the entire cohort or when looking at groups of patients obtained as described in Section 3.1. Indeed, when attempting a clustering of patients, those with a higher mortality risk rate play a crucial role and in some cases they force a certain type of resulting division. This is why we will firstly analyse the obtained results when clustering the entire cohort. This will be the Step 1 of our analysis. Then we will compute again the clustering and reconstructing algorithm over subgroups of our cohort, every time removing those patients who died at i-th hospitalisation (i.e., in Step 2 we remove patients who died at first hospitalisation, in Step 3 we remove patients who died at their second one, and so on). In the final step of the current analysis (Step 6), all patients are alive at the end of the study. This enables us not only to recognise the impact of dying patients over the clustering algorithm, but also to monitor the group membership of each survived patient over time.

4.1 Step 1

Let's consider the selected cohort. We firstly divide it into three subgroups. This partition is obtained through the

Table 1 Characteristics of groups: size (number and percentage), mortality rate (percentage), mean and median of the length of stay (LOS) from the first admission to next event (in days), mean and standard deviation of age (in years), intensive therapy (number and percentage), procedures (number and percentage), comorbidities (number and percentage)

Properties	Group 1	Group 2	Group 3
Group size	14,872 (43.36 %)	7,820 (22.80 %)	11,606 (33.84 %)
Mortality rate	93.19 %	55.78 %	22.09 %
Mean (med) LOS 1st ev.	352.5 (169)	482.4 (316.5)	2,075 (2,104)
Mean age (sd)	81 (9.76)	76 (10.50)	73 (11.34)
Intensive therapy	3,300 (22.19 %)	2,469 (31.57 %)	2,956 (25.47 %)
Procedures	2,306 (15.51 %)	2,451 (31.34 %)	2,720 (23.44 %)
Comorbidities	12,802 (86.08 %)	7,237 (92.54 %)	8,755 (75.44 %)

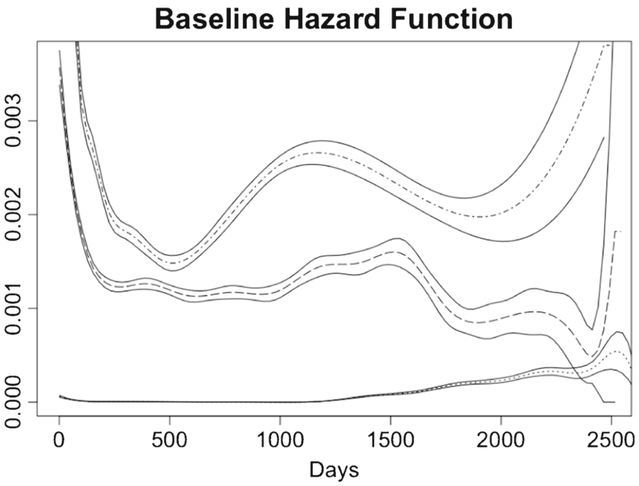


Fig. 2 Baseline hazard functions with 95 % confidence bands. *Dot dashed line*: terminally ill group, dashed line: sick group, *dotted line*: healthy group. *Solid line*: confidence bands for each baseline hazard function

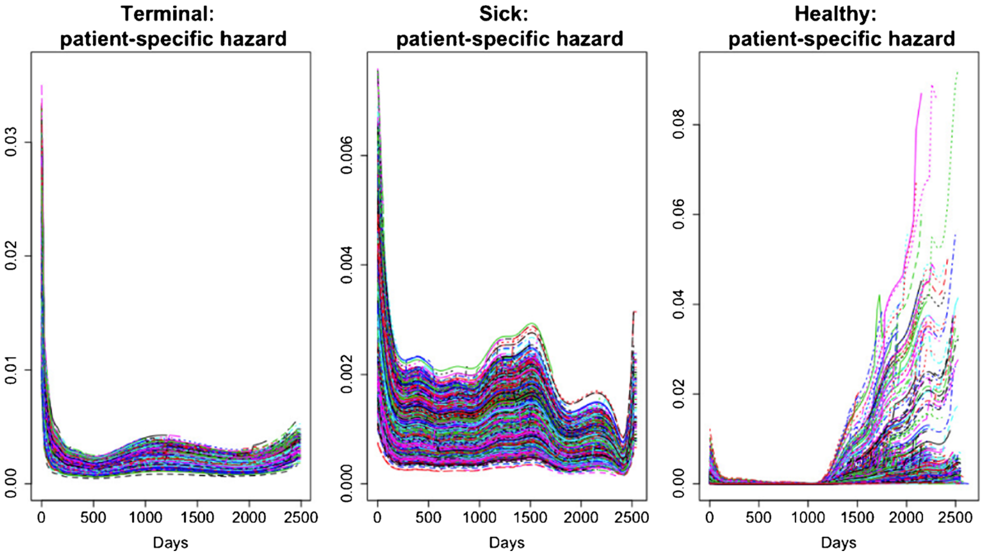
algorithm implemented in PHM package [18] and briefly explained in Section 3.1. Note that the clustering algorithm does not take into account information concerning the type of event we are considering (no information on death of patients are made available to the algorithm). Once the three groups are obtained, we can look back at information extracted from the SDO of every patient and analyse the characteristics of each group.

In Table 1 we give some details of the characteristics of the groups. The first and most interesting data available from this table is the mortality rate: it seems that the algorithm classifies patients mainly according to their mortality risk. Patients in group 1 have the highest mortality rate, while patients in group 3 have the lowest. We can label these groups as: terminally ill (group 1), sick (group 2) and healthy (group 3). Notice that the choice of these names is to give an easy and immediate description of the prototype patient in the group. Other characteristics can be found in Table 1. Among these it is interesting to recognise the mean age trend, which is equal to 81 for the terminally ill group (group 1), and it decreases down to 73 for the healthy group (group 3). Then age results to be another important feature to drive the clustering. It is reasonable since elderly patients are exposed to more serious heart conditions than younger patients.

Once the division of patients into three groups is obtained, we are ready to evaluate patient-specific hazard functions according to the model presented in Section 3.2 and implemented in [23].

In Fig. 2 are shown the baseline hazard functions for the three groups. They perfectly represent the hazard's functional trend expected in each group: terminally ill group has the highest risk of a new event throughout the study time period (dot dashed line), healthy group is nearly null compared to the other groups (dotted line), and the sick group is characterised by an in between trend (dashed line). Terminally ill groups baseline shows a behavior such that its higher values are at extremities: this is due to an increased probability of dying towards the end of the study (a characteristic common to all groups due to censored patients) and an increased probability of dying within the first year of the study, since the most fragile patients show the tendency to have in this period the majority of their hospitalizations. The tendency seen in the terminally ill groups baseline is similar to that of sick groups baseline, where we can appreciate differences in the lower probabilities of the sick group, whereas the corresponding confidence bands are wider, since these patients have a less defined behavior throughout time.

Fig. 3 Patient-specific hazard functions



The same functional shape can be found in patients hazard functions. This is because the baseline gives the general trend to the hazard function of a patient. Through frailty terms and covariates we are able to slightly modify this trend and discern patients' specific trends. In Fig. 3 are shown the resulting hazard function of all patients in the study, all plotted together and stratified by group. Each graphic is shown in its own scale, as otherwise certain properties would be masked. As expected, frailty terms and covariates significantly change the baseline hazard function values. In terminally ill group it is evident an initial high risk that decreases towards the end of the study time period, revealing what was shown in Table 1: the majority of patients in group 1 die before the end of the study time period, and the LOS in the first state (from first admission to next event) is for half patients in this group less than 169 days, equivalent to 5 months. The same analysis can be conducted to discuss patient-specific hazard functions for the other two groups. It is mostly interesting to examine the results of healthy group. A significant change of trend is observable after 1,000 days (2.7 years). Indeed, 99.78 % of patients have a LOS in the first state longer than 1000 days, and, eventually, new events happen during the last four years of observation. This way we are able to justify the great increment in the value of patient-specific hazard functions shown in Fig. 3. In Table 2 the estimates of the covariates coefficients, as well as the variability of frailty terms are reported.

Observing the coefficient estimates in Table 2, we see that age increases the instantaneous risk for all the groups, particulary the healthy one. Intensive therapy does the same in sick and terminal groups, whereas represents a decreasing risk factor for healthy people. Procedures provide benefits whenever patient go through them. Comorbidities are relevant for the healthy group, increasingly substantially the instantaneous risk for a patient presenting them during her/his hospitalizations. Finally, the variability of the frailty within the healthy group is definitively higher than the other two.

4.2 Further steps

After having analysed the results for the selected cohort, it is interesting to give an estimate of the goodness of the model and of the obtained results. The idea is that dying patients influence considerably the clustering process, hence the reconstruction of patient-specific hazard functions. To overcome this problem, step by step we remove from the initial cohort patients dying before the n-th event. This means that at Step 2 we remove patients whose event after the first admission is death. At Step 3 we remove also patients whose event after their second admission is death. At Step 6, we remove all dying patients from the initial cohort.

Table 3 Characteristics of groups in steps from 2 to 6: size (number and percentage), mortality rate (percentage), mean and standard deviation of age (in years)

Group	Terminally ill				
Step	2	3	4	5	6
Cluster Size	6,169	4,836	3,076	2,204	1,834
Cluster Size (%)	28.51 %	28.39 %	20.56 %	15.74 %	13.57 %
Mortality Rate	86.03 %	66.38 %	45.68 %	20 %	0 %
μ(Age)	79.21	76.66	74.65	72.60	71.46
sd(Age)	9.89	10.23	10.65	10.91	10.96
Group	Sick				
Step	2	3	4	5	6
Cluster Size	5,000	2,953	2,889	2,796	2,548
Cluster Size (%)	23.11 %	17.34 %	19.31 %	19.97 %	18.86 %
Mortality Rate	40.24 %	8.30 %	1.45 %	1.79 %	0 %
μ(Age)	74.82	72.28	71.83	71.88	71.69
sd(Age)	10.76	10.88	10.77	10.80	10.96
Group	Healthy				
Step	2	3	4	5	6
Cluster Size	10,470	9,244	8,994	9,003	9,130
Cluster Size (%)	48.38 %	54.27 %	60.12 %	64.29 %	67.57 %
Mortality Rate	7.72 %	0.71 %	0 %	0 %	0 %
μ(Age)	71.87	71.22	71.15	71.15	71.17
sd(Age)	11.25	11.34	11.37	11.36	11.34

In Table 3 we show the characteristics of each group through subsequent steps. From this table we are then able to evaluate whether the algorithm is able to produce a good division of initial cohort.

Firstly it is interesting to look at the dimension of each group through subsequent steps (see also Fig. 4): for the terminally ill group, which is each time identified through the mortality rate index, it decreases considerably from Step 1

Table 2 Estimates (SE) of the covariates coefficients and variance of frailty terms

Estimate (SE)	Group 1 - Terminal	Group 2 - Sick	Group 3 - Healthy
Age	0.1713 (0.0086)	0.2291 (0.0099)	1.0405 (0.0345)
Intensive therapy	0.1826 (0.0206)	0.1988 (0.0242)	−0.1404 (0.0704)
Procedures	−0.3249 (0.0245)	−0.4204 (0.0261)	−0.1953 (0.0736)
Comorbidities	0.0059 (0.0209)	−0.0607 (0.0256)	0.5848 (0.0644)
Frailty variance	0.0889	0.0897	2.7956

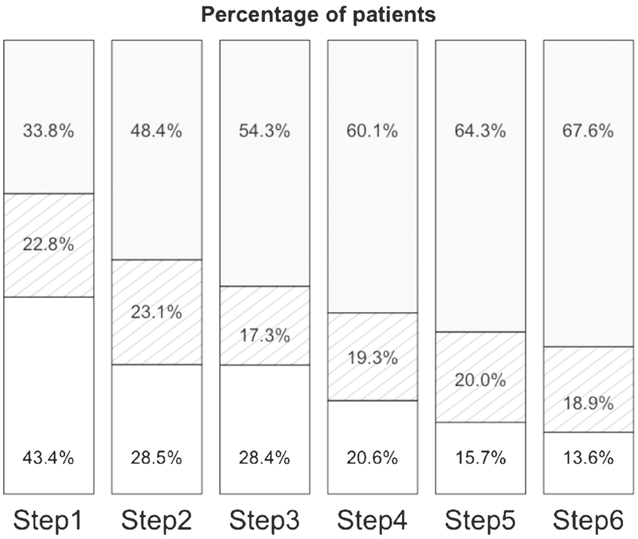


Fig. 4 Evolution of groups' percentage from Step 1 to Step 6. *White bar*: Terminally ill group. *Dashed bar*: Sick group. *Grey bar*: Healthy group

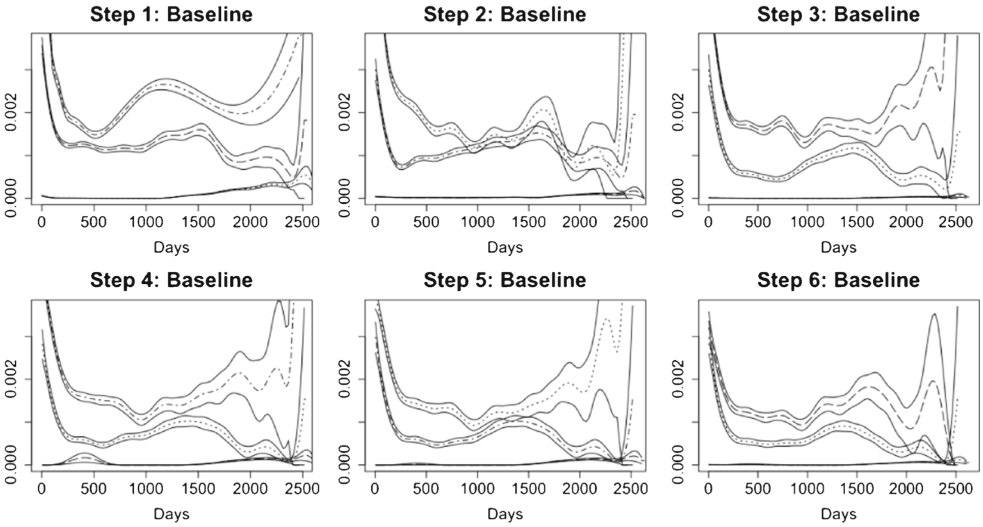
to Step 2 (reduction of the percentage dimension equal to 14.85 %), but it also decreases step by step, reducing to the smallest group by Step 6. This is expected once we remove dying patients, as we are in fact removing terminally ill patients. The opposite result is found for the healthy group, where we observe in Table 3 an increase in its percentage dimension, with a corresponding almost constant group size. Once again this result confirms what expected when removing patients. The sick group is mostly stable both in the size and percentage, except for an initial reduction when removing those patients dying after the first or second admission.

Another interesting feature that can be deduced from Table 3 is that the majority of dying patients are older than the surviving ones. We find that the mean age of dying patients is 81.11 years, where the mean age of surviving patients is 71.31. Notice that the range of ages in the two conditions are the quite similar: [19; 104] vs [18; 99].

From this initial analysis we can state that the clustering algorithm is able to divide patients sufficiently well. It is now important to understand what differences arise when reconstructing hazard functions. Of course the presence of dying patients significantly influence the shape and trend of the estimated baseline hazard function within each group. We can clearly observe this through Fig. 5. The baseline hazard function for the healthy group remains mostly constant throughout the considered steps. In particular, as expected, its values are close to zero compared to those of the other groups (range of values [7.2e − 23; 5.4e − 4]). There is, on the other hand, an evident reduction in the risk of a new event, within the first year of observation, for terminally ill and sick groups. This is due to the removal of dying patients, which as expected considerably influence the baseline hazard functions. Towards the end of the study time period, we can see instead an increase in the risk for terminally ill and sick groups, due to correspondent increase in the percentage of censored patients within the group (once dying patients are removed, the remaining ones are censored).

The corresponding reconstructed hazard functions are obtained applying the model in Eq. 5. We show only the result for Step 6 (see Fig. 6), as for all other steps the results are similar. Moreover, these are interesting results, being the only ones computed over a completely surviving population. We can promptly observe the difference from the results shown in Fig. 3 for terminally ill and sick groups. First, observing the terminally ill group, the range of values for the hazard functions is sensibly reduced: in Fig. 3 there is a high initial pick due to dying patients whose mean LOS in

Fig. 5 Baseline hazard functions with 95 % confidence bands of groups in steps 1 through 6. *Dot dashed line*: terminally ill group, dashed line: sick group, *dotted line*: healthy group. *Solid line*: confidence bands for each baseline hazard function



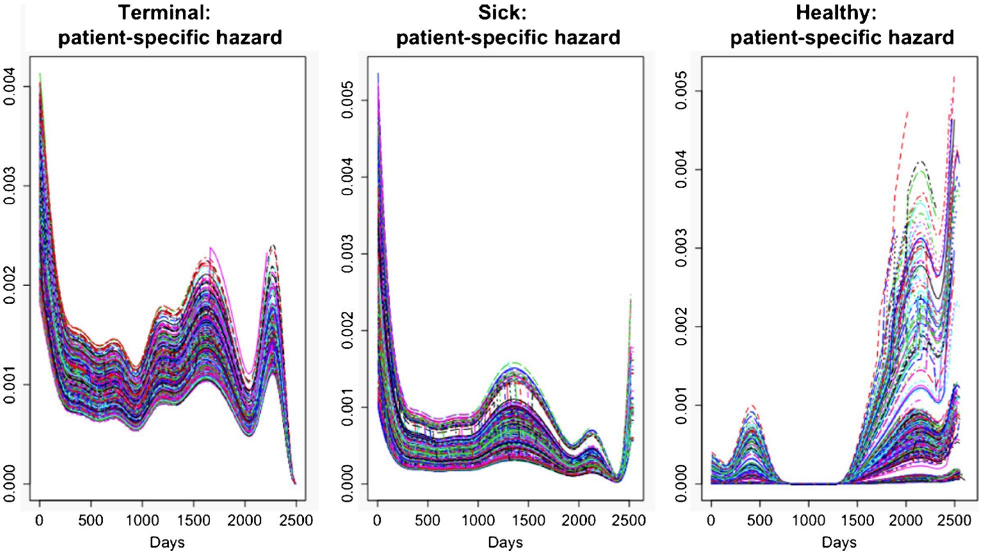


Fig. 6 Patient-specific hazard functions of groups obtained at Step 6, when all dying patients in the initial cohort have been removed

the first state is less than 169 days; once these patients have been removed, we are able to better observe the hazard trend of patients classified as terminally ill. Second, the results for the sick group differ in the variability of the obtained functions: at Step 1 sick group hazard reconstruction was characterised by a great variability, which is significantly reduced at Step 6. Finally, for what concerns the healthy group, also in the result at Step 6 we find the same significant change of trend after 2.7 years, which appeared evident in Fig. 3: due to the regrouping after the removal of dying patients, we observe a new bump in the first segment of these functions, which is the symptom of a reallocation of certain patients.

4.3 Model assessments

On the basis of the results obtained when dividing into three groups the selected cohort, step by step, we are able to understand and give an estimate on what is the movement of patients among groups from one step to the next. Removing dying patients, surviving patients could be assigned to a different group of risk: they may seem healthier than dying patients at Step 1, hence being assigned to the healthy group, but at successive steps it may be that their condition is not perfectly aligned with that of true healthy patients, hence being assigned to sick or terminally ill groups. The same can happen with patients initially considered as terminally ill, or sick, whose condition is then revalued once dying patients are discarded.

In Fig. 7, we show the probabilities that a patient, who was at previous step assigned either to terminally ill, sick or healthy group, is now assigned to terminally ill, sick or healthy group, or is now dead. From these plots it is evident that, once removed the great majority of dying patients (after steps 1 and 2), the probability of being reassigned to the same risk group increases. In particular for the

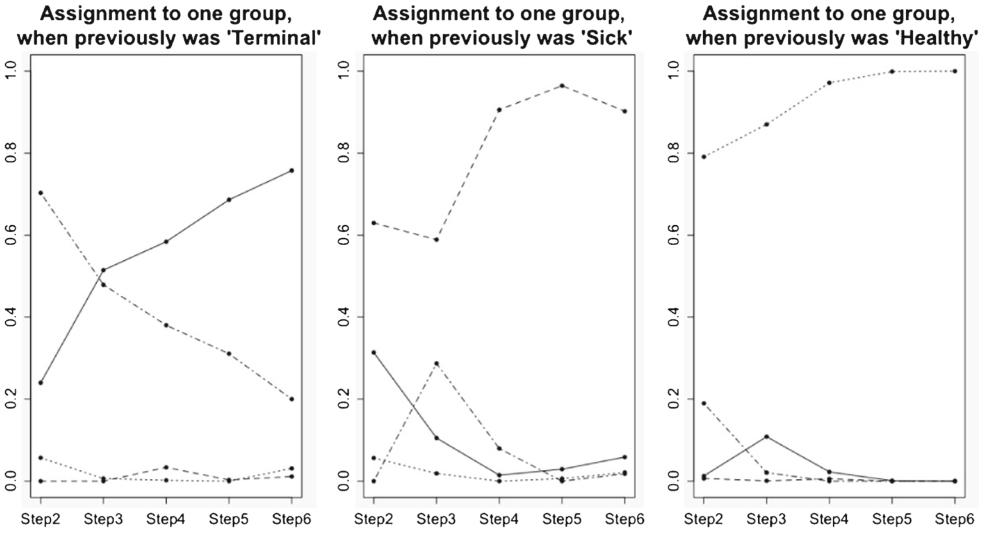


Fig. 7 Probability plots of being assigned to one group, knowing which was the one of provenance at previous step. *First plot* shows the probabilities of being assigned to either the terminally ill, sick, healthy or dying patients group, given that at previous step the patient was assigned to the terminally ill group. The same is for the other two plots, one for patients previously assigned to sick group and one for those who were before assigned to healthy group. *Solid line*: Terminally ill group. Dashed line: Sick group. *Dotted line*: Healthy group. *Dot dashed line*: dead patients

Table 4 Misclassification matrix: known labels on the rows and assigned labels on the columns.

	Group 1 - Terminal	Group 2 - Sick	Group 3 - Healthy
Terminal	11874	2752	246
Sick	97	7545	178
Healthy	0	72	11534

healthy group, at final steps the probability of being healthy once considered healthy is approximately equal to 1. If we observe terminally ill patients plot, once more we find what was already clear from previous analyses: after Step 1, the majority of these patients dies; then the trends of solid and dot dashed lines, being one the new terminal patients and the other that of dead patients, are specular, with a growing tendency for the red line, meaning that step by step the algorithm is able to correctly classify terminally ill patients. Sick patients have characteristics in between the extreme groups (terminally ill and healthy): for this reason these are the most difficult patients to identify as sick. In every step, the algorithm assign again to the sick group the majority of previously considered sick patients; the remaining 10 % is assigned to other groups.

Finally, we can state that dying patients have a great influence at first steps of our algorithm, although once removed it becomes clear to which group each patient is related to. In spite of this evidence, the misclassification error at first steps is not compromising the obtained results for the hazard reconstruction, as the number of misclassified patients remains low compared to the comprehensive cohort dimension.

4.4 Prediction for a new patient

In order to evaluate the goodness of our prevision model, we perform a cross validation analysis over the same dataset through which we built the three clusters (see Section 3.1). The obtained misclassification matrix, with known labels on the rows and assigned labels on the columns is reported in Table 4.

The Actual Error Rate (AER) value obtained for this misclassification matrix is $AER = 0.0975$.

Now that we assigned the new patient j_{new} to one of the three clusters, we can reconstruct her/his hazard function, based on the baselines of the three clusters, obtained in Section 3.2. Using the parameters β of the regression term in Eq. 5 and the variance term for Eq. 6, computed in Section 4.1, we can evaluate the hazard function for patient j_{new}. In Fig. 8, we show three examples, one for each cluster, of correctly classified patients: in the plot are shown in dashed line the original hazard function, and in solid line the new computed hazard function. It is evident from the plots that, there is a slight difference between the two curves, the original one and the new one. This is due to the new value of the frailty term v_j randomly generated from Eq. 6. Of course the difference is mostly emphasized in the healthy group, where the variance of the frailty term distribution is higher (see Table 2).

If we take a look at Fig. 9, we can see some examples of misclassification, one for each type of incorrect assignment of a patient. We can see that in each of the shown cases, there is a considerable difference between the two lines (dashed line: original hazard function, solid line: new hazard function). In particular, it seems that when misclassifying a sick patient, the reconstruction of the hazard function is more similar in the case that the patient is now assigned to the

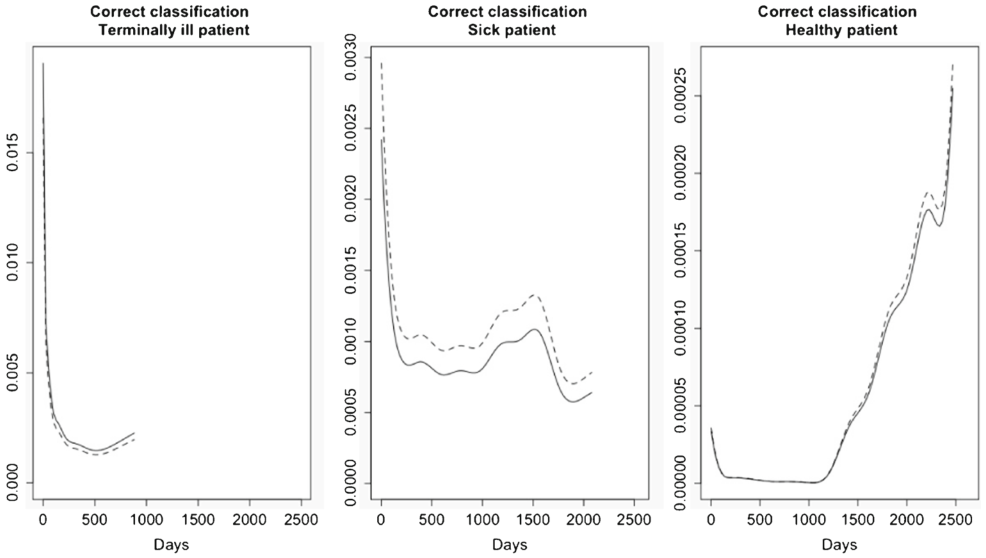


Fig. 8 Examples of correct classification of patients in the three clusters. The original hazard function is the *dashed line*, the new computed hazard function is the *solid line*

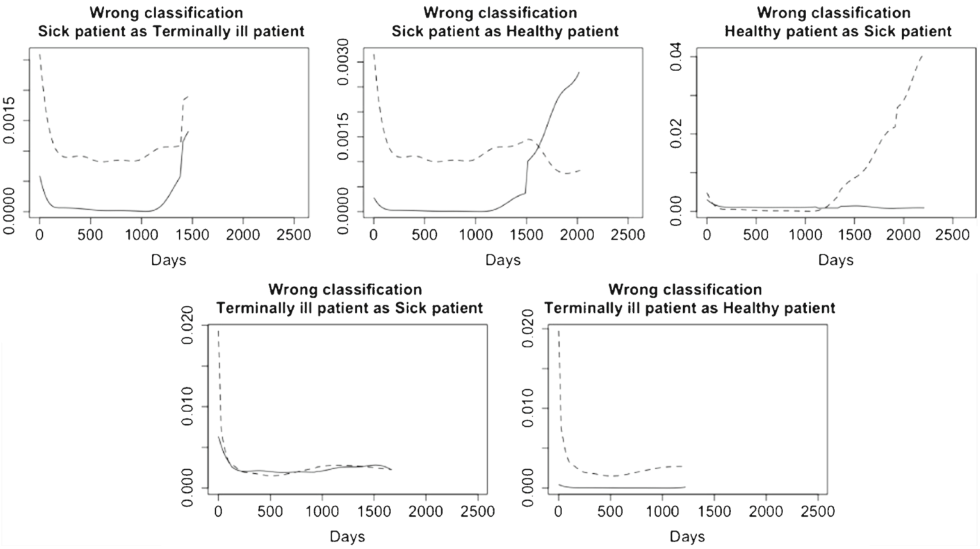


Fig. 9 Examples of wrong classification of patients in the three clusters. The original hazard function is the *dashed line*, the new computed hazard function is the *solid line*

terminally ill group then to the healthy group. This is because, if we recall Fig. 3, we can see that the behaviour of the two functions is much more similar between the groups of sick and terminally ill patients, even though terminally ill patients seem to have a smaller group variance. The same behaviour is found when looking at terminally ill patients who have been misclassified. For healthy patients, who are misclassified only as sick patients, we recover the opposite behaviour, where is even more evident the difference between dashed and solid lines, due to the deep difference in the two clusters.

5 Discussion and concluding remarks

In this work we analysed data on hospitalizations of patients affected by Heart Failure (HF) in Regione Lombardia. We considered patient's histories, in terms of hospitalizations, as trajectories of a non-homogeneous counting process, modelling the inter-times between hospitalizations as mixtures of independent Weibull distributions. This allowed us to make inference on which different latent groups of disease progression patients belong to. Moreover, a method for carrying out predictions for a new patient is proposed. In fact, the aims of the work were twofold: first, to identify latent pattern of disease evolution starting from the hospitalization process, that is the only process administrative data allow to observe. Second, once the evolution patterns are identified, the model becomes a tool for prediction of disease evolution and the corresponding healthcare consumption.

These groups differs mainly in risk of experiencing a new event. Moreover, we can also efficiently predict the group a new patient will belong to and consequently her/his disease progression. This can help the health-care management system to predict health-care consumption, and to optimize the allocation of resources requested in managing health care process of patients affected by the disease object of study. We think that this model could be easily extended and used in all chronic pathologies in which the patient history is characterised by many subsequent events.

One of the main novelty of the work is the use of administrative data for epidemiological purposes. Administrative database and routinely-collected data, infact, have great potential for clinical research, since they are population based and combine information from multiple centers. In so doing, they could capture complete health system use. Moreover, they are usually inexpensive. In particular we model administrative data (i.e. patients histories) in terms of hospital admissions. We model the inter-times between hospitalizations as mixtures of independent Weibull distributions and via EM algorithms we can estimate the related parameters pointing out groups of different behaviours in the readmission process. Once the evolution patterns are identified, the model becomes a tool for prediction of patient disease progress and the corresponding healthcare consumption.

Further developments concern the study of the changing points in the hospitalisation process of each patient, making use of the estimated hazard and the dynamic clustering over interval times. In fact, it is likely that a change in disease status is reflected by a different dynamic in the hospitalizations process. If we were able to identify these changes, this would improve the diagnosis of the disease and the ability of a clinician to predict its evolution. Moreover, a better understanding of the disease trend from a general point of view, might allow hospitals to plan and address in a more proper way the needs of future hospital admissions, improving the efficiency of clinical facilities and, consequently, of the collective welfare.

Acknowledgments This work is within the Strategic Program "Utilization of Regional Health Service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalized for heart failure". The authors wish to thank Regione Lombardia - Healthcare division for having funded and sustained the project, Lombardia Informatica S.p.A. for having provided data, all the physicians who collaborated to the project and the MRC Biostatistics Unit of Cambridge (in particular prof. Sylvia Richardson) for inspiring and supporting the work on methodology.

References

1. Amarasingham R, Moore BJ, Tabak YP et al (2010) An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 48(11):981–988
2. Andersen PK, Keiding N (2002) Multi-state models for event history analysis. Stat Methods Med Res 11:91–115
3. Barbieri P, Grieco N, Ieva F, Paganoni AM, Secchi P (2020) Exploitation, integration and statistical analysis of Public Health Database and STEMI archive in Lombardia Region. Complex data modeling and computationally intensive statistical methods - Series Contribution to Statistics. Springer, pp 41–56
4. Berta P, Seghieri C, Vittadini G (2013) Comparing health outcomes among hospitals: the experience of the Lombardy Region. Health Care Manag Sci 16:245–257
5. Campbell MJ, Jacques RM, Fotheringham J, Maheswaran R, Nicholl J (2012) Developing a summary hospital mortality index: retrospective analysis in English hospitals over five years. Br Med J 344:e1001
6. Chun S, Tu JV, Wijeysundera HC et al (2012) Lifetime analysis of hospitalizations and survival of patients newly-admitted with heart failure. Circ Heart Fail. doi:10.1161/CIRCHEARTFAILURE.111.964791
7. Cook RJ, Lawless J (2007) The statistical analysis of recurrent events, Springer - Statistics for Biology and Health
8. Garbe S, Primer E (2007) Administrative health databases in observational studies of drug effects-advantages and disadvantages. [review]. Nat Clin Pract Rheumatol 3:725–732
9. He J, Ogden LG, Bazzano LA et al (2001) Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study. Arch Intern Med 161:996–1002
10. Hougaard P (1999) Multi-state models: a review. Lifetime Data Anal 5:239–264
11. Ieva F, Jackson CH, Sharples LD (2015) 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. to appear
12. Ieva F, Gale CP, Sharples LD (2015) Contemporary roles of registries in clinical cardiology: when do we need randomized trials? Expert Rev Cardiovasc Ther 12(12):1383–1386
13. ISTAT - Istituto Nazionale di Statistica. http://demo.istat.it/
14. Joynt KE, Jha AK (2011) Who has higher readmission rates for heart failure, and why? Implications for efforts to improve care using financial incentives. Circ Cardiovasc Qual Outcomes 4:53–59
15. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S (2011) Risk prediction models for hospital readmission, a systematic review. J Am Med Assoc 306(15):1688–1698
16. Krumholz HM, Merrill AR, Schone EM et al (2009) Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission. Circ Cardiovasc Qual Outcomes 2:407–413
17. Lloyd-Jones D, Adams RJ, Brown TM, Carnethon M, Dai S, De Simone G, Ferguson TB, Ford E, Furie K, Gillespie C, Go A, Greenlund K, Haase N, Hailpern S, Ho PM, Howard V, Kissela B, Kittner S, Lackland D, Lisabeth L, Marelli A, McDermott MM, Meigs J, Mozaffarian D, Mussolino M, Nichol G, Roger VL, Rosamond W, Sacco R, Sorlie P, Roger VL, Stafford R, Thom T, Wasserthiel-Smoller S, Wong ND, Wylie-Rosett J (2010) American heart association statistics committee and stroke statistics subcommittee. Heart disease and stroke statistics–2010 update: a report from the American Heart Association. Circulation 121(7):e46
18. Mair P, Hudec M (2008) mixPHM: Mixtures of proportional hazard models. R package version 0.7.0
19. Mair P, Hudec M (2009) Multivariate Weibull mixtures with proportional hazard restrictions for dwell-time-based session clustering with incomplete data. J R Stat Soc Ser C Appl Stat 58:619–639
20. McMurray JJ, Petrie MC, Murdoch DR, Davie AP (1998) Clinical epidemiology of heart failure: public and private health burden. Eur Heart J 19 Suppl:P9
21. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2009. [Online] http://www.R-project.org
22. Rondeau V, Commenges D, Joly P (2003) Maximum penalized likelihood estimation in a Gamma-Frailty model. Lifetime Data Anal 9(2):139–153
23. Rondeau V, Mazroui Y, Gonzalez JR (2012) Frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametric estimation. J Stat Softw 47:1–28
24. Ross JS, Chen J, Lin Z et al (2010) Recent national trends in readmission rates after heart failure hospitalization. Circulation 3:97–103
25. Shulan M, Gao K, Dea Moore C (2013) Predicting 30-day all-cause hospital readmissions. Health Care Manag Sci 16:167–175
26. Shams I, Ajorlou S, Yang K (2015) A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or COPD. Health Care Manag Sci 18:19–34
27. Van Walraven C, Austin P (2012) Administrative database research has unique characteristics that can risk biased results. J Clin Epidemiol 65:126–131
28. Wirehn AB, Karlsson HM, Cartensen JM et al (2007) Estimating Disease Prevalence using a population-based administrative healthcare database. Scand J Public Health 35:424–431