POLITECNICO DI MILANO
School of Systems Engineering
MSc of Mathematical Engineering

Graduation thesis in Statistics

Hazard reconstruction and clustering for better prognosis of disease progression in heart failure

Advisors:

Prof. Anna Maria Paganoni

Dr. Francesca Ieva

Candidate:

Teresa Pietrabissa

782064

Academic Year 2012-2013

Ai miei fratelli.

Abstract

Nowadays Heart Failure (HF) is considered the leading cause of repeated hospitalisations in patients aged over 65. The resulting longitudinal dataset and its analyses are consequently becoming of a great interest for clinicians and statisticians worldwide. We analysed HF data collected from the administrative databank of an Italian regional district (Lombardia), concentrating our study on the days elapsed from one admission to the next for each patient in our dataset. The aim behind this project is to identify groups of patients, conjecturing that the variables in our study, the time segments between two consecutive hospitalisations, are Weibull distributed in each hidden cluster. Therefore, the comprehensive distribution for each variable results in a Weibull Mixture Model. From this assumption we developed a survival analysis in order to estimate, through a proportional hazards model, the corresponding hazard function for the proposed model and to obtain the desired clusters. We find that the selected dataset, a good representative of the complete population, can be categorized into three clusters, corresponding to "healthy", "sick" and "terminally ill" patients. Furthermore, we attempted a reconstruction of the patient-specific hazard function, adding a frailty parameter to the considered model.

Sommario

Ai nostri giorni lo scompenso cardiaco (HF) è considerato la causa principale delle numerose ospedalizzazioni in pazienti di età oltre i 65 anni. I dati longitudinali che si ottengono e le analisi condotte sugli stessi, stanno diventando di grande interesse per i clinici e gli statistici. In questa tesi vengono analizzati dati sullo scompenso cardiaco provenienti dalla banca dati amministrativa di regione Lombardia. Ai pazienti presi in considerazione, è stato diagnosticato lo scompenso cardiaco e questa diagnosi può essere ricavata da precise codifiche presenti in tale banca dati (si vedano i Capitoli 1 e 2). Solo i pazienti dimessi dopo la prima ospedalizzazione entro la fine del 2006 sono stati presi in considerazione ai fine dello studio, che ha avuto una durata di 5 anni, fino al 31 Dicembre 2010. In questa tesi, i pazienti con più di 5 ospedalizzazioni sono stati esclusi dalle analisi, in quanto tale riduzione del dataset permette una più agevole analisi dei dati, senza causare per questo una considerevole perdita di informazioni.

Uno dei principali scopi di questa tesi, è quello di identificare gruppi di pazienti. Le variabili prese in considerazione per ottenere i risultati sperati, sono variabili temporali caratteristiche dei dati longitudinali, che nel nostro caso corrispondono ai giorni che intercorrono tra una ospedalizzazione e la successiva. Chiameremo queste variabili *intertempi*, e verranno indicate come T_{ih}, dove i rappresenta l'i-esimo paziente, e h è la corrispondente h-esima ospedalizzazione. Per identificare i gruppi di pazienti tra quelli in analisi, abbiamo ipotizzato che gli intertempi corrispondenti a ciascuna ospedalizzazione, ovvero T_1, T_2, T_3, T_4 e T_5, siano distribuiti secondo misture di Weibull. Questo risultato è chiaro se si immagina di conoscere i gruppi di pazienti che vogliamo identificare. In particolare, ipotizzando che ogni intertempo proprio di un gruppo di pazienti sia una variabile casuale distribuita secondo una legge Weibull, risulta evidente che la rispettiva variabile casuale globale è distribuita seconda la mistura delle Weibull delle variabili nei gruppi considerati. Fissando il numero di gruppi di pazienti pari a $K = 3$, il modello risultante per ciascun intertempo T_h è della forma:

$$f(t_h) = \sum_{k=1}^{K} \pi_k f(t_h; \lambda_{kh}, \gamma_{kh})$$

dove

$$f(t_h; \lambda_{kh}, \gamma_{kh}) = \frac{\gamma_{kh}}{\lambda_{kh}^{\gamma_{kh}}} \, t_h^{\gamma_{kh}-1} exp(-(t_h/\lambda_{kh})^{\gamma_{kh}}).$$

Secondo il modello appena riportato, T_h ha funzione densità $f_T(\cdot)$ che è, come già detto, una mistura di Weibull: π_k sono i pesi di ciascuna misturante, pari alla numerosità del cluster corrispondente, e $f(t_h; \lambda_{kh}, \gamma_{kh})$ è la funzione densità della variabile aleatoria T_{kh}, ovvero l'intertempo tra la h-esima e la $(h+1)$-esima ospedalizzazione per i pazienti del cluster k.

Partendo da questa assunzione, è stata condotta un'analisi di sopravvivenza per stimare, attraverso il modello proportional hazards model (PHM), la funzione di hazard, rischio di ri-ospedalizzazione, di ciascun gruppo e per ottenere quindi i cluster di pazienti. L'analisi è stata condotta avvalendoci del software R [26], con cui abbiamo eseguito tutte le analisi discusse in questa tesi. In particolare, è stato utilizzato il pacchetto `mixPHM` [16], che stima i parametri della densità di ciascun intertempo, permettendo la ricostruzione della relativa funzione di hazard, e assegna ciascun paziente ad uno dei cluster ottenuti: una volta stimata la divisione in gruppi, viene aggiornata la stima dei parametri e vengono riassegnati i pazienti ai nuovi cluster, ripetendo il procedimento in modo ciclico fino a che una qualche condizione di iterazione non sia più soddisfatta. L'algoritmo utilizzato per raggiungere questo risultato è l'algoritmo EM, un algoritmo iterativo che permette di ottenere la classificazione desiderata. Applicando questo modello alla popolazione in analisi, che può essere considerata rappresentativa della popolazione globale, dai risultati appare evidente che vi sono tre distinti gruppi di pazienti: pazienti "sani", pazienti "malati" o pazienti "terminali". Le etichette che abbiamo assegnato a ciascun cluster sono il risultato di una dettagliata analisi delle caratteristiche di ciascun gruppo ottenuto. In particolare, dalle analisi è emerso che è possibile identificare un gruppo di pazienti caratterizzato da un indice di mortalità molto alto, in cui i pazienti che ne fanno parte hanno, complessivamente, poche, ma frequenti, ospedalizzazioni. Queste caratteristiche ci hanno portato a identificare questo cluster con il gruppo dei pazienti "terminali", ovvero pazienti in uno stadio della malattia molto avanzato. Il cluster identificato con i pazienti "sani", ad una analisi descrittiva, mostra proprietà opposte a quelle del primo gruppo: i pazienti che fanno parte di questo cluster rimangono (quasi) tutti vivi fino alla fine del tempo di osservazione, con un risultante indice di mortalità

molto basso. I pazienti di questo cluster hanno poche ospedalizzazioni, registrate, l'una dall'altra, a distanza di lunghi archi temporali. L'ultimo cluster ha proprietà comuni agli altri due, ed è, per questo, stato identificato come il gruppo dei pazienti "malati". In questo cluster sono presenti tutti quei pazienti la cui malattia non può più essere considerata allo stadio iniziale, ma neppure mostrano le caratteristiche proprie di pazienti allo stadio finale. Tutte queste considerazioni sono discusse approfonditamente nel Capitolo 3.

Identificati i gruppi in cui è possibile suddividere il dataset in analisi, il secondo scopo della tesi è stato quello di ricostruire, per ciascun paziente, le funzioni di rischio di ri-ospedalizzazione su tutto l'arco temporale dello studio. Il modello utilizzato per ottenere questo risultato è un modello non parametrico, facente parte della classe dei Frailty Models. Questi sono una naturale estensione del modello non parametrico di Cox [7], in cui viene introdotto un parametro, detto frailty, utile a stimare l'eterogeneità tra i soggetti in analisi. Il pacchetto utilizzato per questo scopo è il `frailtypack` [29]. Il pacchetto stima in modo non parametrico la baseline hazard function del modello: questa è la funzione di rischio che rappresenta l'andamento della popolazione in analisi. Per differenziare ciascun paziente e ottenere la sua specifica curva di rischio, il modello calcola la hazard function del paziente i come segue:

$$h_{ij}(t|v_i) = h_0(t)v_i exp(\boldsymbol{\beta}'\mathbf{z}_{ij})$$
$$v_i \overset{\text{i.i.d.}}{\sim} log - Normal(0, \sigma^2)$$

dove $h_0(t)$, come già detto, è la baseline hazard function; v_i rappresenta il termine di frailty inserito nel modello, variabile aleatoria tempo-indipendente che segue la distribuzione log-Normale; $\boldsymbol{\beta}$ è il vettore dei coefficienti del termine regressivo del modello e viene stimato dall'algoritmo; $\mathbf{z}_{ij}$ è la matrice delle covariate introdotte nel modello, sulla base delle quali valutare il termine regressivo. Come viene spiegato dettagliatamente nel Capitolo 4, le covariate che abbiamo inserito nel modello sono sia tempo-dipendenti, che costanti. Questo modello è stato applicato su ciascuno dei cluster trovati con il metodo di clustering adottato nella prima fase dell'analisi del dataset. I risultati, ottenuti con la ricostruzione funzionale e paziente-specifica della funzione di rischio, confermano le ipotesi fatte sui gruppi precedentemente identificati. L'andamento che si ottiene rispecchia perfettamente quelle che sono le caratteristiche dei tre gruppi: nel dettaglio, le funzioni di rischio dei pazienti classificati come "terminali" hanno valori maggiori nei primi anni di osservazione, e raggiungono la soglia della probabilità nulla di ri-ospedalizzazione entro la fine dello studio. Inoltre la variabilità stimata

tra questi pazienti risulta molto bassa, un dato che rispecchia perfettamente le proprietà intrinseche dei pazienti di questo tipo. Per quanto riguarda le funzioni di rischio di ri-ospedalizzazione dei pazienti "sani", queste hanno una forma funzionale speculare a quella ottenuta per i pazienti "terminali", ancora una volta un segno chiaro della differenza tra i pazienti che compongono questi due cluster. Tali pazienti hanno valori alti di probabilità di ri-ospedalizzazione solo verso la fine dello studio, quando l'incertezza della stima si fa maggiore, a causa della perdita di informazioni dovuta al censuramento degli intertempi (tutti i pazienti che sono vivi alla fine dello studio, hanno i corrispondenti intertempi T_{iH_i} censurati, con H_i il numero di ospedalizzazioni del paziente i). Infine, come prima, per i pazienti "malati" otteniamo funzioni di rischio che hanno proprietà di entrambi gli altri gruppi: tali funzioni hanno valori molto bassi verso la fine dello studio e valori più alti nei mesi iniziali, ma si distinguono dal cluster dei pazienti "terminali" in quanto il loro andamento funzionale non risulta monotono decrescente, proprietà attribuibile alle funzioni di rischio ottenute per l'altro gruppo di pazienti.

In conclusione, dalle analisi effettuate sul dataset originale, ne deduciamo che è possibile trovare e distinguere gruppi di pazienti con caratteristiche predominanti e differenti: riconoscere cluster diversi di pazienti e valutarne il rischio di una nuova ospedalizzazione permette di tracciare, a livello temporale, l'evoluzione della malattia di un paziente affetto da scompenso cardiaco. In particolare, nel Capitolo 5, abbiamo sviluppato un'ulteriore analisi, eliminando, di volta in volta, i pazienti deceduti, al fine di avere delle stime di rischio meno sensibili ai pazienti terminali, e più affidabili per quanto riguarda la maggioranza dei pazienti con questa patologia. La possibilità di identificare gli stadi della malattia basandosi sulle conoscenze pregresse e sulla storia di ciascun paziente, consentirebbe di dare al medico uno strumento aggiuntivo per pianificare eventuali visite preventive, e agli ospedali il modo di programmare la disponibilità al ricovero di nuovi pazienti sulla base della domanda prevista. Un tale progresso nel sistema di diagnosi dello scompenso cardiaco, permetterebbe un servizio migliore alla comunità e un miglioramento nell'approccio alla cura della malattia. Nel Capitolo 6, sono riassunti i risultati ottenuti e discussi i possibili sviluppi futuri.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is the result of several analyses on a dataset concerning patients diagnosed with heart failure (HF) or chronic heart failure. One of the first aim behind this thesis, is to find, within the analysed dataset, groups of patients with distinctive characteristics, which could help clinicians monitor their disease evolution. Furthermore, we want to evaluate the increment in the risk of a new event, i.e. a new hospitalisation, for patients that were, in their past history, admitted several times for HF. At last, computing patient-specific hazard functions could allow to estimate the probability of a new hospitalisation for a single individual, based on the information collected before and during the study.

In this introduction we summarize the characteristics of heart failure disease, and give a brief and basic explanation of the kind of data we will be working with throughout the rest of the thesis (see Section 1.1, Section 1.2 and Section 1.3). In the following chapters we will discuss the theory behind the models we applied and the results obtained. In particular, in Chapter 2 we present the dataset we worked on, outlining the way data were collected, its properties and describing covariates that are to be considered for the thesis purposes and that will play a fundamental role for the computation of the proposed models. In Chapter 3, we present the theory behind `mixPHM` package, the properties of the considered model and, consequently, the transformations that are to be done to our dataset in order to be able to estimate the parameters of the desired model. Moreover, we present the results obtained when running the package with our data: we show the result of the clustering method, based on the idea that the underlying distribution of times to the next hospitalisation is a Weibull Mixture Model (see Section 3.1). The model distinguishes three groups among considered patients, whose properties will be widely discussed in Section 3.3. In Chapter 4, we present a methodology

to attempt a patient-specific hazard reconstruction through non-parametric models: the theory, the use of `frailtypack` package, the manipulation of data and the obtained results, are presented in chapter's sections. Deeper analyses are then conducted in Chapter 5, where we apply, once more, the previously discussed models, and corresponding algorithms, over a precisely conjectured group of patients, selected from our dataset. Finally, in Chapter 6, we go through the obtained results and analyse the overall outcome, reviewing the conclusions discussed in previous chapters and concluding on further developments.

1.1 Heart Failure

Heart failure is a term conjectured to identify 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, 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 [10]. Professor Packer defined it as 'a complex clinical syndrome characterised by abnormalities of left ventricular function and neurohumoral regulation, which are accompanied by effort intolerance, fluid retention and reduced longevity' [21]. Despite the great amount of medical literature that have been published to define and label HF and chronic HF, up to these days there is not a common manner to identify and diagnose this disease. Nevertheless, this is, in spite of newer medical techniques adopted to improve patients' condition, the leading cause of hospitalisation in elderly subjects. People affected by heart failure are estimated to be 23 million worldwide.

There exist several codes to identify and label patient's condition as HF. We analysed patients records extracted from the project "Utilisation of Regional Health Service databases for evaluating epidemiology, short- and medium-term outcome, and process indexes in patients hospitalised for heart failure". To consider, within the study case, the majority of HF patients, admissions diagnosed in Major Diagnostic Category (MDC) 01 - Nervous System, 04 - Respiratory System and 05 - Circulatory System, that happened in the Northern Italy regional district of Lombardia, have been included. To identify HF patients, a list of ICD-9-CM codes has been created as the union of codes from "Heart failure mortality rate" by AHRQ-IQI [12] and from CSM-HCC Model Category 80 [25, 24]. Based on these categories, a dataset

of patients diagnosed with heart failure has been created. The resulting dataset is made of longitudinal data (see Section 1.2) and its properties are widely discussed in Chapter 2.

1.2 Survival Data

Survival data are a particular category of statistical data, which describe the time to some event: in our case, data represent the time to the next hospitalisation. As for their nature, survival data are positive real valued variables having a continuous distribution [11]. To collect this kind of data, it is necessary to identify a starting point, $t = 0$, common to every individual in the study case, from which times to the next event are measured. As we will later show, for each patient in our dataset we fixed the starting time point to be equal to the time of first admission, which is an event common to every patient, since they all experienced it in order to fit in our study case.

Recalling that our dataset is made of patients diagnosed with chronic heart failure, they are to be considered as high-risk patients, who naturally experience frequent events, i.e. hospitalisations. Our patients are followed along five years, and the collection of records from each patient give rise to a particular structure of the survival data, which earns them the name of longitudinal data. These are to be considered as trajectories of a stochastic process. In particular, the resulting data represents the life history of a single individual, who can experience the same event (hospitalisation for HF) several times. When dealing with longitudinal data, we can hypothesise that patients' hospitalisations are independent. As a result, the transition times between two events will be an increasing sequence of independent random times, and we will allow only the last survival time to be censored. Censored times means that we have partial information on a patient. As in most of the study cases in literature, if a patient is alive at the end of the study time period, we do not know her/his complete lifetime, since it exceeds the time of observation. For this reason, we follow the individual i over the interval $(0; C_i)$, where C_i is the censoring time for patient i and it is a fixed random time. Events (new hospitalisations) happen within this time interval, and the corresponding time is T_{ih}, where $i = 1, \dots, n$ and $h = 1, \dots, H_i$. Notice that, for the inner nature of the disease, H_i, i.e. the number of hospitalisations along the observed lifetime of patient i, is a random variable. Patients move through these H_i states, called S_{ih}: the sequence of states, proper to each patient i, represents the increasing number of events the individual experienced. Moreover, we need to know the transition type from one state

to the next. Referring to U_i as the lifetime span of patient i, and c_i the observed censoring time, the last survival time is $T_{iH_i} = \min(U_i, c_i)$, and the corresponding transition type is $D_{iH_i} = \mathbb{I}\{T_{iH_i} = U_i\}$. D_{iH_i} is then equal to 1 if the patient died within the time of observation and it is equal to 0 if the last patient's transition time is a censoring.

Throughout this thesis project, we will discuss survival times' distributions in terms of the hazard function, which is defined as the probability of the next event (say a new hospitalisation) within a short interval, conditionally on the fact that it was not yet happened at the beginning of the considered time interval. In our case, this means that the hazard function is computed as the probability of a patient, or a group of patients, to experience the $(h+1)$-th hospitalisation in the time interval $\Delta t \to 0$, knowing that at the considered time point t the number of admissions was equal to h. Properties of the hazard function will be furthermore discussed in Section 3.1.

1.3 Working with big data

> *«How big must big data be?»*
> *«Statisticians were very helpful in the past, because in the past collecting data was so difficult, so time consuming, so costly. So statisticians told us that if we have any population and we want to understand that population, if we do a random sample of 1200 of them, we can understand the entire population. And that is a really good short-cut, to just do a sample. But we can not, with just this sample, go into details or answer questions after we have collected the sample. It is very different if we collect all of the data. That is what the power of big data is. [...] It is not the absolute number that counts. It is the relative number, relative to the question that we have at hand, relative to the population I want to study.»*
>
> *I Big Data Secondo Viktor Mayer-Schönberger.*
> *Interviw to V. Mayer-Schönberger by F. Pedrocchi. [18]*

The dataset we will be working on, represents the northern Italy population diagnosed with heart failure (see previous section), whose first admission ended within the year 2006. We follow and record their admissions history up to December 31st, 2010. In particular, the selected dataset takes into

account for those patients that had a maximum of five hospitalisations during the observation time (see Section 2.1 for details). This choice permits us to group patients in the study case, according to their total number of hospitalisations H_i: many of the analytical results, presented in this thesis, were obtained considering each patient within the corresponding group, based on the number of observed admissions. Furthermore, we will compute many of the results conditionally on the number of hospitalisations a patient experienced. The resulting dataset, as we will later discuss (see Chapter 2), is made of 13785 patients, corresponding to 27392 records. For each event of patient i, we collect all the information available from the SDO (hospital discharge papers). For this reason, the obvious consequence is a difficulty in manipulating the dataset: each time we will try to fit our data to a conjectured model, we will need to adjust the dataset format to the most suitable one for our intents, and, moreover, we will need to select among all information, those useful for the proposed analyses. In Appendix B, we present the code used to generate and compute all the results discussed in Chapters 2, 3, 4 and 5. R software [26] was used to develop all the analyses, and we took advantage of several of its packages in order to compute the desired algorithms. These will be presented in the following chapters.

Chapter 2

The Dataset

Today the role of healthcare databases in the study of health and disease conditions in defined populations has become fundamental, thanks to the great availability of these data. Every time a patient is being admitted to a hospital, this event gets registered in the administrative database, along with several information concerning the sex of the patient, the age and other significant details. For this thesis project we used a dataset collected from the major project called "Utilisation of Regional Health Service databases for evaluating epidemiology, short and medium term outcome, and process indexes in patients hospitalized for heart failure". This dataset collects hospitalisations' information from patients affected by chronic Heart Failure (HF) in the regional district of Regione Lombardia, Italy. Data include all patients whose first discharge happened within 2006 and their disease evolution is followed up to December 31st, 2010, covering an overall time of five years. Information on potential death events, happened before the end of the study, were obtained from the admissions database to the National Registry of deaths. Patients are included in our database if they show a HF code in any of the six diagnosis fields of the SDO (Scheda di Dismissione Ospedaliera, i.e. hospital discharge paper). The comprehensive number of patients in the study is 15856, equivalent to 36949 records. Moreover, other restrictions were applied to the dataset: patients younger than 18 years, together with those patients who were admitted and discharged the same day or whose records contained errors, were removed from the dataset, leading to a final number of 15298 patients and 35224 records.

2.1 Selection criteria for the observation units

What is of central importance when analysing an administrative database, is to properly select the most suitable observation units in order to obtain a correct and more robust interpretation of the results. The criteria chosen for the study will affect the outcome of any study, leading to different - but not divergent - images of the diseases.

Patients in our dataset have all at least one hospitalisation. There are three possible events after their first hospitalisation: they could die, while they are still in hospital or once they have been discharged - but we will consider no difference between these two cases - they could have no other admissions or they could have a new hospitalisation event. We decided to set the maximum number of consecutive hospitalisations for each patient to be equal to five. This is because, analysing the dataset made out of the population with a maximum of five hospitalisations over the study time period, we are indeed analysing the 94.41% of the complete dataset, see Figure 2.1. Also, as we can see from Figure 2.2, increasing the maximum number of admissions from four to five gives us the last significant expansion of the dataset we are about to analyse.

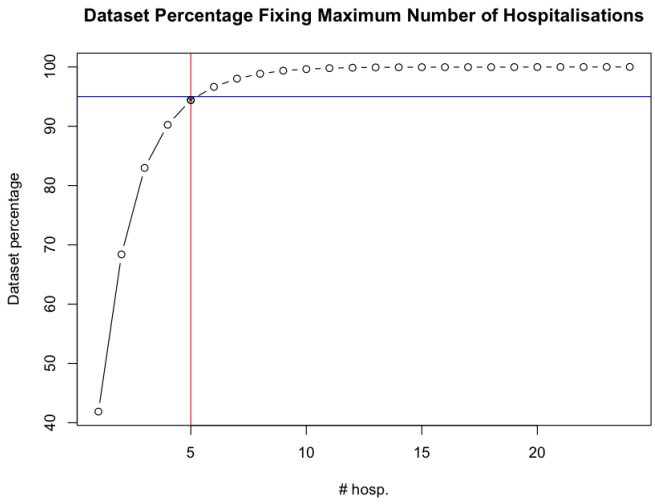


Figure 2.1: Percentage of the dataset analysed when fixing a maximum number of hospitalisations: the blue line is the 95% line, the red line highlights the point where we set a maximum of five hospitalisations.

After the very first analysis, we decided also to consider in our dataset just patients who didn't experienced any shock during the five years of observation. A circulatory shock is a life-threatening medical condition character-

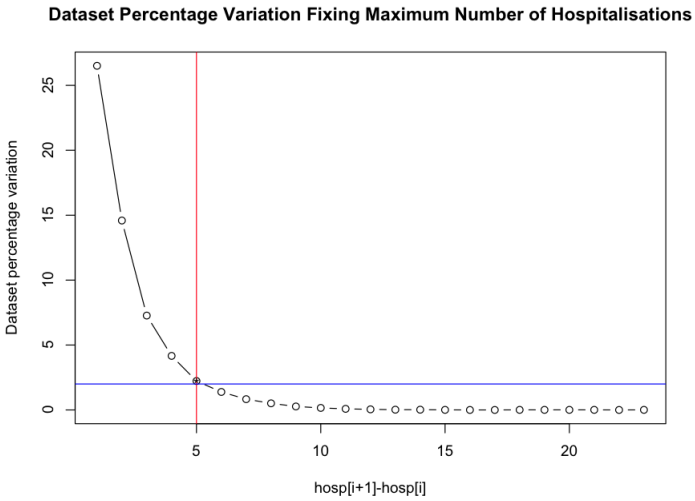


Figure 2.2: Percentage of variation between two consecutive maximum hospitalisations numbers. The blue line is the 2% line, the red line highlights the point where we set a maximum of five hospitalisations.

ized by low blood pressure, rapid heartbeat and poor end-organ perfusion. We chose to exclude these patients from our study because their behaviour is thoroughly unlike the one of the majority of patients in this trial. The most evident difference can be seen in the mortality rate, as these patients show a higher rate value compared to that of the complete dataset - 84.12% against 44.46%. Moreover, the number of patients with this medical condition in the initial dataset is equal to 724 (equivalent to 4.73% of the collected data).

As a result of all these choices, the dataset reduces to 13785 patients, equivalent to 27392 events. This is the observation units collection on which we will conduct our analysis.

2.2 Analysed features

Between all the information collected from the record of a patient admission, we selected some of these to conduct our analysis with regard to specific factors. Every patient is identified by an encrypted ID code, so that we can follow her/his admissions history. For each hospitalisation we collect the age at the date of admission, the sex of the patient, dates of admission, discharge and, if applicable, death. We have a boolean variable stating whether the patient died before the end of the study period or not. Five other information for each patient are collected: these are all boolean variables which, for each

patient's admission, state whether a specific condition or treatment is applicable. Then we merged all the information collected for each specific variable in such a way that it will be equal to 1 throughout the patient's history if, in at least one admission, the variable is equal to 1. First of all, we collect the information regarding the circulatory shock. As we already discussed, this induced us to draw a distinction between two groups of patients, i.e. those who do have at least one shock event in their hospitalisation history, during the time of analysis, and those who don't. Another variable of interest for our study is the boolean variable "CABG", Coronary Artery Bypass Surgery, commonly known as heart bypass surgery. "CABG" is a surgical procedure performed to relieve angina, chest pain caused by insufficient oxygen supply, and to reduce the risk of death from coronary artery disease. The other three boolean variables are "PTCA", "ICD" and "STENT". "PTCA" stands for Percutaneous Transluminal Coronary Angioplasty, commonly known as coronary angioplasty, and it is a non-surgical procedure used to treat the abnormal narrowed coronary arteries of the heart found in coronary heart disease. This type of clinical presentation is often caused by the build up of the cholesterol-laden plaques. "ICD", Implantable Cardioverter-Defibrillator, is a small battery-powered electrical impulse generator that is implanted in patients who are at risk of sudden cardiac death due to ventricular fibrillation and ventricular tachycardia. This device include electrode wire that pass through a vein to the right chambers of the heart, usually lodging in the apex of the right ventricle. "STENT" corresponds to the surgical procedure of placing a surgical device called stent, which is a mesh tube inserted into a natural passage in the body to prevent or counteract a disease-induced flow constriction. These are all the variables we will be using during our analysis of the dataset and in the attempt to reconstruct the patient-specific hazard functions (see Chapter 4).

2.3 Descriptive analysis of the dataset

We want here to give an explanatory description of the characteristics of this dataset. First of all we recall that we selected 13785 patients, each with a maximum of five hospitalisations and no shock events during the observation time of this study. The dataset is distributed over the five hospitalisations as expected: as we can see from Figure 2.3, there are very few patients, compared to the other groups, that have five hospitalisations and the maximum number of patients localises in the first three groups, corresponding to those patients who have a maximum of one, two or three admissions. This is probably due to two main causes: patients who are at the

early stages of their disease, over the examined time window experience few events, while, on the other hand, there are those patients who are at the end of their disease evolution and die during one of their first hospitalisations. These two opposing conditions merge together giving the results we see in Figure 2.3. The observed trend of the number of patients' function is monotonically decreasing, representing the behaviour outlined.

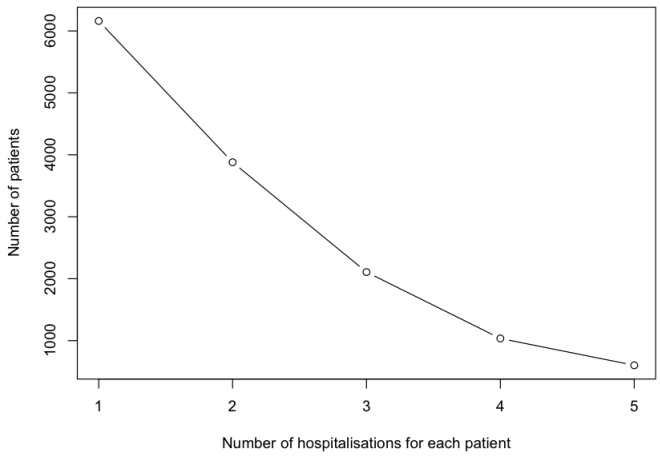


Figure 2.3: Number of patients for each maximum number of hospitalisations. Each patient is counted just in the group corresponding to her/his maximum number of hospitalisations.

It is important to analyse the death index of our dataset. The mortality rate is equal to 42.15%, and as we recall from Section 2.1, this percentage has diminished compared to the one of the original dataset (44.46%). This is because firstly we shrank the dataset to those patients who have a maximum number of hospitalisations equal to five, which slightly reduces the mortality rate to 44.11%. Moreover, we do not consider those patients who experienced a shock event during at least one of their admissions. These patients have a high mortality rate and this leads to the major reduction in the percentage. As a matter of fact, we can see from Table 2.1 that we are really diminishing the mortality rate in each of the considered groups of patients, resulting in the reduced overall percentage. Of course the reduction of the mortality rate is considerably higher in the very first hospitalisations, as patients who experienced at least one shock happen to have a mean number of admissions equal to 1.94. Hence we obtained from the tests of equal proportions that there is a significant difference between these two groups, with or without shock-patients, only in the first three cases. When we consider a maximum number of hospitalisations equal to four or five, the number of patients with at least one shock is not influential over the equivalent dataset.

Max	With Shock	Without Shock	p-value
1	44.48%	42.85%	0.03389
2	43.28%	41.39%	0.0466
3	43.59%	41.10%	0.0516
4	43.88%	41.06%	0.1006
5	47.88%	45.35%	0.2016

Table 2.1: Mortality rate table. First column shows mortality rates for each maximum number of hospitalisations for the dataset with patients who experienced a shock. Second column shows mortality rates for each maximum number of admissions for the dataset without patients who experienced a shock. Third column shows p-values for the null hypothesis that the proportions from the two groups are the same.

Another interesting feature of our dataset is the patients' age. As said at the very beginning of this chapter, patients younger than 18 years, being it their age at the first admission, were not included in the study. This way, the minimum age is 18 while the maximum is 103 years. Over the five groups, patients with the same number of events, there is a modest difference in the distribution of the patients' ages, as we can see from Figure 2.4. In particular, the mean value for the age variable over the whole dataset is equal to $\mu = 75.77\, years$, with a standard deviation equal to $\sigma = 12.69\, years$, while, if we divide all patients according to the maximum number of admissions, we find that the second group is the one resembling the most to the complete dataset ($p - value = 0.6348$, obtained with a Kruskal-Wallis rank sum test). For all the other groups there is evidence to state that the age distribution is different compared to that of the whole dataset. This result is probably due to the distribution of patients over the five groups: actually the mean number of hospitalisations for the complete dataset is equal to 1.987, implying that the equivalence to the entire dataset can only be find in the second group. Despite this difference between the groups and the total dataset, nevertheless we can consider the mean and standard deviation of the complete dataset as representative of all the groups.

For these reasons it is very fascinating to look at the histogram of the age variable (see Figure 2.5). Before the most frequent class, the frequency increment has an exponential trend while, after that, it decreases rapidly showing the expected trend due to the ageing of the patients and the consequent rise of the mortality rate, resulting in an obvious reduction of the number of observation units that could have the specific characteristics required by our study trial. This result is perfectly consistent with what is today well known in HF epidemiology studies. In fact, we can find in literature that the average age of patients with this kind of disease is around

75 years, and that this pathology is the leading cause of hospitalisation in people older than 65 [15].

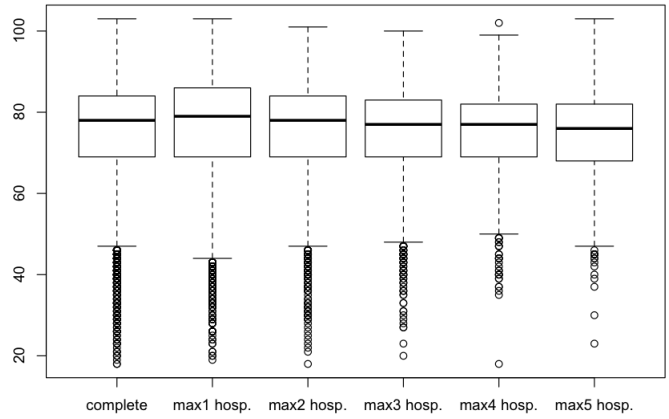


Figure 2.4: Age boxplot: the first one is the boxplot of the age variable for our dataset, the others are the boxplots for the same variable but in the five groups.

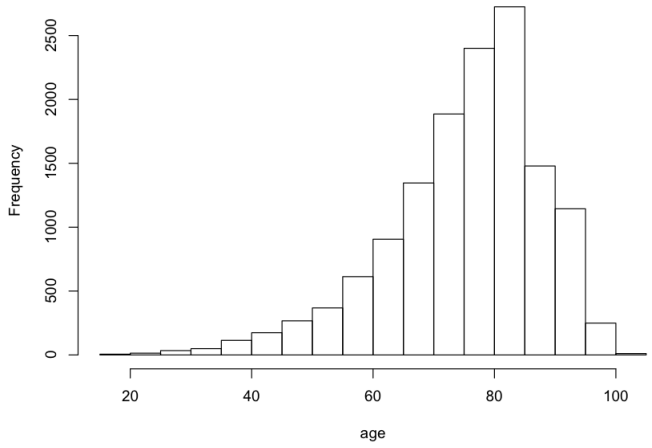


Figure 2.5: Histogram of patients' age at first hospitalisation from the complete dataset.

Another relevant factor in this analysis is the sex of the observation units [22]. First of all, it is important to state that our dataset is slightly unbalanced between the two genders, with 46.2% of men and 53.8% of women. In medical literature we find that there are significant differences between men and women affected by chronic HF, mostly because of the main causes that led to this pathology in the two kind of subjects. Moreover, it seems that women manifested with heart failure at older age than men, a trend that can

also be find in our dataset (see Figure 2.6). Conducting a Wilcoxon test over our dataset, we obtained evidence in support of the alternative hypothesis that men's mean age is smaller then that of women ($p-value < 2.2\times10^{-16}$). Another relevant equivalence between what is well known in cardiology literature and our analyses, is that women appeared to undergo less surgery interventions, as we find that between the patients who had "ICD", 83.2% of them were men and similar statistics are find for the other factors: 76.17% of units with "CABG" are of male gender as the 67.93% of patients who underwent "PTCA".

Finally we would like to briefly discuss the mortality rate in the two genders. Men and women have around 10% difference in the mortality rate (37.15% vs 46.43%), but still in conformity with the mortality rate of our dataset. This result is probably the outcome of the discard of patients who had a circulatory shock, as there is a higher percentage of men who experienced this kind of pathology.

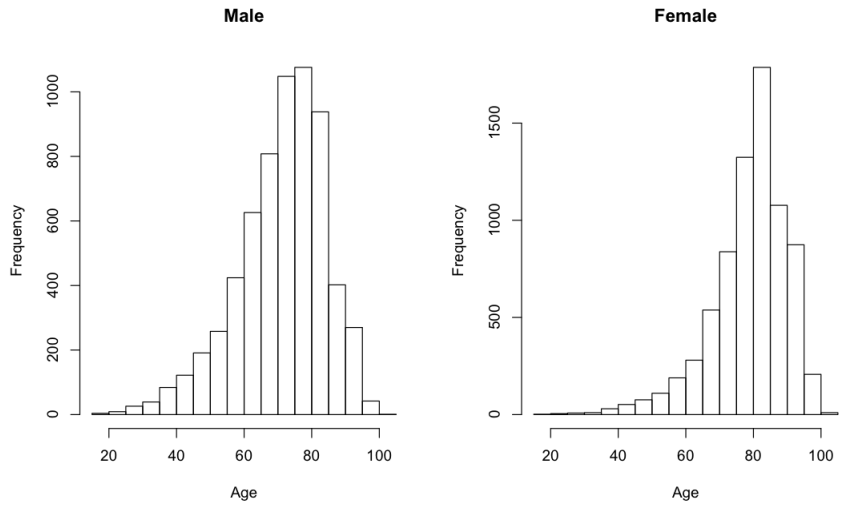


Figure 2.6: Histogram of age for female and male gender from complete dataset.

Chapter 3

Patients clustering

In our analysis we used the R package `mixPHM` [16]. The package fits multiple variable mixtures of various parametric proportional hazards models using an EM-Algorithm. It was originally implemented for the study of dwell-time-based sessions clustering with incomplete data [17]. The underlying model is a proper model for our dataset as it allows missing data in the maximum likelihood equation for a mixture model. In particular, the authors solved this issue introducing in the model a "prior" probability estimated by the corresponding relative frequency. This way we are able to estimate the likelihood for the model (E-step) and then compute the M-step: the package allows for several possibilities of achieving the maximisation step for the classification of our data. Moreover, we see a strong evidence in the correlation between "survival times" in medical statistics and "dwell times" in web usage mining, and for this reason we apply the same methodology to our dataset. We used this package in order to get a clustering of our observation units based on the time elapsed between two consecutive events.
Before the discussion on the method and its results, we need to make some basic assumptions on our dataset: we will consider the time intervals between two consecutive events to be independent and will hypothesize that they are Weibull distributed with different parameters over the resulting clusters.

3.1 The model: Weibull Mixtures and Proportional Hazards Models

Survival analysis deals with the duration times until some event occurs, where in our study the new event corresponds to a new admission. The analysed survival times in our model are the times between two hospitalisations, meaning that for each patient we have a maximum of five possible

time intervals: between the first and the second admission, between the second and the third, and so on, up to the fifth hospitalisation where the following event is the end of the study. Of course there are patients who have less than five admissions: for these patients the subsequent event could be the decease or the end of the study, if they survive. This said, from now on we will refer to the variable "hospitalisation time" as the time interval between two successive events and will denote it as T_h, with $h = 1, \ldots, H$, where $H = 5$. For example, the time of the second hospitalisation will be the survival time between the second and the third admissions. As already stated, we assume for each survival time between two events to follow a Weibull distribution with density function $f(t) = \gamma/\lambda^{\gamma}\, t^{\gamma-1} exp(-(t/\lambda)^{\gamma})$, where λ is the scale parameter and γ the shape parameter. Notice that the model presuppose that we know the number K of clusters in which we will try to split our dataset. We will choose $K = 3$, but will come back to this choice and its reasons in Section 3.2. In addition, as we are stating that each time segment is Weibull distributed, each different hospitalisation time has its own parameters, which are different from the parameters of the previous and the following T_h and from one cluster to the other. This means that the model for this kind of conjecture is a *Weibull Mixture Model*, such that for each hospitalisation time the resulting mixture density has the following form:

$$f(t_h) = \sum_{k=1}^{K} \pi_k f(t_h; \lambda_{kh}, \gamma_{kh}), \tag{3.1}$$

where

$$f(t_h; \lambda_{kh}, \gamma_{kh}) = \frac{\gamma_{kh}}{\lambda_{kh}^{\gamma_{kh}}}\, t_h^{\gamma_{kh}-1} exp(-(t_h/\lambda_{kh})^{\gamma_{kh}}). \tag{3.2}$$

t_h is a possible realisation for the h-th hospitalisation time T_h, $\forall h = 1, \ldots, H$, and $\pi_k > 0$ are the mixing proportions that satisfy the condition $\sum_{k=1}^{K} \pi_k = 1$. This way the parameters of the Weibull mixture model are $2(K \times H) + K$, corresponding to K weights π_k for $k = 1, \ldots, K$, $K \times H$ scale parameters λ_{kh} for $k = 1, \ldots, K$ and $h = 1, \ldots, H$, and $K \times H$ shape parameters γ_{kh} for $k = 1, \ldots, K$ and $h = 1, \ldots, H$. Hence the model's parameters are of the following form:

$$\Lambda = \begin{pmatrix} \lambda_{1,1} & \ldots & \lambda_{1,H} \\ \vdots & \ddots & \vdots \\ \lambda_{K,1} & \ldots & \lambda_{K,H} \end{pmatrix} \tag{3.3}$$

where λ_{kh}, for $k = 1, \ldots, K$ and $h = 1, \ldots, H$, is the scale parameter for the h-th hospitalisation time T_h in the k-th cluster;

$$\Gamma = \begin{pmatrix} \gamma_{1,1} & \cdots & \gamma_{1,H} \\ \vdots & \ddots & \vdots \\ \gamma_{K,1} & \cdots & \gamma_{K,H} \end{pmatrix} \tag{3.4}$$

where γ_{kh}, for $k = 1, \ldots, K$ and $h = 1, \ldots, H$, is the shape parameter for the h-th hospitalisation time T_h in the k-th cluster, and

$$\Pi = \begin{pmatrix} \pi_1 & \ldots & \pi_K \end{pmatrix} \tag{3.5}$$

where π_k, for $k = 1, \ldots, K$, is the weight parameter for the k-th cluster.

We herein recall that a hazard function for any distribution with density function f and cumulative function F is computed as the ratio between the density function and the survival function, $h(t) = f(t)/(1 - F(t))$. The resulting hazard function from this model choice on our dataset for the h-th hospitalisation time T_h in the k-th cluster is of the following form:

$$h(t_h; \lambda_{kh}, \gamma_{kh}) = \frac{\gamma_{kh}}{\lambda_{kh}^{\gamma_{kh}}} t_h^{\gamma_{kh}-1} \tag{3.6}$$

Moreover, in [17], the authors of `mixPHM` package took advantage in their model from an important class of models in the survival analysis: proportional hazards models (PHMs). In Appendix A the reader can find a brief summary of the origin of this class of models for survival data. The main idea behind these kind of models is that the hazard function $h(t)$ is equal to the product of a baseline hazard function $h_0(t)$ and a factor quantifying the influence of some predictors. In our model we are not interested in adding, at this stage of our analysis, any kind of covariates, as we want to obtain a clustering method of our data based uniquely on the hospitalisation times we acquire from the admissions' data of each patient. For this reason we set the covariates vector equal to the unit vector $\mathbf{1}$. Notice that this procedure is also suggested by the authors for situations like ours. This way the hazard function in our study project becomes of the form:

$$h(t_h|k; \mathbf{1}) = \gamma_{kh} t_h^{\gamma_{kh}-1} exp(\mathbf{1}\beta^{(k,h)}) \tag{3.7}$$

As a result of this new interpretation of the model, the scale parameter becomes of the form $\lambda_{kh} = [exp(\beta^{(k,h)})]^{-1}$.

We now come to the resolution of the problem of missing data. We deal in our dataset with right-censored data, as not all our patients die before

the end of the study, leading to a natural right censoring for their survival times. Moreover, as a comparison between survival times and dwell times, all of our patients visit the first "state", corresponding to the hospitalisation time between the first and the second admissions. But not all of them visit the other "states", as already seen in previous chapter. In order to overcome the missing data problem and the consequent limitation of the EM-algorithm used in the clustering process, the authors introduced into the model the probability $\tau_h(i|k)$ that a patient i, knowing that she/he is in the k-th cluster, is admitted h times. This way, in addition to the parameters of the Weibull mixture model, the package will estimate these probabilities too, for an overall number of parameters equal to $2(K \times H)+K+(K \times H) = 3(K \times H)+K$. Using these probabilities we are able to compute the likelihood for patient i being in cluster k for each hospitalisation time h individually: if the h-th hospitalisation was not reached by the i-th patient we compute the likelihood as $1 - \tau_h(i|k)$, and we therefore multiply, where applicable, the known likelihood value by $\tau_h(i|k)$.

3.2 Using the package

In order to use package `mixPHM`, we had to build a times matrix suitable for the package requests. To compute the mixture hazards model and the clustering of our observation units, as explained in previous section, we used the package function `phmclust`. It allows, as valid argument, a matrix of dimension $n \times H$, where the matrix elements are survival times. "`NA`" values are allowed, too.

In order to compute the required survival times we decided that, for each patient, her/his first survival time started at calendar date corresponding to their first admission, even if there are several patients whose first hospitalisation started before January 1st, 2006. For this reason the overall study period is equal to 1931 days, corresponding to 5 years (from the beginning of 2006 to the end of 2010) and few months (the ones before 2006). Our survival times are then computed as follows: the realisation $t_1(i)$ for first hospitalisation time T_1 of patient i is computed as the days elapsed from the date of the first admission to the date of next event, which can be a new hospitalisation, the death of the patient, or the end of the study (December 31st, 2010). If the patient died after the first admission or experienced no other hospitalisations, her/his following survival times are marked as "`NA`". If the patient has another hospitalisation event after the first one, her/his next survival time, corresponding to realisation $t_2(i)$, is determined as the elapsed

days between the second admission date and the following event, and then the computing procedure repeats for the remaining hospitalisation times T_3, T_4 and T_5. Remark that for the last hospitalisation time variable T_5, the last events are of two kinds only: a unit can die after its last hospitalisation or can survive until the end of the study. No other events are allowed, as we decided to consider the population with a maximum of five hospitalisations over the study time period.

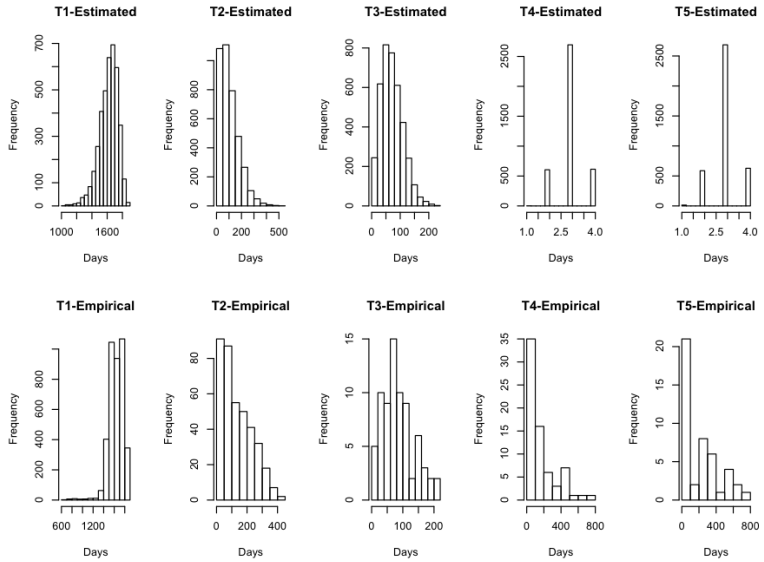
Following this procedure, we are able to build the survival times matrix necessary to use package `mixPHM`. Our matrix will then be made of $n = 13785$ rows, one for each patient in our study trial, and $H = 5$ columns, one for each possible hospitalisation. It is very important to remember that in our dataset every patient has the first hospitalisation, as if, in a comparison with the dwell-times problem, every user would visit a specific page, for example the "home page" of a particular website. This is a situation which is not so far from being a good approximation of what happens in reality, and for this reason our dataset is suitable to be studied by means of the proposed model. All the other matrix columns will have some values equal to "`NA`", in such a way that the number of "`NA`" found in each column will increase from the second to the last column, according to the characteristics of the dataset.

There are other arguments requested by function `phmclust` in order to compute the hazard mixture model and to obtain the desired clusters. Here we will discuss only the three most relevant ones. First of all we need to state the number of clusters in which we want to split our dataset, which can be done by setting the value of argument K. As already declared, we set $K = 3$ in order to divide our units into three natural groups: "healthy", "sick" and "terminally ill" patients. Notice that every patient in our dataset enters because her/his first hospitalisation ended within 2006, but, as for any disease, the diagnosis can be done at very early stages or at final ones. This is why we expect to obtain three clusters with this algorithm, and we presume they will exactly represent the natural groups previously highlighted.

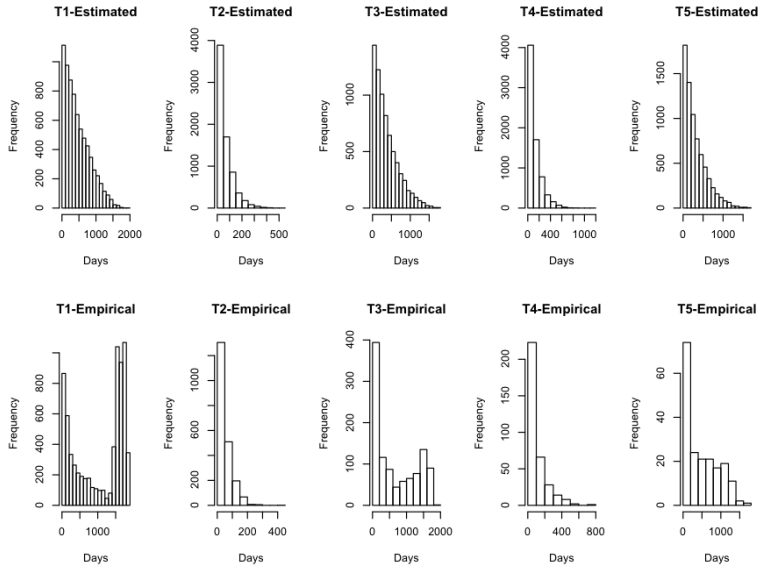
Another important parameter to be set in order to correctly compute the algorithm, is the underlying distribution of the hazard function $h(t)$. The `phmclust` function allows for different distributions: Weibull, Exponential and Rayleigh. To verify that our survival times real distribution is better fitted, among all the possibilities, by a Weibull distribution, we computed the algorithm and the clusters for each of the possible distributions. Here we discuss the results obtained with the fit of Weibull and Exponential distributions, which were the most suitable ones. These two distributions have a strong connection: when we set the shape parameter of the Weibull distribution equal to 1, it corresponds to the Exponential distribution. In

order to understand why we decided to consider the survival times Weibull distributed, we take as example the third cluster obtained running the algorithm firstly imposing the Weibull distribution, then the Exponential distribution (see Figure 3.1). First of all notice that variables T_4 and T_5 have too many observations with "`NA`" values, a characteristic that excludes the possibility of making any kind of considerations about their distribution. For this reason we took into account only the first three hospitalisation times, T_1, T_2 and T_3. From the examined figure, we can truly appreciate the correctness of the Weibull distribution fitting the empirical distribution of the hospitalisation times from our dataset. In particular, using the Exponential distribution, we lose the ability to estimate the right tails of empirical distributions, corresponding to censored observations. This characteristic can be better appreciated if we look at the distribution of the first hospitalisation time T_1 in the provided example (see Figure 3.2): if we use the Weibull distribution, the algorithm perfectly approximates the shape of the corresponding empirical distribution, whilst the Exponential distribution fails to estimate the right tail.

Finally we imposed no proportional restrictions to our model, setting the `phmclust` argument `method` equal to `separate`. This way we force our model to take into account for no covariates in the proposed proportional hazards model. The corresponding model is the one reported in (3.7).



(a) *Weibull distribution for survival times in the third cluster.*



(b) *Exponential distribution for survival times in the third cluster.*

Figure 3.1: Comparison between the use of Weibull or Exponential distribution. First row shows the histograms of estimated survival times for each hospitalisation time variable obtained when computing the algorithm under the hypotheses that empirical survival times are Weibull or Exponential distributed. Second row shows corresponding histograms of empirical survival times from the units which have been assigned to the third cluster under these conditions.

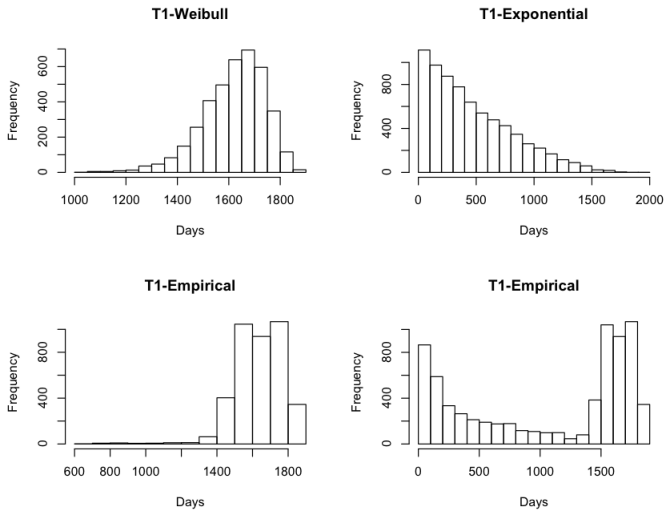


Figure 3.2: Comparison between the goodness of fit for the empirical distribution using a Weibull or Exponential distribution. These plots are the same from Figure 3.1 in the first column.

3.3 Analysis of the results

Once we run `phmclust` function, several values are returned for the utility of the user. The most important for our analysis are: the estimated shape and scale parameters, $\hat{\lambda}_{kh}$ and $\hat{\gamma}_{kh}$ for $k = 1, \ldots, 3$ and $h = 1, \ldots, 5$, and the final deterministic assignment of all observations to the most suitable cluster k, with $k = 1, \ldots, 3$. Here we report the results we obtained for the model parameters:

$$\hat{\Lambda} = \begin{pmatrix} 331.24877 & 670.59967 & 315.22627 & 331.22789 & 272.83983 \\ 176.20118 & 33.28450 & 1565.97558 & 239.28680 & 1.67868 \\ 1696.36323 & 144.98267 & 84.64472 & 2.70787 & 2.70787 \end{pmatrix} \tag{3.8}$$

$$\hat{\Gamma} = \begin{pmatrix} 0.86552 & 1.01331 & 0.77148 & 0.76640 & 0.68695 \\ 0.66403 & 0.96890 & 10.92267 & 4.28698 & 3.46154 \\ 16.48662 & 1.29069 & 1.99473 & 5.9754 & 5.91754 \end{pmatrix} \tag{3.9}$$

Once we obtained the estimated parameters and the corresponding division of all patients into three clusters, we should first of all analyse the cluster-profiles shown in Figure 3.3. From this plot we can promptly see that the first cluster is some sort of mean of the other two clusters, as all the mean values of survival time for each hospitalisation time variable T_h in this cluster fall into the interval $[350; 700]$, being this the central band of all

the possible mean values from the analysed dataset. Notice that this cluster has only two values grater then the other two clusters, for T_3 and T_5. We will show, while analysing in deep the results, that this result is the natural consequence of the particular characteristics of second and third clusters. Having said that, we are already able, at this very early stage of our analysis, to appreciate that the second cluster is the one with the lowest values in the first two survival time variables, skimming the lower bound value of zero for the mean survival time of the second variable T_2. On the other hand, the third cluster reaches the highest mean value of the whole dataset in the very first survival time variable, approaching the upper bound of 1931 days with a mean survival time value of 1634.903 for T_1. The shown peculiarities of the three clusters are already anticipating the very deep nature of the obtained groups: in a first attempt to give them names, the first cluster would be that of "sick" patients, the second one of "terminally ill" patients and the last one that of "healthy" subjects. We will try to appreciate the characteristics of each cluster in the following subsections.

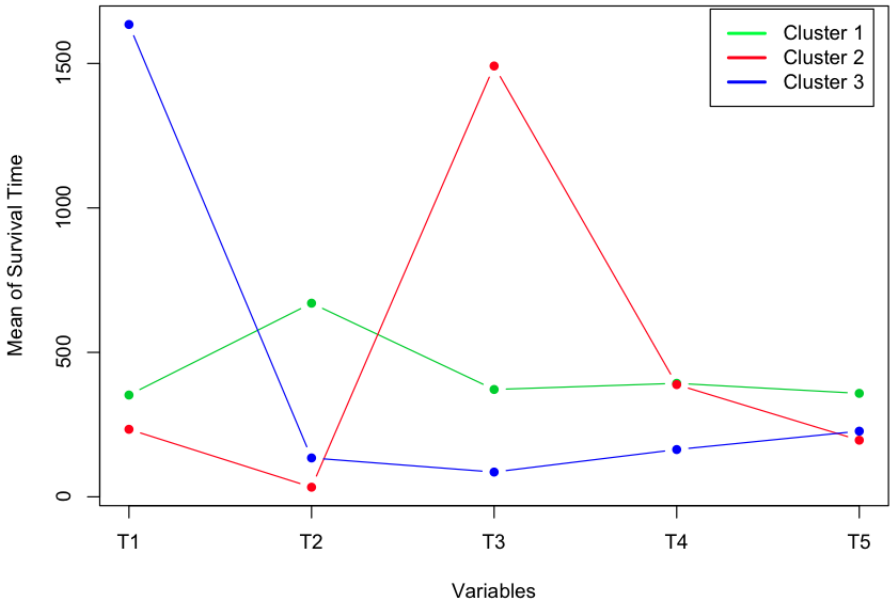


Figure 3.3: Cluster-Profiles: Mean survival time for each variable T_h, $h = 1, \ldots, 5$, and for each cluster k, $K = 1, \ldots, 3$.

3.3.1 First cluster: "sick" patients

First of all we need to take a deeper look into the first cluster. This is the most numerous one, with a total of 5885 patients corresponding to 42.69%

of the dataset. In this cluster, there is the same number of men and women. In Figure 3.4 and Figure 3.5 we can see the age distribution from patients in the first cluster. In this cluster, as for the total dataset, we find that the mean age is coherent with what is reported on medical journals, being it equal to $\mu_1 = 74.35\, years$ with standard deviation $\sigma_1 = 11.95\, years$. Conducting a Wilcoxon test to compare the total dataset age distribution and the age distribution of this cluster, we obtain a strong evidence in favour of the alternative hypothesis that the cluster's age distribution is shifted to the left of that of the total dataset ($p - value = 3.238 \times 10^{-4}$ imposing as alternative hypothesis that the true location shift is less than -1, the first value for which we obtain a significant $p - value$).

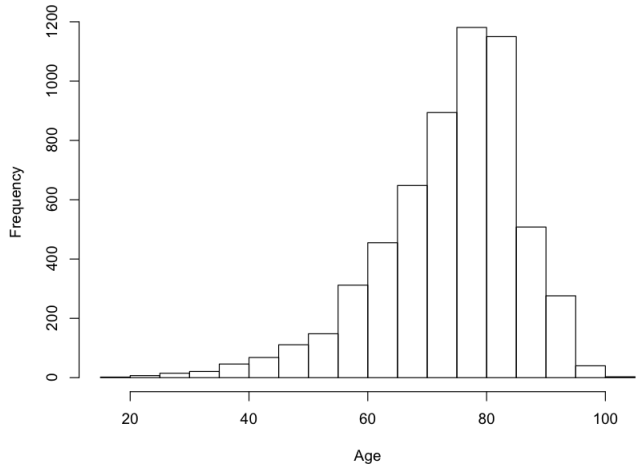


Figure 3.4: Age histogram for patients in the first cluster.

There is a very interesting feature of this cluster, which can be appreciated in Figure 3.6: all patients who were identified as being part of this cluster have at least two consecutive hospitalisations. This is relevant in our analysis because it is showing that this cluster is formed by "sick" units, as their mortality rate (34.15%) is close to that of the complete dataset, but they all have the characteristic of living longer throughout the study time window. As we can see from Figure 3.7, the percentage of patients who died during the h-th admission is an increasing function, asymptotically going to the 50% value, with a monotonic trend except for the forth hospitalisation time T_4, which slightly diverges from the functional behaviour. This is exactly what we would expect from a general analysis of a complete dataset of patients affected by chronic heart failure and this is why we consider this first cluster the "mean" cluster, as it represents the trend of the complete dataset.

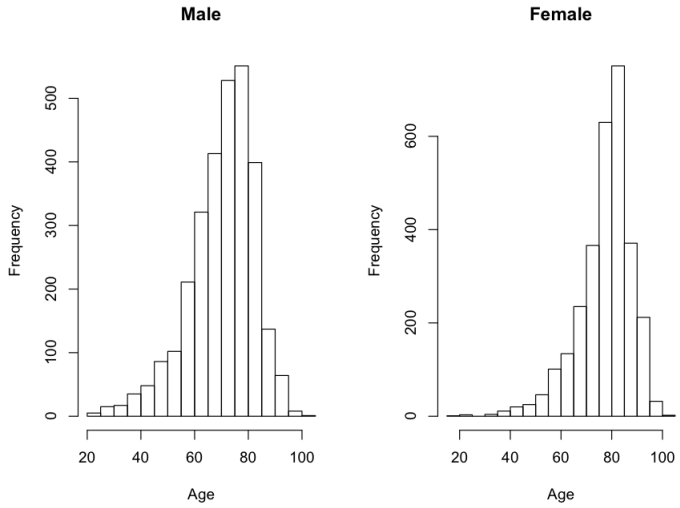


Figure 3.5: Age histogram for men (on the left) and women (on the right) in the first cluster.

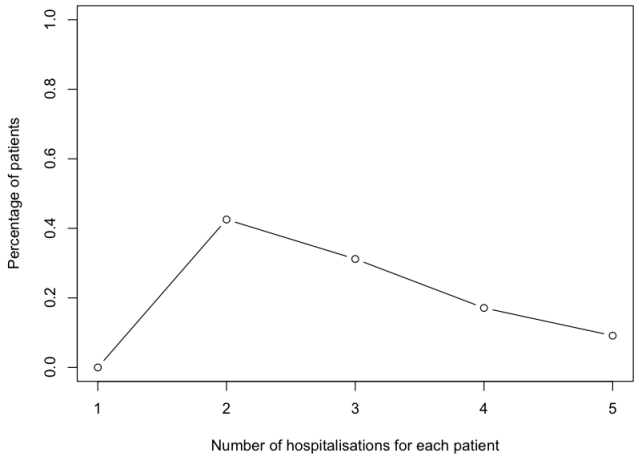


Figure 3.6: Percentage of patients in the first cluster who have only one hospitalisation over the study time period, or just two admissions, or three and so on. It is of a great importance to notice that in this first cluster no patients have just one admission, but they all have at least two.

3.3.2 Second cluster: "terminally ill" patients

Analysing the second cluster, which is the 28.91% of complete dataset, what is immediately evident is that this cluster represents and is constituted almost only by "terminally ill" patients. This is clear if we take a look at the mortality rate of subjects belonging in the group and at its distribution through the five hospitalisation time variables. The mortality rate of

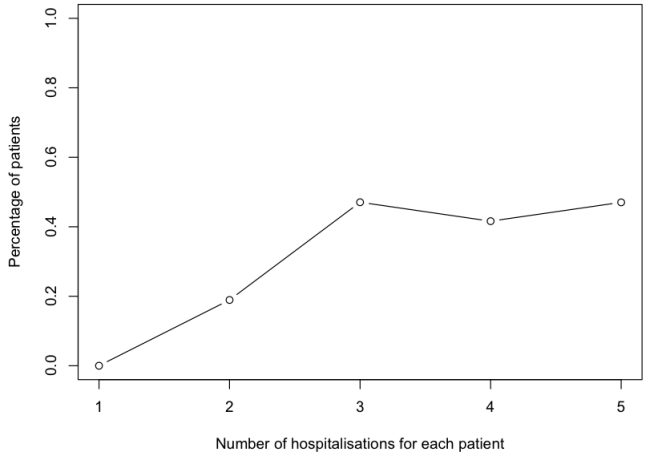


Figure 3.7: Percentage of patients in the first cluster who died during their first hospitalisation over the study time period, or during their second one, or third and so on.

the cluster is equal to 92.87%, considerably higher than the mortality rate of the complete dataset (see Section 2.3). Moreover, as we can see in Figure 3.8, in the first two hospitalisation time variables all patients die, meaning that they had just one or two admissions not because they are healthy subjects, but because they are at the terminal stage of their disease. In particular, those who have just one or two hospitalisations cover the 92.72% of the cluster's units.

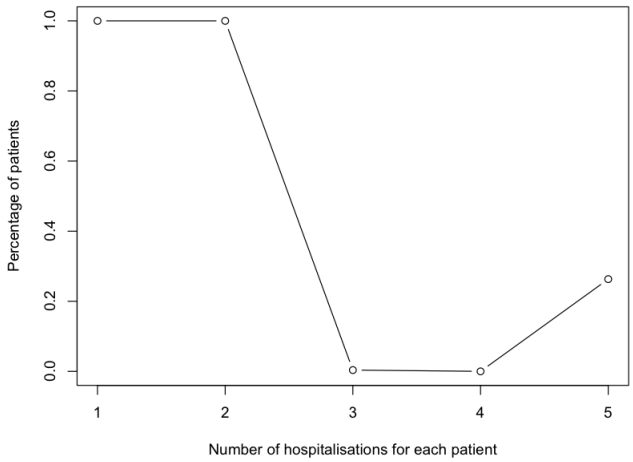


Figure 3.8: Percentage of patients in the second cluster who died during their first hospitalisation over the study time period, or during their second one, or third and so on.

Now that we convinced ourselves that patients in this cluster correspond to "terminally ill" ones, it is even more interesting to analyse the distribution of patients' age. As a matter of fact, we can see from Figure 3.9 that the population of this cluster is older than that of the complete dataset. A strong evidence to confirm this idea can be found examining the mean age of this cluster, equal to $\mu_2 = 82.14\, years$ with standard deviation $\sigma_2 = 10.28\, years$. The Wilcoxon test, comparing the age distribution of the complete dataset and the second cluster, gave strong evidence ($p - value = 4.568 \times 10^{-6}$) in favour of the alternative hypothesis that the true location shift is greater than -5, being this the first value for which we obtained a significant $p-value$. This means that, being the population of the cluster significantly older than what it would be for a general chronic HF dataset, the natural consequence is that the mortality rate increases compared to that of the complete dataset.

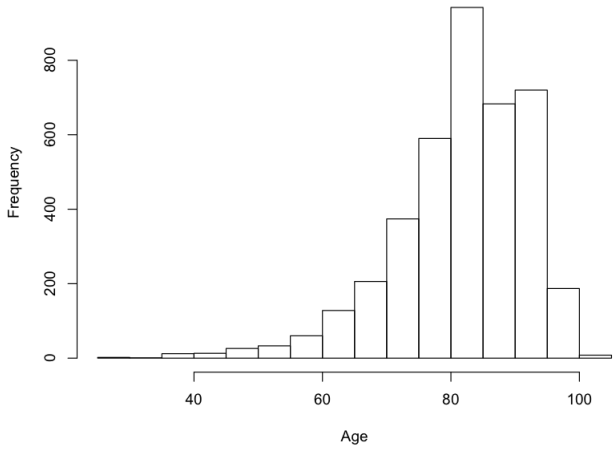


Figure 3.9: Age histogram for patients in the second cluster. It is clear that this population is considerably older compared to the population of the complete dataset (see Figure 2.5).

In this cluster women are more than men, as it is in the complete dataset. In particular, 60.5% of the unites are of female gender and 94.73% out of these died within the first two admissions. The high mortality rate of women in this cluster, compared to that of men (equal to 90.04%), is even more evident and has a natural explanation if we look at Figure 3.10, where we can see from the age distribution of men and women that women are considerably older than man in this group. Actually, the mean value for the women's age is $\mu_{2F} = 84.36\, years$ while that for men is $\mu_{2M} = 78.76\, years$. The related 90% Wilcoxon test shows evidence ($p - value = 1.409 \times 10^{-6}$) in favour of the alternative hypothesis that the true location shift for men's age distribution compared to that of women is less than -4, being this the

first value to return a corresponding significant $p - value$. Not only we see that men are younger, but also that, into this cluster, units of male gender preserve some of the characteristics of the complete dataset, although they have a higher mortality rate.

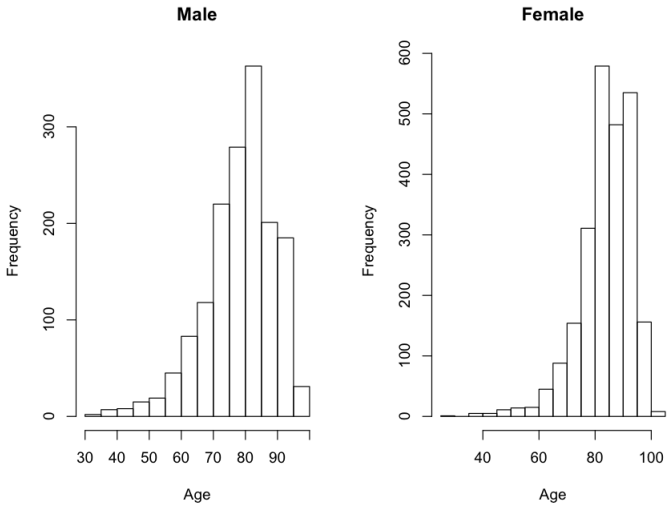


Figure 3.10: Age histogram for men (on the left side) and women (on the right side) in the second cluster.

3.3.3 Third cluster: "healthy" patients

Finally we need to take a deep look into the last cluster and, as said, we expect it to be the cluster of "healthy" patients. In order to do so, first of all we need to analyse the very basic characteristics of this cluster and compare them with the same characteristics of the complete dataset and of the other clusters. Cluster number three is the 28.4% of the entire dataset, with 3915 units being assigned to it. Out of these, 99 died before the end of the study time period, a number equivalent to 2.53% of the population in this group. We find immediate proof to what we were expecting: subjects in this cluster are the ones that are healthier than the ones assigned to the other clusters, as the very majority of them remains alive during all five years of observation. From this very first analysis we are already able to state that this is, indeed, the cluster of "healthy" patients. In order to reinforce this statement, we need, first of all, to look at Figure 3.11, where is shown the percentage of patients who died during their h-th hospitalisation. The function is an increasing function, starting from very low values (percentage of dead patients for T_1 is equal to 0.31%) and asymptotically reaching the

upper bound 33.33%. This trend is even more indicative of the healthiness of the subjects in this cluster if we compare it with Figure 3.12. As we can see, the majority of patients (90.22%) in this third group have only one hospitalisation and, as already discussed, just few of them died. The remaining 9.78% of the cluster is distributed between the other hospitalisation time variables, in such a way that the high values of the mortality rate obtained for T_3, T_4 and T_5 and reported in Figure 3.11, are completely insignificant to the understanding of the cluster's characteristics. Moreover, we can say that the representative units of this cluster are those experiencing only the first hospitalisation, corresponding to variable T_1, during the study time period.

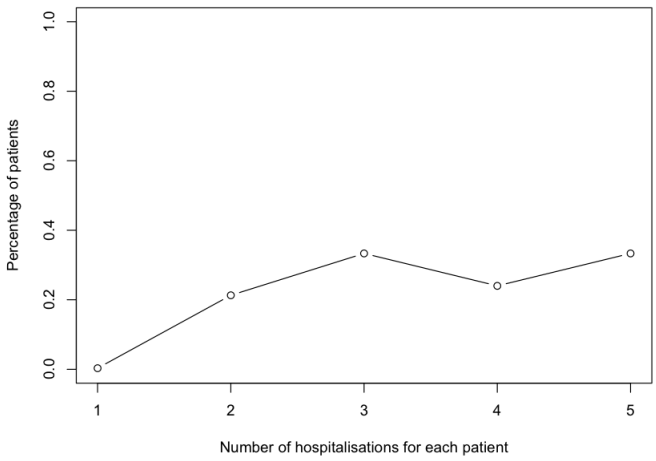


Figure 3.11: Percentage of patients in the third cluster who died during their first hospitalisation over the study time period, or during their second one, or third and so on.

At this point, it becomes relevant to see if, together with the characteristic of being constituted by patients in their first disease stage, this cluster also has the peculiarity that the mean age of its units is lower than that of the complete dataset and of the rest of the clusters. As expected, we obtain that for this cluster $\mu_3 = 71.42\, years$, which is a considerably lower mean age than that of the complete dataset. The corresponding Wilcoxon test gives evidence ($p - value = 1.536 \times 10^{-7}$) for the alternative hypothesis that the true location shift of the cluster's age distribution compared to that of the complete dataset is less than -3, being this the first value to return a corresponding significant $p - value$. Moreover, the mean value is considerably lower than that of the second cluster, being it the "terminally ill" patients cluster ($p - value = 4.535 \times 10^{-5}$ for the corresponding Wilcoxon test with true location shift less than -9, the first value for which we obtained a significant $p - value$). This result can also be seen trough the age histogram of

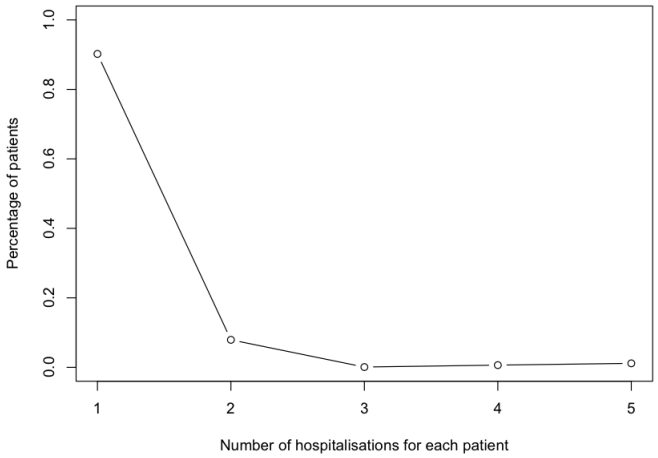


Figure 3.12: Percentage of patients in the third cluster who have only one hospitalisation over the study time period, or just two admissions, or three and so on.

the third cluster (see Figure 3.13), compared to that of the complete dataset (see Figure 2.5).

Finally, we want to extend our analysis to the comprehension of the gender distribution in this third cluster. Remember that, as we already said in Section 2.3, women happened to present with heart failure at older age than men. Here we find the same result, with an even more interesting outcome: mean age for women ($\mu_{3F} = 75.71\, years$) is perfectly matching what is expected to be for the average of patients with this disease (Wilcoxon test to compare women age of this cluster and age of complete dataset gives no evidence for the alternative hypothesis that true location shift is not equal to 0, $p - value = 0.56$); on the other hand, for men the mean age is considerably lower than what expected, being it equal to $\mu_{3M} = 66.65\, years$ (Wilcoxon test for the comparison between men and women age gives evidence, $p - value = 1.219 \times 10^{-7}$, in favour of the alternative hypothesis that the true location shift is less than -7, being this the first value for which the test results in a significant $p - value$). This characteristic behaviour of men age distribution in this cluster reflects what is known in medical literature and can also be appreciated through the age histogram for the two genders in Figure 3.14. As for the complete dataset, there are more women than men, this being the reason why the overall mean age is equal to $\mu_3 = 71.42\, years$, with a higher standard deviation $\sigma_3 = 13.49$.

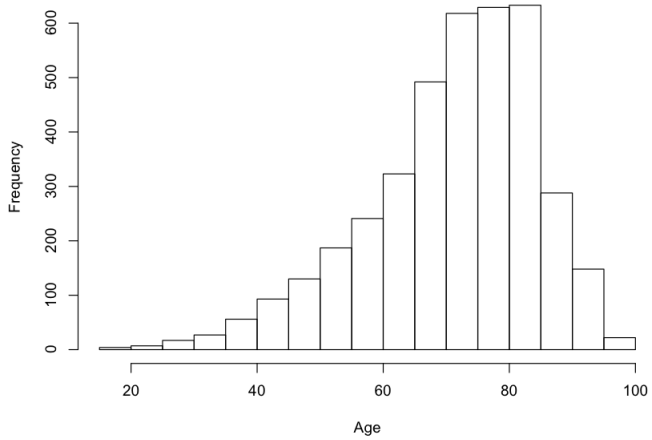


Figure 3.13: Age histogram for patients in the third cluster.

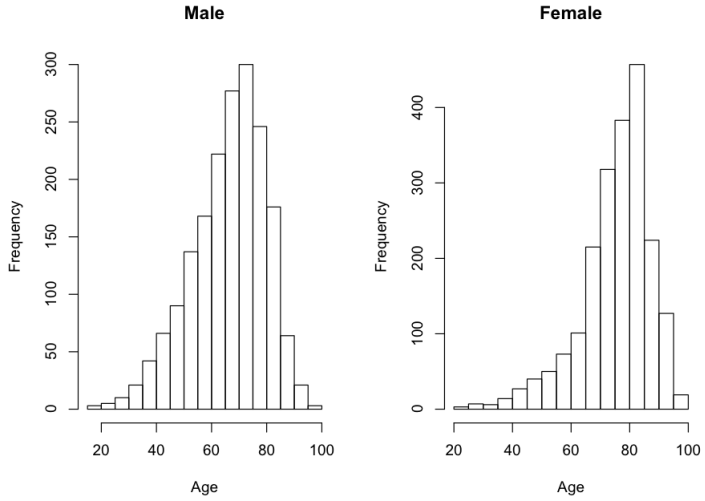


Figure 3.14: Age histogram for men (on the left side) and women (on the right side) in the third cluster.

3.4 Survival analysis and comparison between the three groups: Kaplan-Meier estimator

To give a more refined estimate and explanation of the behaviour of patients in the obtained clusters, we took advantage of the Kaplan-Meier estimator in order to conduct a survival analysis for each cluster and to compare the results.

3.4.1 Definition and properties

The Kaplan-Meier estimator, also known as the product-limit (PL) estimator [14], was introduced by E. L. Kaplan and P. Meier to compute the estimate of a population probability of surviving beyond t, i.e. $P(t) = \mathbb{P}(T > t)$, where T is the survival time variable for the population. The main idea which stimulated this theoretical approach is that the reduced-sample (RS) estimate of the same survival probability function is not suitable enough when we deal with right-censored data, like in most of medical datasets.

To define the Kaplan-Meier estimator, first of all we need to divide the observation time $(0; t)$ into several disjoint intervals, namely $(0; u_1)$, $(u_1; u_2), \ldots, (u_{k-1}; t)$. To identify these intervals, there are mainly two strategies: first, one can determine u_j in such a way that only one observation unit "dies" in the corresponding interval $(u_{j-1}; u_j)$; second, the same strategy can be applied with "loss" events, imposing that only one observation is "lost" in the interval $(u_{j-1}; u_j)$. In this section we talk about "death" and "loss" as two possible conditions for the observation units, where in general they do not strictly relate to the literal meaning. In our study instead, the death of a unit corresponds to the actual death of our patient, while the loss represent the censored subject, for whom we have no information regarding her/his death event as she/he survives beyond December 31st, 2010. Remember that one can always choose a different method to determine the intervals $(u_{j-1}; u_j)$, with the insight of using the correct approximation for the resulting estimator.

Once we selected the desired set of intervals, the algorithm requires to compute the proportion of items alive after u_{j-1} and that survive beyond u_j. We will call this proportion p_j. Finally the survival probability for the considered population is equal to the product of $p_j, \forall j = 1, \ldots, k$.
Lets call δ_j and λ_j the number of deaths and losses, respectively, in the j-th interval, $(u_{j-1}; u_j)$. Being n_j the number of items under observation after

u_{j-1}, the estimate of p_j is equal to:

$$\hat{p}_j = \frac{n_j - \delta_j}{n_j} = \frac{n'_j}{n_j}, \tag{3.10}$$

where n'_j is the number of units under observation after the δ_j deaths. It is obvious that if an interval contains only losses, the corresponding estimate for p_j is equal to $\hat{p}_j = 1$. The resulting PL estimator is the product of all the $\hat{p}_j$ computed in the different intervals, i.e.:

$$\hat{P}(t) = \prod_{j=1}^{k} \frac{n'_j}{n_j}, \tag{3.11}$$

where we assumed that $u_0 = 0$ and $u_k = t$, that $n'_j = n_j - \delta_j$ and that $t < t^*$, if t^* is the greatest observed lifetime and it corresponds to a loss unit. In the case that one wishes to consider only one death for each interval, then the corresponding PL estimator is of the form:

$$\hat{P}(t) = \prod_{r:t'_r \leq t} \frac{N - r}{N - r + 1}, \tag{3.12}$$

where we define t'_r as the ordered sample of death or loss times from our population, such that $t'_1 \leq t'_2 \leq, \ldots, \leq t'_N$, N being the number of observation units at the beginning of the study, i.e. at $u_0 = 0$. We let r run through those positive integers for which t'_r is a time of death, not loss. This simplified model reduces to the previous one if we suppose to cancel same factors rising in (3.12) when two successive r values are computed. Moreover it reduces to the reduced-sample estimator if there are no losses in the observed population, $\hat{P}(t) = n(t)/N$.

As a result to this theoretical construction, it is evident that we can compute $\hat{P}(t)$ for every available t, hence leading to a step function with its discontinuities at the times corresponding to the deaths of one or more patients. The obtained estimator for the survival function is consistent and slightly biased. It can be considered unbiased if the probability of an indeterminate result (when we find $t \geq t^*$) is small (for details see [14]).

3.4.2 Comparing Kaplan-Meier estimator for the three groups

To compute the Kaplan-Meier estimator for our dataset, we took advantage of function `kaplan.meier` from package `spatstat` [2]. This function computes the Kaplan-Meier estimator for all the intervals in which we decided to

divide the overall time of observation, returning a vector of values, each corresponding to the updated estimate of the survival function at that specific time t.

In our dataset, death events correspond to the death of a patient, while the losses are equal to those subjects which survive through the observation time interval and are, for this reason, right-censored. In order to obtain the intervals out of the period during which we monitored patients, we decided to set the division points u_j at the time when at least one death event happens. This way we are obtaining a high quality estimate of the survival function for each time t.

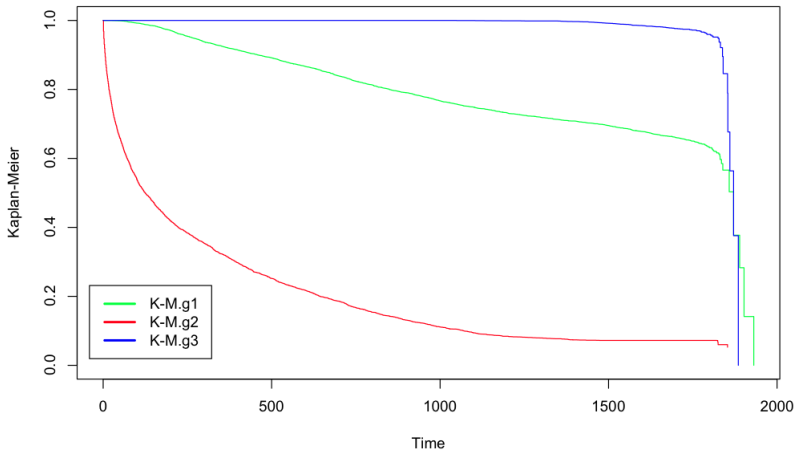


Figure 3.15: Kaplan-Meier estimator function for the three clusters over the five years of observation.

From Figure 3.15, we can see the trend of the survival function estimated through the Kaplan-Meier algorithm for the three clusters: "sick", "terminally ill" or "healthy" patients. First of all it is important to notice that, as we could have expected from the previous analyses (see Sections 3.3.1, 3.3.2 and 3.3.3), the survival function of the second cluster, corresponding to the group of "terminally ill" patients, tends to 0 rapidly, reaching after the first 125 days a probability value of surviving equal to 0.5. On the other hand, "healthy" patients group is characterised by the opposite behaviour, as the obtained survival function is equal to 1 throughout the first three years of observation. Moreover, for the third cluster the survival probability function descends below 0.9 after five years of observation and, when approaching the censoring time, the expectation of surviving beyond the greatest observed lifetime diminishes rapidly. Of course the first cluster, "sick" patients, has a trend which is in between that of the other clusters.

The results from this analysis are perfectly in line with what we could have expected, strongly affirming that the second cluster is that group where the clustering algorithm, used in Section 3.2, collected all patients being characterized by short lifetime values in the first hospitalisations corresponding to death events. At the same time in the third cluster we gathered patients with long lifetime values being right-censored, resulting in the trend just described for the corresponding Kaplan-Meier estimator.

Finally, it could be of some interest to notice that the first cluster, being a mixture of the other two, has a survival trend which resembles better that of the "healthy" patients. This behaviour inspired the interest in deepen the analysis regarding patients surviving after each hospitalisation. This analysis is carried out in Chapter 5. Moreover, it is relevant to notice that the used algorithm doesn't know whether a survival time is censored or not, i.e. if a patient is alive by the end of the study time period. For this reason, despite the good clustering quality obtained, we will see in Chapter 5 that there are some patients who have been misclassified.

Chapter 4

Patient-specific hazard reconstruction

This thesis aims, among others, to reconstruct the hazard function under Cox model [7] for each patient in the study trial. This problem represents a great challenge in statistical literature, and many already confronted with it. Some examples can be found in articles by Peña and Hollander [23], Baraldo, Ieva, Paganoni and Vitelli [3] or Rotolo, Munda and Legrand [20]. An equally great amount of R packages has also been created, see [9], [19] and [1] for examples. Between all the multitude of possibilities, we decided, in order to evaluate patient-specific hazard functions, to take advantage of `frailtypack` package [29]. This package allows to compute Frailty Models, which are a simple extension of the Cox proportional hazards model [28]. The difference from the Cox model is in a multiplicative term which adds random effects in order to estimate the heterogeneity in the studied population. The package was originally written using Fortran 77, and then implemented for use in statistical software R. This package depends on `survival` package [31], [30].

4.1 The model

Frailty models were firstly introduced to model survival data in 1959 by R. E. Beard [4]. The term frailty comes from medicine, referring to feeble people which are characterised by having an increased risk for morbidity and mortality [8]. As a matter of fact, in frailty models the frailty term is introduced as a 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. The shared frailty model is of the form:

$$h_{ij}(t|w_i) = h_0(t)exp(\boldsymbol{\beta}'\mathbf{z}_{ij} + w_i) \tag{4.1}$$

where $h_0(t)$ is the baseline hazard function (see Appendix A), $h_{ij}(t|w_i)$ is the hazard function for patient i at time points t corresponding to her/his j-th hospitalisation, conditionally on the term w_i, and $\mathbf{z}_{ij}$ is the vector of covariates for patient i during the j-th hospitalisation with $\boldsymbol{\beta}$ being the fixed regression coefficient vector. The model (4.1) can also be rewritten as:

$$h_{ij}(t|v_i) = h_0(t)v_i exp(\boldsymbol{\beta}'\mathbf{z}_{ij}) \tag{4.2}$$

where we set $v_i = exp(w_i)$. This term is called the frailty of patient i. The values for v_i in the model are the actual sampled values from a density distribution f_V. In our model we decided, among all possibilities, to consider v_i, for $i = 1, \ldots, n$, independent and identically distributed from a log-Normal distribution, i.e.:

$$v_i \overset{\text{i.i.d.}}{\sim} log - Normal(\mu, \sigma^2), \tag{4.3}$$

$$f_V(v) = \frac{1}{v\sqrt{2\pi\sigma^2}} exp\Big(-\frac{1}{2\sigma^2}(\log v - \mu)^2\Big). \tag{4.4}$$

Consequently, the correspondent term w_i is a realisation from a Normal distribution with mean μ and variance σ^2. In particular, we chose a zero-mean Normal distribution for w_i, leading to the obvious update in the density function for the frailty term v_i:

$$f_V(v) = \frac{1}{v\sqrt{2\pi\sigma^2}} exp\Big(-\frac{1}{2\sigma^2}(\log v)^2\Big). \tag{4.5}$$

This frailty distribution has been largely used for frailty models, although its Laplace transform (used to estimate Kendall's coefficient of concordance, a measure of dependence for bivariate parallel data) is theoretically intractable [11]. There are, anyway, approximation methods in order to estimate the Laplace transform and its derivatives, so that one can, nevertheless, conduct probability evaluations on this model. The correctness of the log-Normal choice can be seen through values of mean and variance of variable V, as the variance of corresponding Normal distribution, σ^2, reaches zero. The mean value for the log-Normal distribution derived from a zero-mean Gaussian is equal to $\mathbb{E}[V] = e^{\sigma^2/2}$; the corresponding variance

is $Var[V] = e^{\sigma^2}(e^{\sigma^2} - 1)$. It is clear that when we let σ^2 tend to zero, the results are as follows:

$$\mathbb{E}[V] = e^{\sigma^2/2} \to 1; \qquad Var[V] = e^{\sigma^2}(e^{\sigma^2} - 1) \to 0. \tag{4.6}$$

The mean value of 1 is the natural restriction usually imposed for other frailty distributions; when the estimated variance of a population tends to zero, this means that there is no heterogeneity among subjects under observation and, as in our situation, in general one wishes to obtain the homogeneity case as a limit case of the model (remember that the frailty parameter has been introduced to estimate the heterogeneity between different statistical units).

The shared frailty model is a particular extension of Cox proportional hazards model, where, as already said, we introduced a random effect parameter to quantify the frailty of patients. As they are, in fact, similar to Cox model, shared frailty models are to be considered non-parametric models, where the baseline hazard function is completely unspecified. This means that with this new approach (compared to that in Chapter 3) we are allowing the model to compute and estimate the most suitable hazard function, without imposing any specific distribution a priori. On the other hand, we differentiate the hazard function for each patient through the frailty term, which is a realisation from a fixed distribution (log-Normal in our case), and over time through the covariate matrix $\mathbf{z}$. Hypothetically, one could let the covariate matrix be time-dependent, in an effort to estimate changes of patients' health through their hospitalisations histories; to build our model, we decided to let only certain covariates be time-dependent, as we will show in next section.

4.2 Using the package

`frailtypack` package computes frailty models for different frailty distributions and different frailty models hypotheses. It depends on package `survival` to handle survival times between events and the kind of event a patient is experiencing. Events are considered as follows: $event = 1$, when a patient is experiencing a new hospitalisation or she/he dies; $event = 0$, when a patient's survival time is censored, for she/he doesn't have new admissions along her/his history nor dies. Survival times are computed as in Section 3.2. In fact, each survival time corresponds to the time elapsed between one admission and the following event, whether this is a new hospitalisation, the death of the patient or the end of the study time period.

The output of function `Surv` from package `survival`, is passed to function `frailtyPenal` of package `frailtypack`. The output is used as the left parameter for a formula object, where on the right we give details of the grouping method for the shared frailty model, i.e. the IDs of patients in our dataset, and values for the covariates we want to use to build our model. For each patient, as stated in Chapter 2, we collected several information: among all we recall the age of each patient at the time of the considered event and whether a subject underwent some kind of surgical or non-surgical practice, like "CABG", "ICD", "STENT" and "PTCA". These are the information that we will use to compute our model. In particular, we will allow patient's age to be time-dependent, as it naturally is, but will rescale it in order to prevent the log-likelihood estimated value to increase dramatically. For what concerns the other used information, as already explained in Section 2.2, they only state whether a patient experienced the considered practice or not during her/his admissions history. Moreover, we will collapse these data into a unique covariate (for simplicity of notation we will, from now on, call it *surgical*) with four possible levels: if a patient underwent no medical practice, then the covariate will report the value *None*; on the other hand if all of the considered practices have been performed, then the corresponding value is *All*; if an Implantable Cardioverter-Defibrillator ("ICD") was implanted in a patient and no other practices were performed, then we will set $surgical = ICD$; the last level corresponds to the covariate value of *Three*, when a patient experienced at least one practice out of "CABG", "STENT" and "PTCA", but had no "ICD".

In order to obtain the desired frailty model using function `frailtyPenal`, which allows for multiple models computation according to values of certain parameters, we need to specify several function's arguments. First of all, to introduce the frailty term into the Cox model, argument `Frailty` has to be equal to `TRUE`, so that the algorithm estimates the frailty term. Moreover, we want to compute a shared frailty model, as it is the best model for the kind of data we are analysing. For this reason we set the other arguments for model selection (`joint` and `recurrentAG`) equal to `FALSE`. All the other arguments of `frailtyPenal` function are used in the estimation process for the baseline hazard function, which is computed and approximated through cubic M-splines, as the estimator of $h_0(t)$ has no analytical solution [28].

The last, but equally important, argument to be set to compute the model through `frailtyPenal` function is `RandDist`, which gives information on what distribution to use for the frailty term. The function allows for two different solutions: Gamma distribution or log-Normal distribution. As already discussed, we decided to use the log-Normal distribution for the frailty parameter in our model.

On the basis of the analysis conducted on our dataset in Chapter 3, we decided to manage the analysis in this chapter according to the results previously obtained. The model and the algorithm we applied to the dataset to estimate the functional shape and value of the hazard functions are strictly connected to the value of the data under analysis. In particular, the model, hence the algorithm, is conjectured in such a way that the baseline hazard function is constructed through patients' survival times and evaluated through all the information from the considered patients. Then the patient-specific hazard functions are reassembled multiplying the baseline hazard function with the regression and frailty terms, following the theoretical model (4.2). For this reason, keeping all patients together would interfere with our aim to rebuild the less constrained patient-specific functional shape, as the model itself impose to the shape of the hazard function for each patient to be adaptable to that of the estimated baseline. Splitting patients into the previously found groups, which have a strong distinctiveness, would allow at least to differentiate the baseline hazard functions for patients that show a completely different hospitalisations history evolution. Considering this aspect of the model and its consequences, we decided to apply the proposed shared frailty model to each of the three previously found groups, not considering patients as if they would come from the same population. Notice that this way, we are allowing the frailty term distribution to vary among groups, as the distribution's variance term will be estimated for each of the considered clusters. As a result, we will have the following model over the complete dataset, where k is the cluster index:

$$h_{ij}(t|v_i;k) = h_0(t)v_i exp(\boldsymbol{\beta}'\mathbf{z}_{ij}); \tag{4.7}$$

$$v_i \overset{\text{i.i.d.}}{\sim} log - Normal(0, \sigma_k^2). \tag{4.8}$$

4.3 Analysis of the results

Function `frailtyPenal` from package `frailtypack` returns as output values several vectors and matrices to reconstruct the patient-specific hazard functions in accordance with the desired and specified model. In particular, a vector of regression terms is returned in output argument `linear.pred`: this vector is of length equal to the number of events in the considered group of patients, i.e. the number of patients plus all of their events next to the first one, when applicable. The value of each element of `linear.pred` vector is equal to the argument of the exponential in the model formula (see

Equation (4.1)): it is the sum of the regression term $\boldsymbol{\beta}'\mathbf{z}_{ij}$ and the frailty term w_i, which is a realisation from the zero-mean Normal distribution. The regression coefficients are fixed values throughout the dataset (for each event and for each patient) and the covariates matrix $\mathbf{z}_i$ corresponds to the record of patient i, including information on age of the patient at the event date and practices performed on her/him throughout the observation time (see previous section). For this reason, $\mathbf{z}_i$ is a matrix of dimensions $j \times 2$, where j corresponds to the number of hospitalisations of patient i. The same value in `linear.pred` can be obtained through arguments `coef` and `frailty.pred`, which contain estimation of regression coefficients and of patients' frailty term, respectively.

Another important value returned by function `frailtyPenal` is a matrix holding the baseline function estimate and its confidence bands. Each of them is evaluated over 100 time points selected as an equally spaced sequence of times starting from $t = 0$ up to the maximum observed survival time in the considered dataset. For this reason, the function returns also a vector with corresponding times where the baseline hazard function was estimated, argument `x1`.

With all these information we are able to reconstruct the hazard function for each patient. The resulting hazard is computed as follows:

$$\hat{h}_i(t_l|v_i;k) = \hat{h}_0(t_l)v_i exp(\hat{\boldsymbol{\beta}}'\mathbf{z}_{il}); \tag{4.9}$$

$$\mathbf{z}_{il} = \mathbf{z}_{ij} \quad for \quad t_{j-th\,event} \leq t_l < t_{(j+1)-th\,event}; \tag{4.10}$$

$$v_i \overset{\text{i.i.d.}}{\sim} log - Normal(0, \hat{\sigma}_k^2). \tag{4.11}$$

First of all, in (4.9), we show the estimated hazard function for patient i, $\hat{h}_i(t_l|v_i;k)$, computed as the product of the estimated baseline hazard function $\hat{h}_0(t_l)$ with the patient-specific frailty term v_i and the exponential of the regression term $\hat{\boldsymbol{\beta}}'\mathbf{z}_{il}$. The hazard function $\hat{h}_i(t_l|v_i;k)$ is computed over the time points as in the returned value `x1`. These, as already said, are 100 points over the observation time period, and here we label them as t_l, where $l = 1, \ldots, 100$. Naturally we have $t_1 = 0$ and $t_{100} = max(t_{observed})$. Then the covariate vector $\mathbf{z}_{il}$ corresponds to the vector of covariates for the i-th patient at time t_l, in conformity with (4.10): at each evaluation time point t_l, the covariate value is equal to its value in the corresponding survival time. While t_l is consecutive to the time of j-th event and previous to the next event time, $j = 1, \ldots, H_i$ with H_i being the maximum number of admissions for patient i, we assume $\mathbf{z}_{il} = \mathbf{z}_{ij}$. Notice that the time of the $(H_i + 1)$-th event for patient i is equal to the date of death or the end of the study.

Finally, the frailty term for patient i is a realisation from the estimated distribution log-Normal, which, in our case, is a one parameter distribution. Parameter $\hat{\sigma}_k^2$ in Equation (4.11), which corresponds to the variance of the zero-mean Gaussian distribution for term w_i in Equation (4.1), is computed to obtain the realisations for patients in the k-th cluster, where $k = 1, 2, 3$.

Once we computed the patient-specific hazard functions, it is also of a great interest to compute the corresponding cumulative hazard functions, as they could be of some help in the comprehension process of the clusters characteristics. The theoretical model for the cumulative hazard function is:

$$\Lambda(t) = \int_0^t h(u)\, du. \tag{4.12}$$

From this theoretical model we are able to compute both the cumulative baseline hazard function and the patient-specific cumulative hazard functions. To compute and estimate these functions on the basis of the obtained baseline hazard function $\hat{h}_0(t_l)$ and patient-specific hazard function $\hat{h}_i(t_l|v_i; k)$, we decided to use a simple trapezoidal rule applied over each pair of time points from vector `x1`. We did not use a more refined numerical technique to estimate the integral in the cumulative hazard definition, because we do not need a precise estimate of the function, but only the qualitative approximation for the integral of the considered hazard function. Consequently, the cumulative baseline hazard function and the cumulative hazard function for patient i are of the following form:

$$\hat{\Lambda}_0(t_l) = \int_0^{t_l} \hat{h}_0(u)\, du \approx \sum_{m=1}^{l-1} (t_{m+1} - t_m) \left[\frac{\hat{h}_0(t_{m+1}) + \hat{h}_0(t_m)}{2} \right]; \tag{4.13}$$

$$\hat{\Lambda}_i(t_l|v_i) = \int_0^{t_l} \hat{h}_i(u|v_i)\, du \approx \sum_{m=1}^{l-1} (t_{m+1} - t_m) \left[\frac{\hat{h}_i(t_{m+1}|v_i) + \hat{h}_i(t_m|v_i)}{2} \right]. \tag{4.14}$$

In (4.13) and (4.14) are shown the estimates for the cumulative baseline hazard function and patient i cumulative hazard function, respectively. In particular, it is important to notice that to compute the value of the cumulative function at any time t_l, with $l = 1, \ldots, 100$, we summed the results obtained with the trapezoidal formula over each time interval $[t_m; t_{m+1}]$, with $m = 1, \ldots, (l-1)$ and $l = 1, \ldots, 100$.

Remembering that, as detailed in Section 3.3, cluster number 1 is "sick" patients cluster, number 2 is "terminally ill" patients group and that the

third cluster is the one of "healthy" patients, it is of a great interest to analyse the resulting baseline hazard function, together with its cumulative, from each of the mentioned groups.

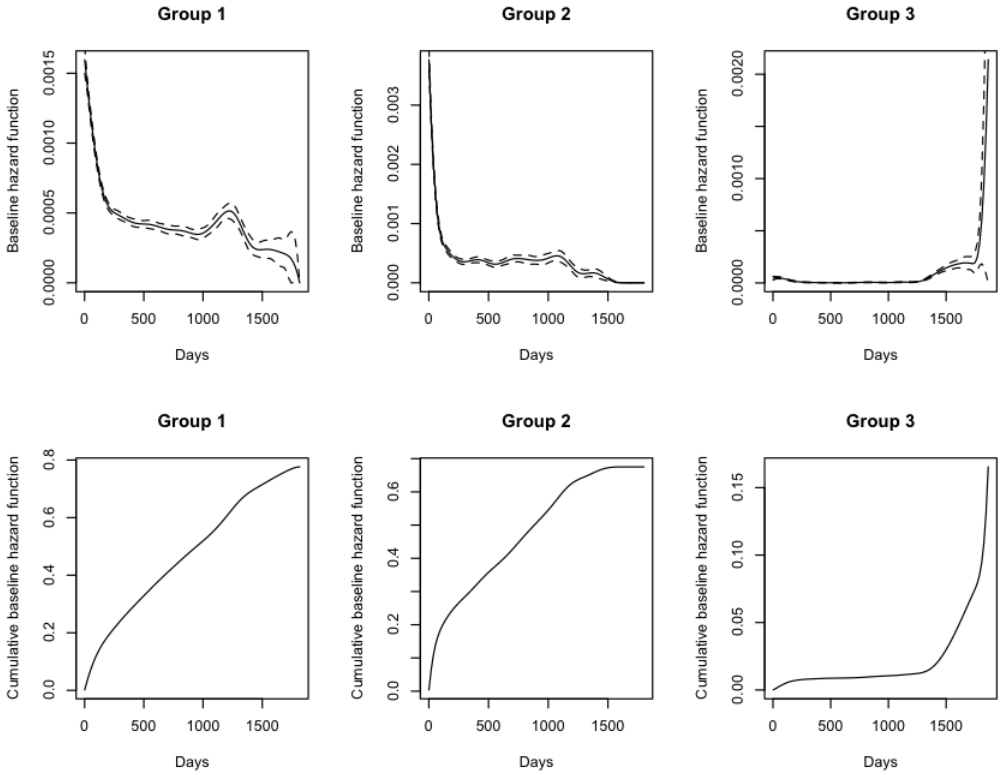


Figure 4.1: Baseline hazard function and the corresponding cumulative baseline hazard function for each of the considered clusters.

In Figure 4.1, we can see in the first row of plots the baseline hazard functions from each cluster and, in the same plots, 95% confidence bands for the estimated baseline hazard functions. First of all, it is important to notice that for each group, the confidence bands give us a good feedback on the correctness of the functional estimation obtained from the applied algorithm. The only group which is characterised by some noise is the third one, especially towards the end of the study time period: this is because, approaching the end of the observed time, the majority of patients in this cluster have censored survival times and for this reason the probability of a new hospitalisation is estimated to be rising rapidly, but with great uncertainty. Secondly, we would like to spend some words on the functional shape of the baseline hazard function of each cluster, in an effort to confront them on the basis of the conclusions reached in Section 3.3. Let's firstly take a look at plots for clusters 2 and 3: these are the two opposite clusters, in

terms of patients that constitute them, representing the most extreme patients' health situations. The baseline hazard function for cluster number 2 is characterised by high risk probability of being re-hospitalised at the very beginning of the study time period, going below the 0.0005 margin within the first six months and reaching the zero limit line after 4.65 years (see Figure 4.2).

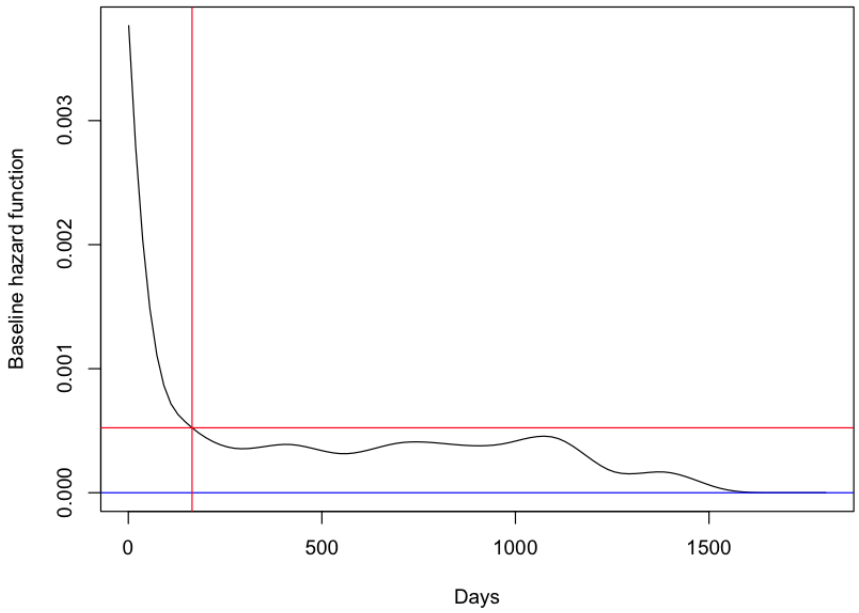


Figure 4.2: Baseline hazard function of second cluster. In blue it is highlighted the zero limit line, reached after 4 years of observation. In red is shown the function's point corresponding to the intersection of 0.0005 probability line and the 164 days bar.

This same result can be seen through the cumulative baseline hazard function of second group, in the second line of plots from Figure 4.1. The cumulative function grows rapidly for the first six months, when the functional gradient diminishes correspondingly to the baseline hazard function's trend. Towards the end of the study, the cumulative baseline hazard function becomes constant, as in the matching baseline hazard function it is reached the zero limit line. Recalling that cluster number 2 is the "terminally ill" patients cluster, with the highest mortality rate, we then find perfect matching with the functional shape of the baseline hazard function and its cumulative. At the end of the study time period the majority of patients in this cluster are dead. This is the reason leading to the natural estimation of the probability of re-hospitalisation equal to zero.

On the other hand, if we analyse the baseline hazard function estimated for cluster number 3, we can immediately appreciate a specular trend compared to that of the second cluster. We need first of all to remember that the third cluster is composed of "healthy" patients, with the lowest mortality rate and the highest number of censored survival times.

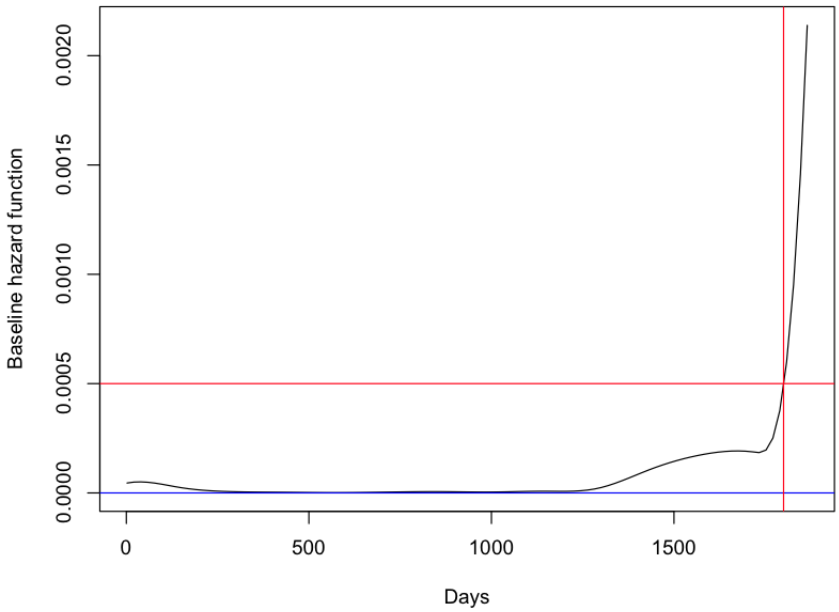


Figure 4.3: Baseline hazard function of third cluster. In blue it is highlighted the zero limit line, from which the function drifts away after 3.5 years of observation. In red is shown the function's point corresponding to the intersection of 0.0005 probability line and the 5 years bar.

In Figure 4.3, we can particularly recognise the rising trend after five years of observation (highlighted through the intersection of the two red lines). If these patients are the ones surviving throughout the study time period, their probability of being re-hospitalised, or being censored, grows rapidly when we reach the observation right limit. In particular it is interesting to notice that this baseline hazard function values are within a smaller interval compared to that of the second cluster ($[1.24 \times 10^{-6}; 2.14 \times 10^{-3}]$ and $[2.26 \times 10^{-18}; 3.77 \times 10^{-3}]$, respectively). For the first five years of observation, the estimated baseline hazard function takes particularly low values, once again because these patients have a zero probability of being re-hospitalised early after their first admission. The same trend can be appreciated through the cumulative baseline hazard function, see plot in the second row from Figure 4.1. The cumulative function has a shape that resembles that of the baseline, as for the first four years the baseline assumes

values which do not modify significantly the computed integral. Like the baseline hazard function, towards the end of the study time period, the cumulative baseline hazard function's gradient increases remarkably.

Finally we should observe the first cluster's baseline hazard function's trend. At a first glance it resembles the most to the baseline hazard function of the second cluster. It reaches the highest values at the very beginning of the study time period, and diminishes down to the zero limit slowly during all the observation window. It has although a lower initial probability of being re-hospitalised compared to that of the second cluster. Moreover, this cluster's baseline hazard function is characterised by a particular decreasing trend: this result matches what we expected to obtain from the baseline hazard function of patients labelled, in Section 3.3.1, as "sick" patients. They have at the beginning a high probability of being re-hospitalised, which decreases rapidly to an almost constant value of 0.0004. This is until after 2.78 years, when the probability of being re-hospitalised increases once more to decrease again rapidly after few months. The reason behind this behaviour is that the majority of these patients dies within the first 1000 days, but there is a certain number of patients who dies in the last interval of time (see Figure 4.4).

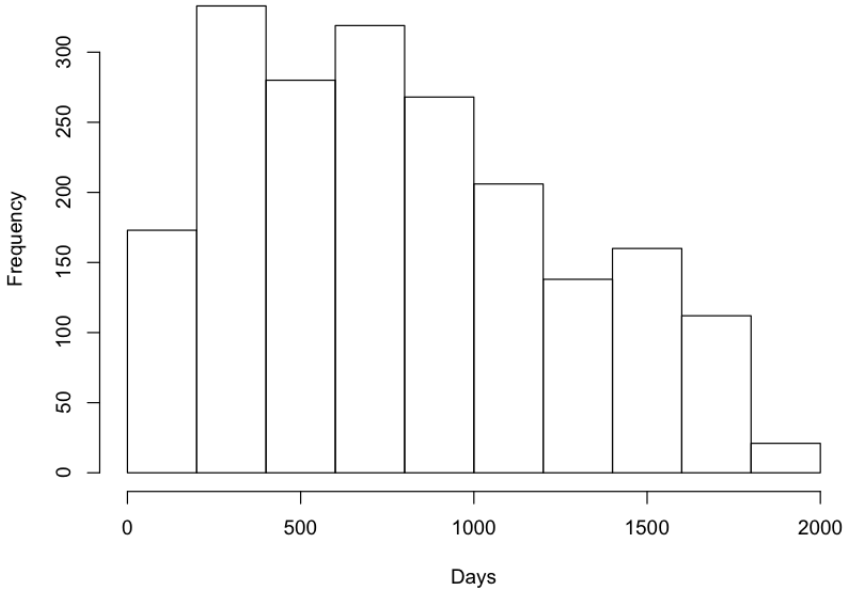


Figure 4.4: Histogram of the time of death for patients in the first cluster.

We find proof to our hypothesis in Figure 4.5, where we attempted to cluster dying patients of this group according to their overall survival time. In this

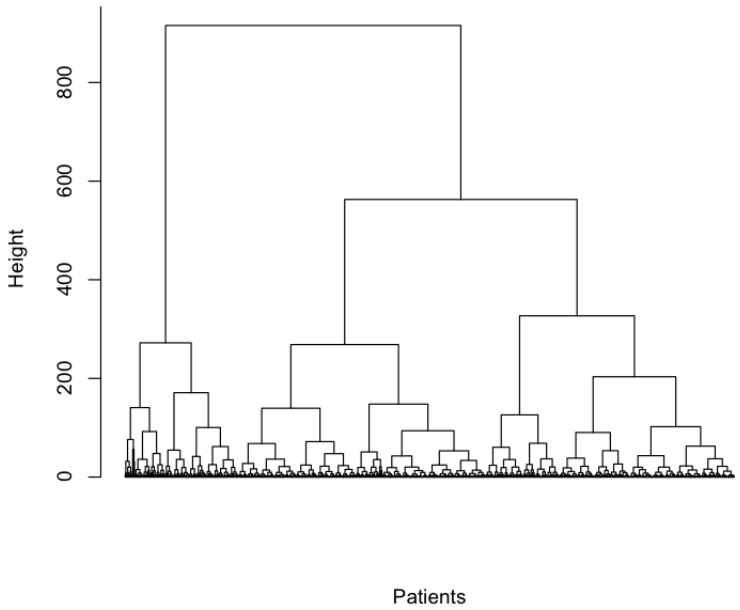


Figure 4.5: Average linkage dendrogram for distances between dead patients' survival time in first cluster.

figure we show the obtained clusters for these patients using an average linkage hierarchical clustering method [13]. From the resulting dendrogram, we clearly see that we can split patients into two or three clusters. In order to explain the baseline hazard function's trend, it is sufficient to state that there are two inner groups of patients. Patients are grouped according to their overall survival time (time elapsed from their first admission to their death) and we find that they are assigned to one or the other cluster conditionally on the fact that their overall survival time, $T_i = \sum_{h=1}^{H_i} T_{ih}$, is greater or smaller than 1278. As a result, we obtained that one of the two groups has a mean survival time equal to 1539.827. This outcome explains the functional shape of the baseline hazard function for "sick" patients cluster (see Figure 4.6). Notice that we are clustering only dying patients. This means that, in previous chapter analyses, we could not have divided the complete dataset into four groups because there are not enough differences among patients of the "sick" cluster to induce us to consider it as two distinctive populations.

Once we analysed the baseline hazard functions for the three groups, we are ready to compute the patient-specific hazard functions and their corresponding cumulatives. In Figure 4.8, we show the results from the estimation of the hazard functions and their cumulatives for patients from group 1. As already said, the functional shape is strictly similar to that of the corresponding baseline hazard function in Figure 4.6. In spite of the

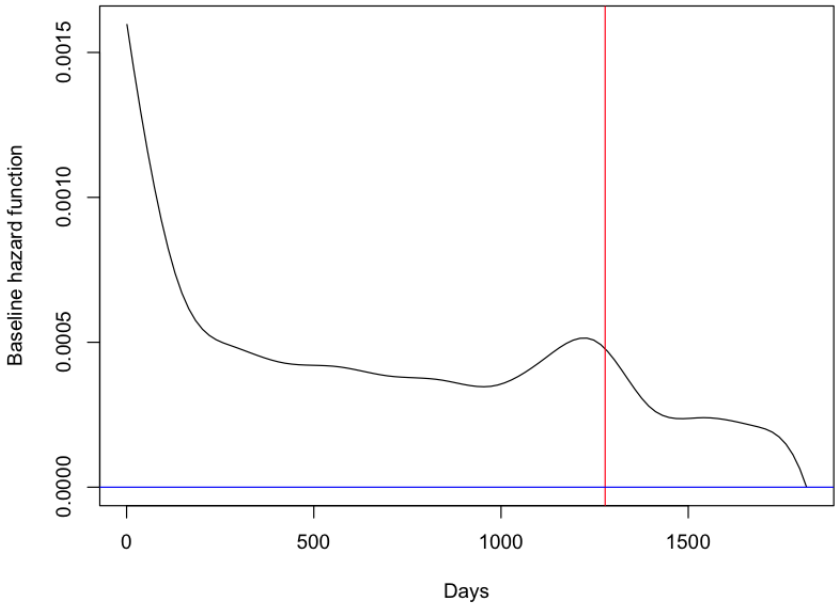


Figure 4.6: Baseline hazard function of first cluster. In blue it is highlighted the zero limit line, reached only at the end of the study time period. In red is shown the time point equal to 1278, the minimum observed survival time for the second group of dead patients within this considered cluster of "sick" patients.

restriction imposed by the model to the functional shape, there is a great variance that can be appreciate among patients' hazard functions, especially through the cumulative hazard functions, a variance that will be further more investigated in next section.

The same results are found also for the obtained hazard functions for patients from the second and third clusters (see Figure 4.9 and Figure 4.10, respectively). Notice that for the third cluster, "healthy" patients cluster, we can perceive not only a great variance among the resulting function for each patient, but also that in the cumulative hazard functions it is possible to identify groups of patients that, towards the end of the study time period, have different functional behaviours. This result is perfectly matching what we already stated when analysing the confidence bands for the baseline hazard function from cluster number three. The uncertainty found in analysing the baseline hazard function is now embodied in the functional variance among patients. Moreover, we will show in Chapter 5 that it is possible to identify different groups of "healthy" patients according to their hazard functions (see Section 5.2.3).

On the other hand, patients from the second cluster, "terminally ill" patients, are characterised by a smaller functional variance, which can be

seen especially through the hazard functions in Figure 4.9a. Patients from this cluster share a particular characteristic: the great majority of them dies within the first year of observation (see Figure 4.7), leading to an obvious reduction in the possible variance expressed in the functional computation.

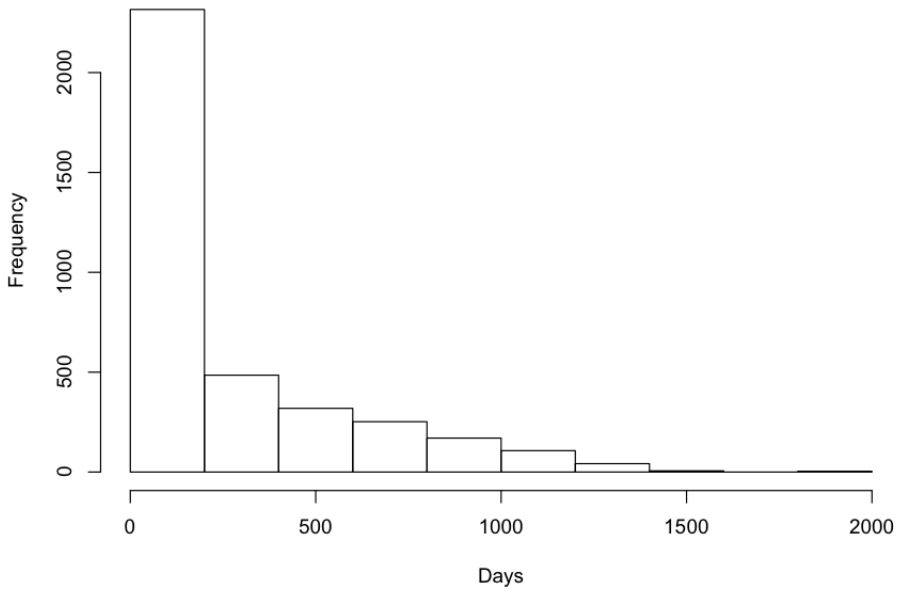
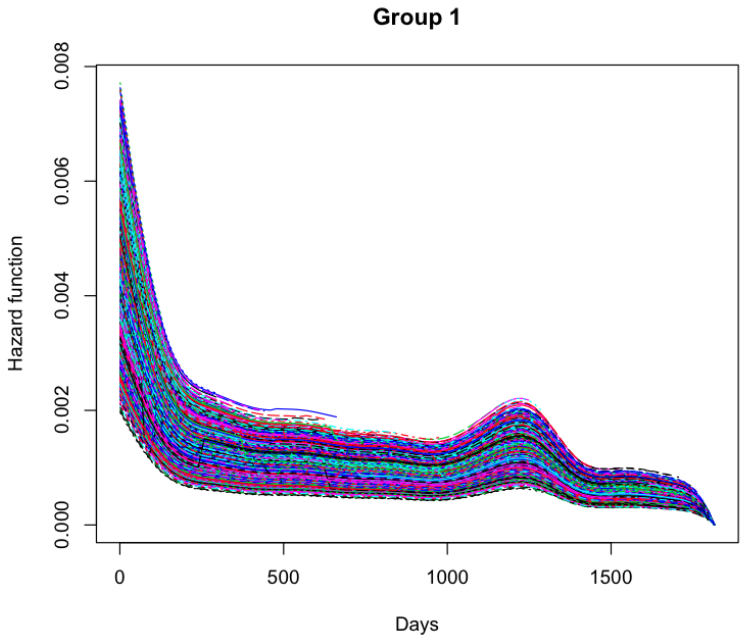
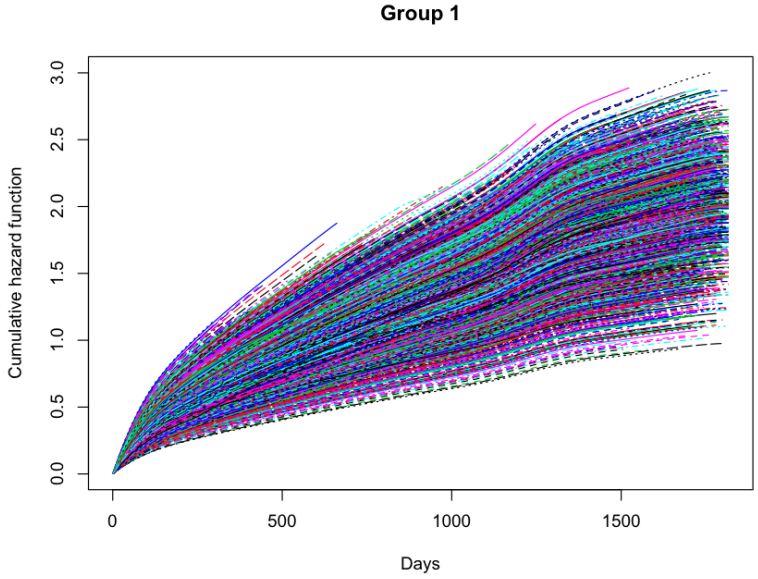


Figure 4.7: Histogram of the time of death for patients in the second cluster.

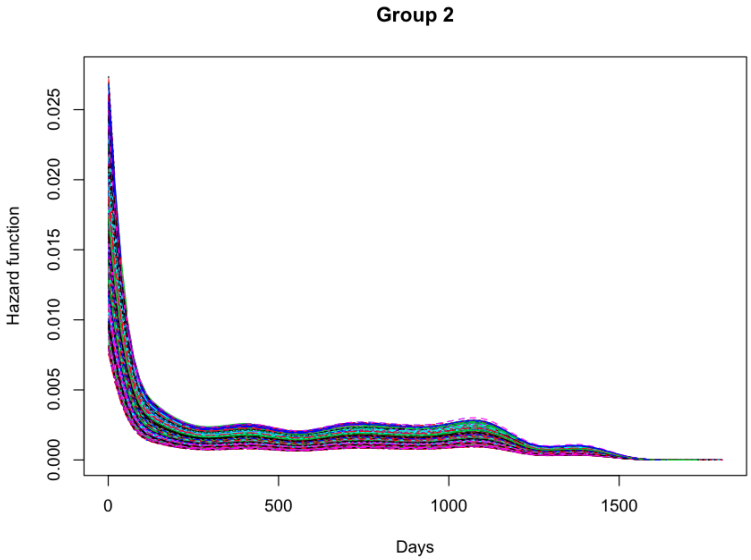


(a) *Patient-specific hazard function.*

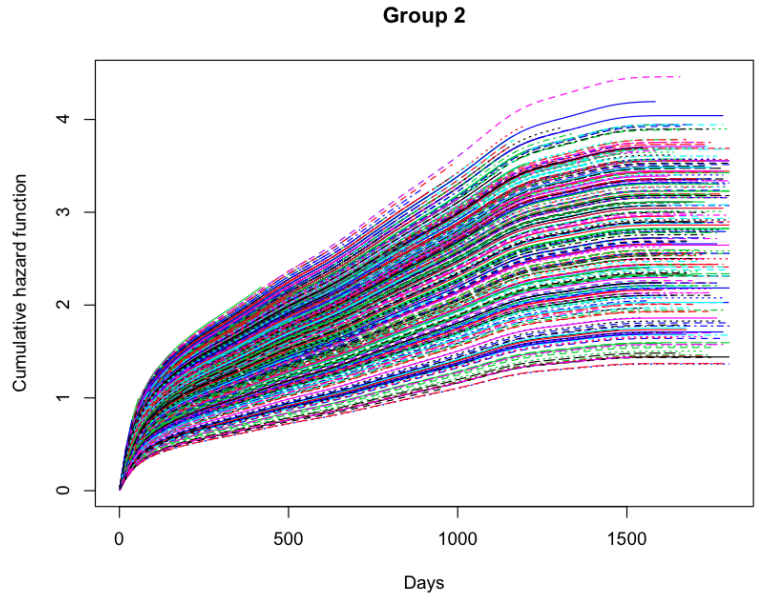


(b) *Patient-specific cumulative hazard function.*

Figure 4.8: Hazard function and cumulative hazard function for patients in cluster number 1. These functions were obtained according to the described procedure, see Equations (4.9) and (4.14).

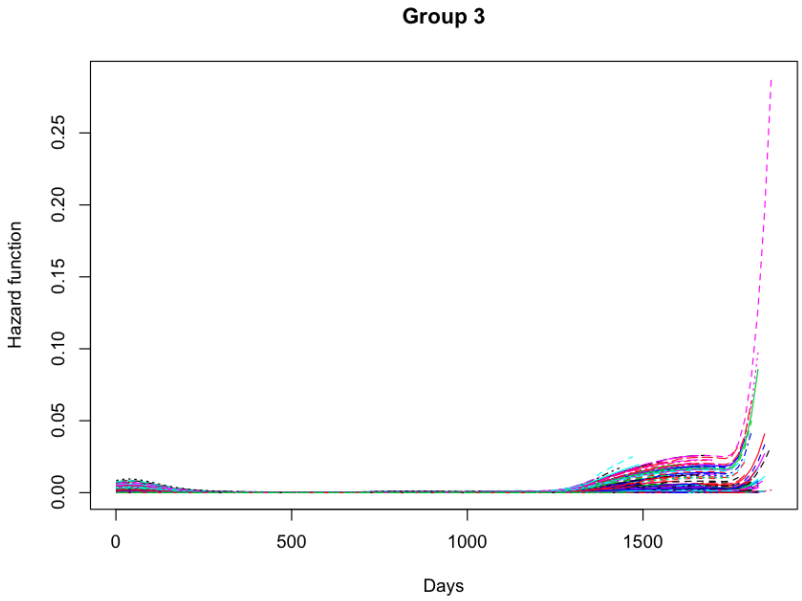


(a) *Patient-specific hazard function.*

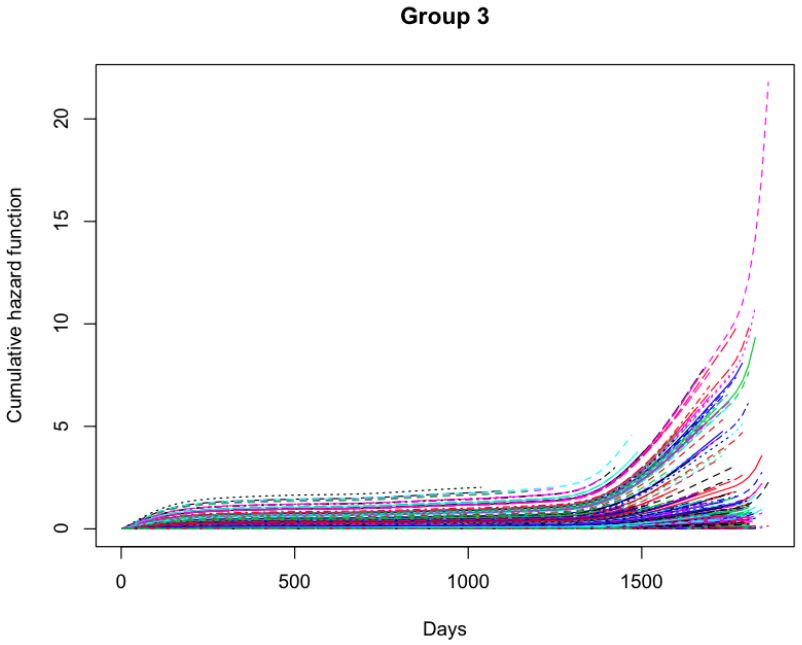


(b) *Patient-specific cumulative hazard function.*

Figure 4.9: Hazard function and cumulative hazard function for patients in cluster number 2. These functions were obtained according to the described procedure, see Equations (4.9) and (4.14).



(a) *Patient-specific hazard function.*



(b) *Patient-specific cumulative hazard function.*

Figure 4.10: Hazard function and cumulative hazard function for patients in cluster number 3. These functions were obtained according to the described procedure, see Equations (4.9) and (4.14).

4.4 Analysis of functional variance within and between groups

We already showed functional shape diversity among groups due to the differences in patients composing the three clusters and leading to the contrasting behaviours of the correspondent baseline hazard functions. What is now interesting to analyse is the variance between groups due to the frailty term. One of the values returned by function `frailtyPenal`, that we used to compute the estimated patient-specific hazard functions for each cluster, is the estimated variance of the frailty term. It is relevant for our analysis to look at this output: variance of the frailty term in the three clusters is equal to 0.08466 for the first group and equal to 0.08482 and 3.9758 for the second and third groups, respectively. What is remarkable, is that the estimated variances for clusters 1 and 2, corresponding to "sick" and "terminally ill" patients, are values similar between them; furthermore they are approaching the zero limit, which is, as we showed in (4.6), the limit of non-heterogeneity among patients in the considered population. What we can appreciate from this result is that, as we already stated, patients from cluster 1 are characterised by a functional trend which resembles that of patients in the second group, not only for the shape of the baseline hazard function, but also for the variability within the patients due to the frailty term. Moreover, there is a very low heterogeneity among patients from clusters 1 and 2, and the variance that we appreciate in the hazard functions from Figure 4.8 and Figure 4.9 is mostly due to the combination of the regression and frailty terms in (4.1). In particular, the variability of patients' age covariate in these two clusters is the one mostly responsible for the results in Figure 4.8 and Figure 4.9.

On the contrary, the estimated variance for the frailty term of cluster number 3 shows that there is a considerably higher heterogeneity within patients of this group. This peculiarity in the third cluster was already shown in the baseline hazard function uncertainty towards the end of the study time period. This characteristic led us to develop an even more detailed analysis on this group of patients, in an effort to find clusters of patients within the considered group (see Section 5.2.3).

Chapter 5

Analysis of patients surviving after each hospitalisation: clustering and functional hazard reconstruction

We showed in previous chapters the analyses over the complete dataset, concentrating the study of the distribution of survival times between two events and the associated event risk, discussing over the theory behind the models we applied, showing and analysing obtained results. Throughout all the analyses on the dataset, we found that we could split it into three groups, each of them identifiable with a certain type of patient: two groups stand for the extreme cases ("terminally ill" and "healthy" patients) and an *in between* group, which shares statistical characteristics with both of the other two. From an analytical point of view (see Sections 3.3 and 3.4), "sick" patients cluster is closer to the "healthy" patients one, as it shares with it a lower mortality rate and a similar trend for the Kaplan-Meier estimator. On the other hand (see Section 4.3), from a functional point of view, i.e. looking at the reconstructed patient-specific hazard functions from each group, characteristics of the "sick" patients cluster are more comparable to the shape and behaviour of the "terminally ill" patients cluster ones. For this reason, we decided to develop a more refined analysis to reinforce our previous assumptions.

Dying patients have the greatest influence over the considered models, as it is evident through the dissimilarities between the most extreme clusters we obtained. At each new hospitalisation, the dataset dimension is reduced by an amount equal to the number of dead patients at the previous admission.

This means that the information in the early stages of the observation time period are richer than those at the end of it, leading to a poorer model estimate compared to that for the first admissions. Moreover, dying patients have a great influence on the computed estimates for the hazard models. To escape these natural limitations, we decided to replicate the same analyses we showed in Chapters 3 and 4 over particularly conjectured subgroups of patients from the complete dataset.

5.1 Surviving after the first admission

First of all, we considered the dataset cleared from the early dying patients. These patients, who died during their first admission, could be patients dying for different causes other than HF or chronic HF. In particular, patients affected by chronic heart failures are likely to have several re-hospitalisations, so we removed from the dataset all patients dying during their first hospitalisation or before the second one. In so doing, the remaining subjects survive through their first admission. Still, there are patients who have only one admission, or just two, or three and so on, up to a maximum of five, but those having only one hospitalisation happen to be alive at the end of the study time period.

Once we selected the desired subgroup of patients from the complete dataset, we applied the already shown models (see Chapter 3 and Chapter 4). We herein report and comment the results, while for the models theory the reader is referred to previous chapters.

The number of patients that have been removed is equal to 2640 (19.15% of total dataset), corresponding to the number of dead patients in the complete dataset who had just one hospitalisation. Trying to split the subgroup into three clusters, as we did for the complete dataset, we obtained the clusters whose profiles are shown in Figure 5.1.

Comparing this figure to the equivalent one for the complete dataset (see Figure 3.3) and making our first hypotheses on the behaviour of the cluster-profiles, especially in the first two survival time variables, it appears clear that we can identify once again the same groups of patients: “sick”, “terminally ill” and “healthy”, in the order. We will support this hypothesis through a deep analysis of the composition of the three obtained clusters. Before that, we would like to underline that, while the third cluster (“healthy” patients) has an almost equal trend to that of the corresponding cluster from the complete dataset, this characteristic can not be highlighted for the other two. In particular, once we removed patients dying before

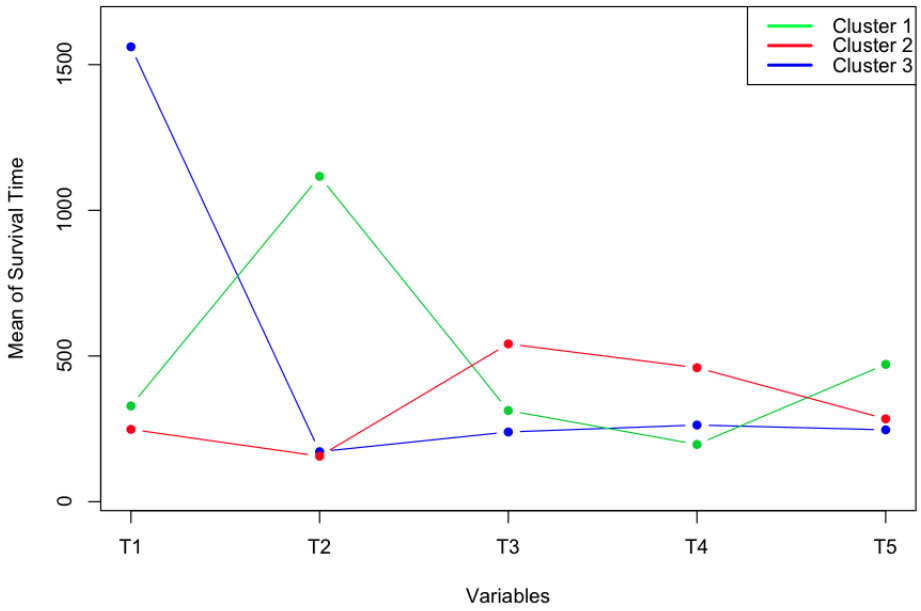


Figure 5.1: Cluster-Profiles: Mean survival time for each variable T_h, $h = 1, \ldots, 5$, and for each cluster k, $K = 1, \ldots, 3$.

their second hospitalisation, the algorithm seems to associate patients with a higher surviving time to the "sick" patients cluster, as expected, leaving the "terminally ill" group with those patients characterised by lower survival times.

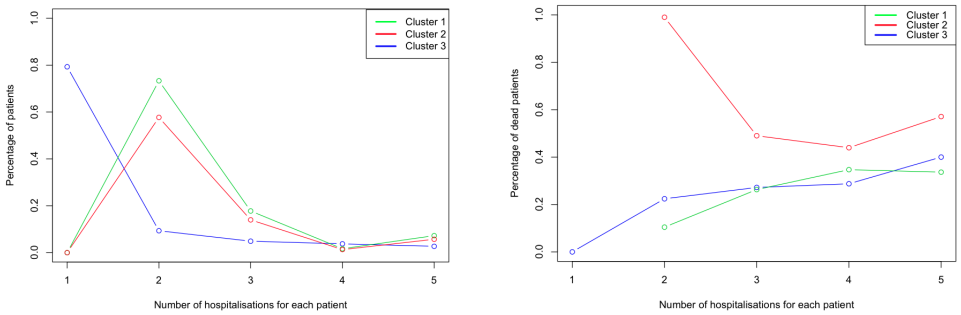
To understand the composition of the three clusters and verify our assumptions, we conducted the same analyses as in Sections 3.3.1, 3.3.2 and 3.3.3.

Properties	**Cluster 1**	**Cluster 2**	**Cluster 3**
Cluster Size	2953 (26.50%)	3752 (28.40%)	4440 (39.83%)
Mortality Rate	15.34%	65.81%	5.59%

Table 5.1: Characteristics of clusters: size and mortality rate

From table 5.1, we see that the mortality rate is the first index to ascribe our characterisation hypothesis of the obtained clusters: the second cluster is that with the higher mortality rate, while the third group registers the lowest one. Moreover, as we can see from Figure 5.2a, the algorithm associates all patients having just one hospitalisation with the third cluster, a behaviour that we naturally would relate to that of healthy patients. On the other hand, patients in the second cluster have the higher mortality

rate throughout the study time period, and we can appreciate this if we take a look at Figure 5.2b. In particular, it is interesting to notice that in correspondence of the second hospitalisation time variable, for the "terminally ill" patients cluster the percentage of patients dying is approximately equal to 1, a fact that reinforces our believes that this is truly the "terminally ill" patients cluster and that, removing patients who died before their second hospitalisation, helped us to achieve a clearer identification of the three clusters and their features. Finally, we find once more the *in between* characteristic proper of the first cluster, which earned it the name of "sick" patients cluster: if we look at Figure 5.2a, the first cluster (green) has almost the same distribution of patients as the second one (red), with a peak at the second hospitalisation time variable T_2. On the contrary, if we move our attention on Figure 5.2b, we see that the first cluster (green), has a mortality percentage trend that resembles that of the third cluster (blue).



(a) *Percentage of patients in each cluster who have only one hospitalisation over the study time period, or just two admissions, or three and so on.*

(b) *Percentage of patients in each cluster who died during their first hospitalisation over the study time period, or during their second one, or third and so on.*

Figure 5.2

At this stage of the analysis, we are ready to proceed with the functional part of it. As we did in Chapter 4, we computed function `frailtyPenal` separately over the three clusters for the same reasons already discussed in Section 4.2. In Figure 5.3 we show the results for each of the considered clusters: plots in the first row show the baseline hazard functions and their 95% confidence bands, while plots in the second row display the corresponding cumulative baseline hazard functions.

First of all, it is evident that the obtained baseline hazard functions are consistent with the results from the previous analyses on the three clusters and that they are comparable to those in Figure 4.1. Once again the con-

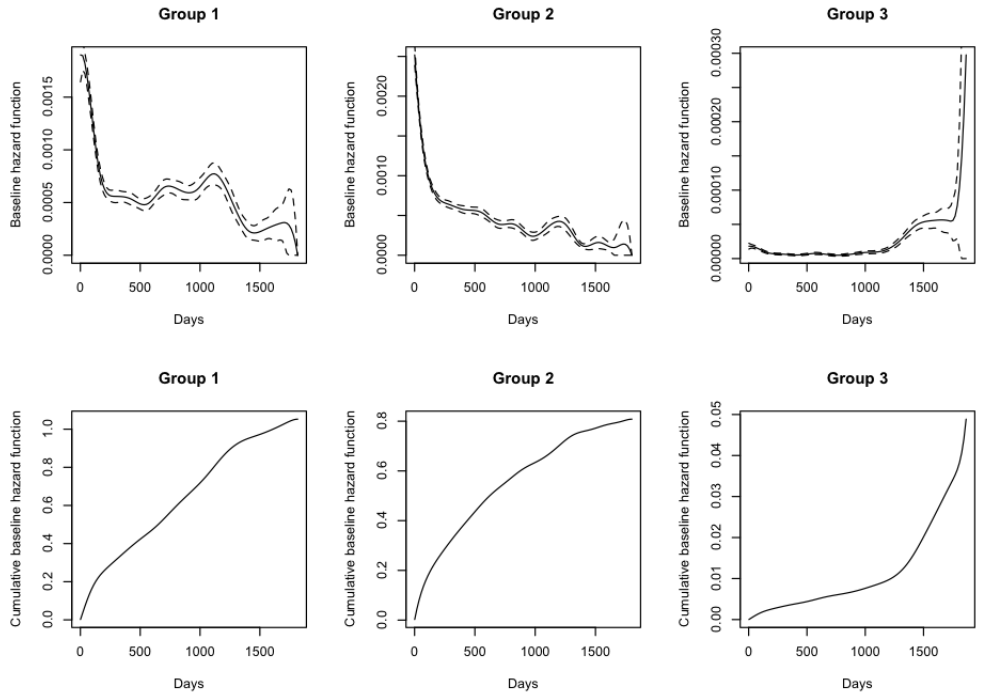


Figure 5.3: Baseline hazard function and the corresponding cumulative baseline hazard function for each of the considered clusters.

fidence bands give us a good feedback on the correctness of the estimated $\lambda_0(t|k)$, with $k = 1, 2, 3$, exception made for the third group ("healthy" patients), for which we find the same result as in the corresponding group obtained from the complete dataset: towards the end of the study time period the confidence bands show an increase in the estimation uncertainty. As for the complete dataset, we find that the baseline hazard function estimation for the first group ("sick" patients) resembles much more that of the second cluster ("terminally ill" patients), being characterised by a higher initial probability of being re-hospitalised. The first cluster differentiates itself from the second one for the numerous peaks along the study time period, which give this function an oscillatory profile. Moreover, this characteristic reinforces the *in between* behaviour proper to this cluster. In Figure 5.4 we show the results obtained when we rebuild the patient-specific hazard functions and corresponding patient-specific cumulative hazard functions, based on the obtained baseline hazard functions. Of course, the mean trend shown in these plots its that already discussed for the baseline hazard functions.

From all these analyses, we can now state that removing patients who died before their second hospitalisation did not change our ability to identify those groups of patients found when clustering the complete dataset. It

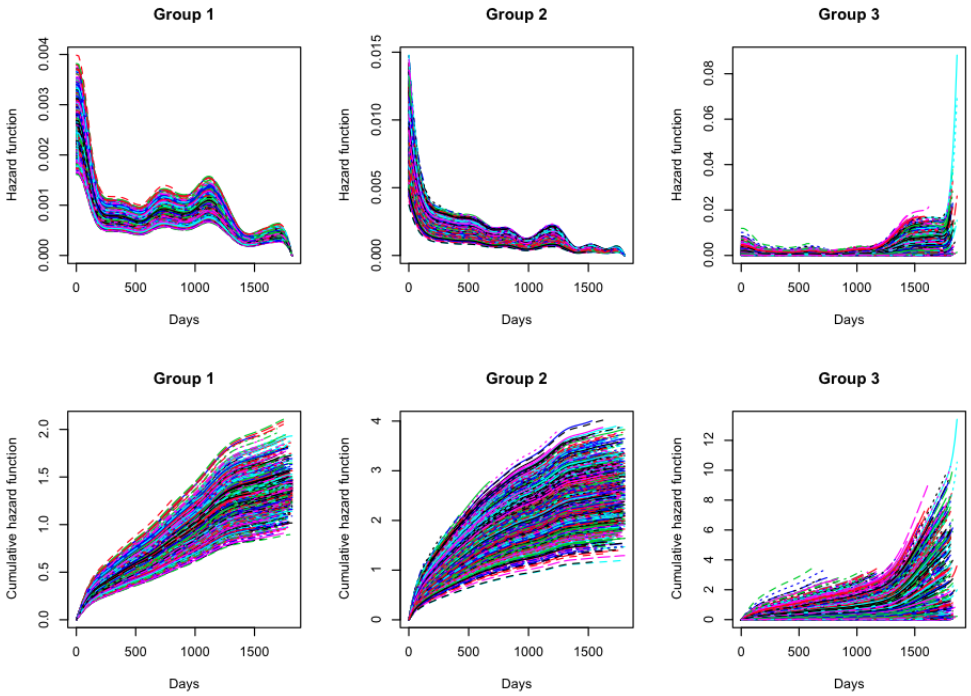


Figure 5.4: Patient-specific hazard functions and corresponding patient-specific cumulative hazard functions for each of the considered clusters.

seems, as a result of this strong characterisation of patients selected for the subgroup, that this choice led to an even more satisfactory division of patients into well distinctive clusters.

This said, on the basis of the remarkable results we obtained, two more ideas have come to our attention: what kind of results could we obtain if removing step by step all patients dying before their h-th hospitalisation? Consequently, what is the movement of patients between obtained clusters? The answers to these questions are stated in the following sections.

5.2 Surviving after the h-th hospitalisation

Once we saw that removing patients dying before their second hospitalisation is improving the algorithm ability to separate different kind of patients, we thought of conducting the same analyses over other properly selected subgroups of patients. In particular, these subgroups are each time selected as a subset of the previous one, as the subgroup of patients analysed in the above section was a subgroup of the complete dataset. We will then consider, as a second subgroup, a subset of patients from the subgroup

of the complete dataset, and so on. We will call each subgroup G_h, for $h = 1, \ldots, 5$. Subgroup G_1 is the first subset of patients selected from the complete dataset, whose properties were already discussed in previous section. Subgroup G_2 is a subset of G_1, obtained removing from it all those patients dying before their third hospitalisation. All the other subgroups are obtained from the previous one, removing each time patients dying before the next hospitalisation. Subgroup G_5 is the last one, which corresponds to the group of patients who remain alive throughout the study time period and, in particular, is a subset of all the previous subgroups. We can summarise properties of the collection of subgroups under analysis as follows:

- $G_{h+1} \subset G_h$, $\forall h = 1, \ldots, 4$;
- $G_h \subset G_{tot}$, where G_{tot} is the complete dataset;
- $\bigcup_{h=1}^{5} G_h = G_1$;
- $\bigcap_{h=1}^{5} G_h = G_5$.

We are now ready to conduct the same analyses as in Chapters 3 and 4 over the subgroups G_h, with $h = 2, \ldots, 5$.

First of all, as we did for the complete dataset and for the first subgroup G_1, we want to take a look at Figure 5.5, showing the cluster-profiles obtained for each subgroup. The similarity of cluster-profiles obtained among all subgroups is remarkable, despite some, yet very little, value variations for variables T_3, T_4 and T_5. Also, notice that if we compare the resulting cluster-profiles for these subgroups to the one obtained for subgroup G_1 (see Figure 5.1), we can appreciate a strong resemblance between all plots. Clusters in green and red, which will later be named "sick" and "terminally ill", happen to have the exact same shape in each of the subgroups, meaning that we are, for each new subgroup, improving the ability of the algorithm to find distinctive clusters, with well defined properties. For what concerns the cluster coloured in blue, this is the only one which is equal among plots in Figure 5.5, but is different from the same cluster obtained from the first subgroup: for variables T_1 and T_2 it appears to have the same shape in every subgroup, including G_1, but in the last four subgroups, i.e. once we removed patients dying before their third hospitalisation, the algorithm selects, in order to build up this cluster, patients who have a maximum of 4 hospitalisations. It is a strong characterisation of this cluster, leading to the natural belief that this is truly the "healthy" patients cluster: this cluster is made of patients with the fewer number of hospitalisations, meaning that

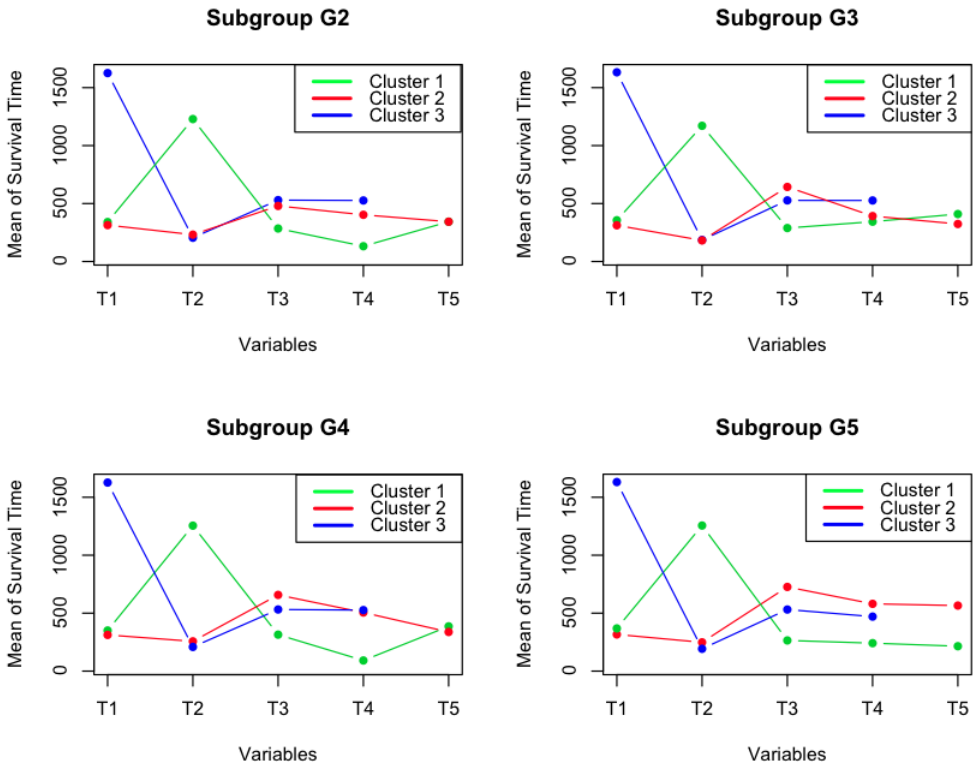


Figure 5.5: Cluster-profiles for subgroups G_2, G_3, G_4 and G_5. It is important to notice the great similarity for the obtained clusters from each subgroup.

they are patients with a lower probability of being re-hospitalised. We will later show that there are other indexes that will prove and reinforce our hypotheses, leading us to the conviction that removing a properly selected group of patients permitted us to identify clusters of patients with strong inner characterisations. Before we deepen into the analysis of clusters obtained for these subgroups, we think it is important to underline a detail that can be appreciated at this early stage of the analysis. We already discussed, in previous section, that removing patients that died before their second hospitalisation changed the inner configuration of the three clusters, permitting to obtain clusters with a stronger characterisation. Now, we can state that, removing patients that died before their third hospitalisation, reinforced even more the features of each cluster, leading to the final characterisation of them, which has no other considerable changes when we remove patients dying before their fourth or fifth admission or before the end of the study time period. We will see that this same conclusion is reached once we analyse the movement of patients among clusters (see Section 5.3).

5.2.1 Descriptive analysis of clusters in the subgroups

Through Figure 5.5, we already showed and underlined the similarities among corresponding clusters obtained in each subgroup. We now show that this resemblance can also be appreciated thanks to descriptive indexes from every cluster and that these will help us define the type of patients assigned to each cluster by the algorithm, over the subgroups.

We already know from previous analyses (see Chapter 3 and Section 5.1) that if we ask to the `mixPHM` algorithm to split our dataset into three clusters, it divides patients according to the requirements. We then choose a label for each patient, in order to identify and summarise the general properties of subjects in a specific cluster. We named them "healthy", if assigned to the cluster of patients with the lowest number of hospitalisations and mortality rate, "terminally ill", if patients are assigned to the cluster having the highest mortality rate. The last group is labelled as "sick", and it is a group showing characteristics of both the other clusters. In Table 5.2 we show the most important characteristics of each of these clusters, and in particular we will use these data to study the evolution of clusters every time we remove patients that died before their h-th admission. The presented indexes in the table are: the cluster size, to show the dimension of each cluster compared to the total number of patients in each subgroup, the mortality rate, which is the most evident clue that enables us to give names to clusters, and the mean and standard deviation of patients' age, to show the evolution of its distribution once we removed dying patients.

First of all, we would like to highlight the trend of the mortality rate index for each cluster through its evolution from one subgroup to the following. As we already stated, this index helps us identify the kind of group we are looking at: the mortality rate of patients in the "healthy" cluster is not only the lowest, but it also reaches the zero limit once we remove patients dying before their third admission. On the other hand, patients in the "terminally ill" cluster have the highest mortality rate in each subgroup, except for G_5, which is the subgroup obtained by removing all dying patients from the complete dataset. It is of some interest to notice that, because we are removing dying patients and this cluster collects those patients who are most likely to die, the corresponding mortality rate index decreases in each smaller subgroup, together with the cluster size. The last cluster, "sick" patients cluster, has a mortality rate which is always between the corresponding values from the "terminally ill" and "healthy" groups. We find, also in this cluster, the same mortality rate's trend as in the "terminally ill" cluster: it diminishes in value from one subgroup to the next. This behaviour finds its natural explanation in the way we selected patients for each subgroup: every

Cluster	**"sick"**			
Subgroup	G_2	G_3	G_4	G_5
Cluster Size	2446	2656	2266	2287
Cluster Size (%)	25.64%	30.62%	27.47%	28.68%
Mortality Rate	5.76%	2.75%	0.97%	0%
$\boldsymbol{\mu}$**(Age)**	71.65 years	71.46 years	71.23 years	71.30 years
sd(Age)	12.06	12.10	13.56	12.12
Cluster	**"terminally ill"**			
Subgroup	G_2	G_3	G_4	G_5
Cluster Size	3151	2136	2048	1792
Cluster Size (%)	33.03%	24.63%	24.83%	22.47%
Mortality Rate	44.91%	29.26%	12.26%	0%
$\boldsymbol{\mu}$**(Age)**	75.02 years	73.37 years	71.99 years	70.96 years
sd(Age)	11.54	11.57	11.65	11.62
Cluster	**"healthy"**			
Subgroup	G_2	G_3	G_4	G_5
Cluster Size	3942	3881	3934	3896
Cluster Size (%)	41.33%	44.75%	47.70%	48.85%
Mortality Rate	0.20%	0%	0%	0%
$\boldsymbol{\mu}$**(Age)**	71.26 years	71.99 years	71.27 years	71.24 years
sd(Age)	12.51	13.54	13.51	13.53

Table 5.2: Table of all the properties of clusters "sick", "terminally ill' and "healthy" obtained in each subgroup (G_2, G_3, G_4 and G_5).

time, removing dying patients, we are reducing the overall mortality rate of the subgroup compared to that of the complete dataset. This result reflects in the gradual reduction of the same index in each cluster (and especially in the "terminally ill" one).

It is also important to look at patients' age distribution for each cluster and its evolution from one subgroup to the next one. As we can see from Table 5.2, "healthy" and "sick" patients clusters have a similar age distribution, a fact highlighted by their mean values, which are around the value of 71 years in every subgroup. Moreover, this similarity is also shown through p-values in Table 5.3, where we carried out a Wilcoxon test with null hypothesis that the distributions of the two considered samples are equal. The shown results demonstrate that these two clusters' age variables are equally distributed over all the considered subgroups. On the contrary, the mean age values for the "terminally ill" patients cluster are considerably higher than the corresponding ones in the other two clusters. It is interesting to

Test	"healthy" vs. "sick"	"terminally ill" vs. "healthy"
$\mathbf{G_2}$	0.92	$< 2.2 \times 10^{-16}$
$\mathbf{G_3}$	0.61	5.79×10^{-7}
$\mathbf{G_4}$	0.24	0.07
$\mathbf{G_5}$	0.34	0.17

Table 5.3: Table of p-values for the Wilcoxon test. First column shows p-values for the Wilcoxon test testing as null hypothesis if ages from "healthy" and "sick" patients clusters are equally distributed. Second column shows p-values for the Wilcoxon test testing as alternative hypothesis if ages from "terminally ill" patients cluster are different from the ones of "healthy" patients cluster. We chose to compare the age distribution of "terminally ill" patients with that of the "healthy" patients because, as shown in the first column, "healthy" and "sick" clusters' ages are equally distributed.

notice that, although this initial characteristic, towards the last subgroups (namely G_4 and G_5) this cluster is reduced to a group of patients whose age distribution is comparable to those of the other two groups (see Table 5.3). This result finds its natural explanation if we recall that dying patients in the complete dataset are, as expected, considerably older than the ones surviving throughout the study time period.

Finally, we come to analyse the size of each cluster. We can notice that, clusters of "sick" and "healthy" patients are both conservative in the number of patients from which they are composed, where "sick" patients cluster covers almost 30% of each subgroup, and "healthy" patients cluster covers 40% to 50% of subgroups. The remaining portion corresponds to "terminally ill" patients. It is important to observe that, in this cluster, the number of patients is always reducing, once more an evident demonstration that patients assigned to this cluster are those about to die. In Figure 5.6, we show the trend of the percentage of patients in cluster k, i.e. the evolution of the percentage of patients in each subgroup assigned to the three clusters. We followed this evolution from the results obtained when dividing in three the complete dataset, to the ones obtained for the smaller subgroup, G_5. We also show (black line) the percentage of dying patients over the complete dataset: for each subgroup, it corresponds to the percentage of patients that have been removed from the complete dataset to obtain the h-th subgroup G_h, where $h = 1, \ldots, 5$. We will show, in Section 5.3, that patients who were selected for a cluster remain in that cluster and that this behaviour is particularly evident for clusters "healthy" and "terminally ill".

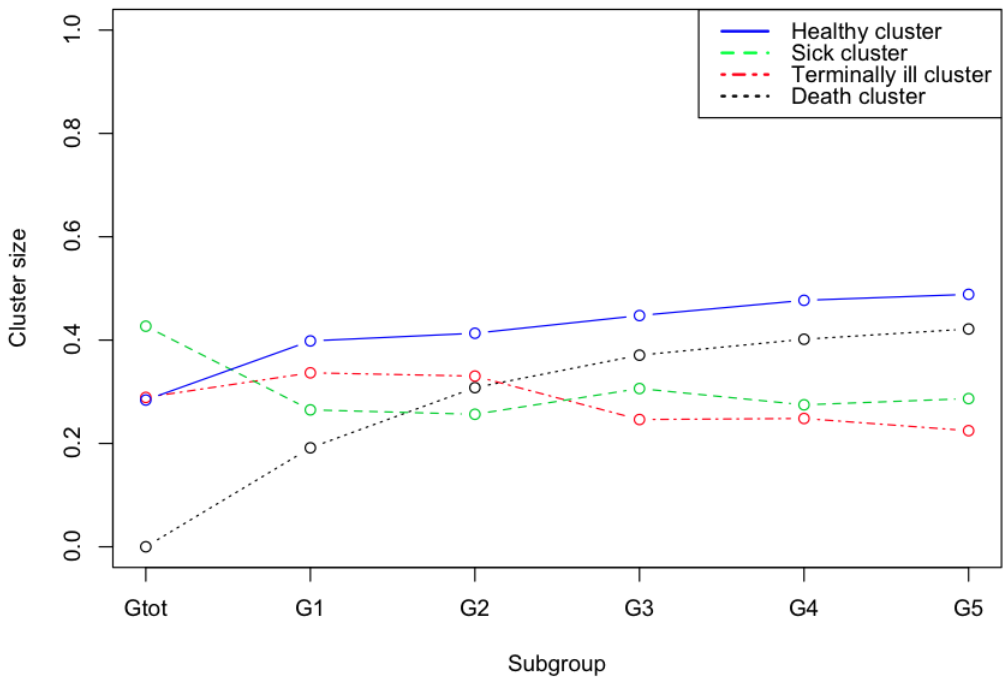


Figure 5.6: Evolution of the size percentage of subgroup G_h represented by each cluster. Trends for "healthy" patients cluster (blue solid line), for "terminally ill" patients cluster (red dotdashed line) and for "sick" patients cluster (green dashed line) are shown. We also added a black dotted line to show the percentage of dead patients in the complete dataset, corresponding to the percentage reduction of patients to obtain the h-th subgroup G_h.

5.2.2 Patient-specific hazard reconstruction within subgroups

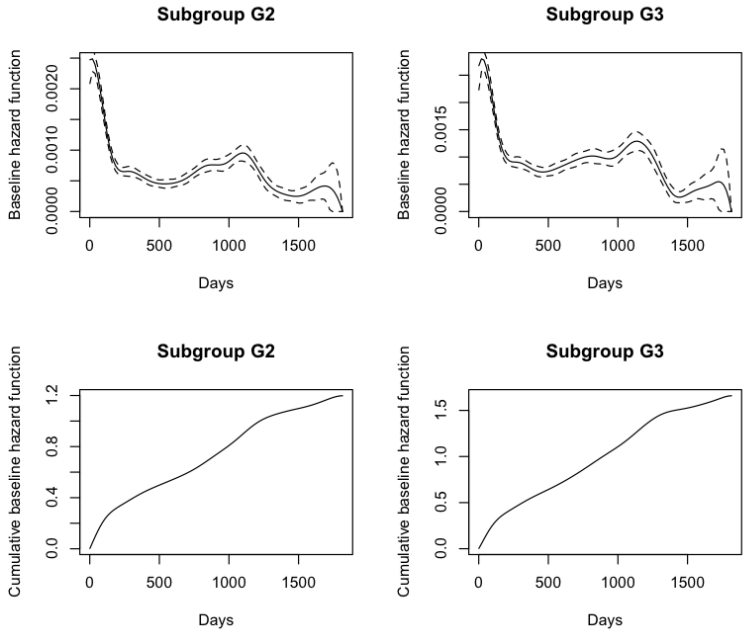
Once again, we want to attempt a reconstruction of the patient-specific hazard functions in each cluster obtained from each subgroup. Applying once more the same techniques presented in Chapter 4, we obtained baseline hazard functions, corresponding patient-specific hazard functions and their cumulatives. Herein we discuss the results obtained for the baseline hazard functions and will only show the obtained patient-specific hazard functions. In Figures 5.7, 5.8 and 5.9, we show the results obtained for the baseline hazard functions of each cluster, comparing the output obtained in each subgroup G_h, where $h = 2, \ldots, 5$. In Figures 5.10, 5.11 and 5.12, we show the patient-specific hazard functions obtained for patients in each cluster, and compare these results for all subgroups. We also show, in each figure, the resulting cumulative hazard functions, corresponding to the estimated baseline hazard function or to the computed patient-specific hazard functions.

First of all, we would like to point out, analysing baseline hazard functions' shapes in Figures 5.7 and 5.9, that these maintain their shape along the four compared subgroups. Moreover, if we compare these plots to the ones obtained from the complete dataset and from subgroup G_1 (see Figure 4.1 and Figure 5.3), we can see that removing dying patients at each hospitalisation is affecting also the functional shape of the estimated baseline hazard functions. In particular, for "sick" patients cluster we appreciate that there is a peak (after 1000 days, i.e. after 2 years and 9 months) in the functional shape that persists throughout all subgroups of patients, with a little variation resulting in the presence of other peaks along the time line. These peaks are the reflection of the possible moments of re-hospitalisation of patients, which in every succeeding subgroup result more evident as they are no longer masked by the presence of dying patients. For what concerns the "healthy" patients cluster, in this case too we are able to appreciate some differences from the results shown in Figure 4.1 and Figure 5.3. The "healthy" cluster, identified from the complete dataset, had a non-zero mortality rate and this fact represents the main reason for the variations appreciable among obtained estimations of the baseline hazard function in each subgroup. The shy bending we saw in the baseline hazard function's shape in Figure 4.1 (after 1000 days, i.e. after half the time of observation), becomes more and more evident in each successive subgroup, with a following increase in the uncertainty to estimate the baseline hazard function (see the range of possible values explored by the confidence bands, for example equal to $[0; 0.0078]$ for the fifth subgroup). It is important, at last, to notice that these variations in the functional shape of the baseline hazard functions can not be equally appreciated in the corresponding cumulative baseline hazard functions, which show the general behaviour of the clusters' re-hospitalisation hazard. As expected, and as already proved through the descriptive analyses in the previous section, the cumulative functions do not show any variation among the considered subgroups for the "healthy" patients cluster.

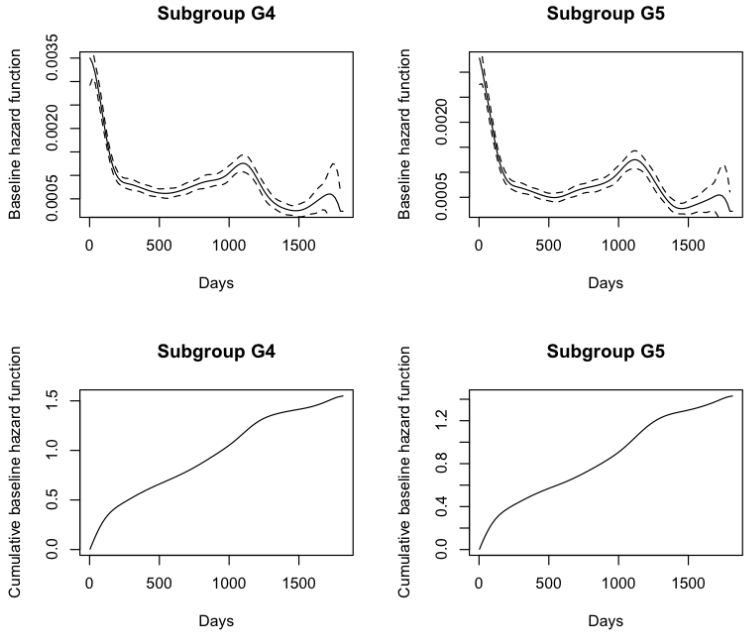
The estimation for the baseline hazard function of "terminally ill" patients cluster is the one, as foreseen, that shows the greatest changes among subgroups. Every time we remove dying patients, the cluster that undergoes major variations is the "terminally ill" cluster, which is, as seen in Section 5.2.1 and in previous chapters, the cluster with the higher mortality rate. We saw, in Figure 4.1, that the estimated hazard function for the "terminally ill" patients cluster distinguishes itself from the other clusters because it reaches the 5×10^{-4} limit within the first 164 days and remains below it for all the study time period. In particular, at the end of the observation period, it reaches the zero limit line. Removing patients dying before their

second or third hospitalisation, introduces a great change in the estimate of the baseline hazard function, a variation that is even more evident when we remove also the rest of dying patients. The baseline hazard functions, for "terminally ill" patients in each considered subgroup, have an initial behaviour that matches that obtained for this cluster in the complete dataset. This trend is maintained until the first half of the study time period, when the confidence bands show an increase in the uncertainty of the final part. What is fascinating about the results from subgroups G_3, G_4 and G_5, is that the baseline hazard function, after 1500 days of observation, increases its value in such a way that it resembles the functional shape of the "healthy" patients cluster. Once again, we find that dying patients where masking the natural behaviour of the other patients forming part of the cluster.

We think it is relevant to show the results obtained once we reconstructed the patient-specific hazard functions. Patient-specific hazard functions computed for "sick" patients and "terminally ill" patients clusters are comparable to the baseline hazard functions estimated for these clusters in each subgroup. In particular, we are able to appreciate from Figure 5.10 and Figure 5.11 the same functional shape as for the baseline hazard functions, with a little variability among patients within each cluster of each subgroup, due to the combination of the effects of the covariates and the variability of the frailty term. This analysis becomes important when we look at the results of the computed patient-specific hazard functions for the "healthy" patients cluster: the consequence of removing dying patients, as stated in Section 5.2.1, is in the ability to identify a cluster of living patients (in fact this is the cluster with mortality rate equal to zero for G_2, G_3, G_4 and G_5). We already discussed the general behaviour of this cluster in each subgroup, but it is now relevant to look at the patient-specific results: as we can see from Figure 5.12, patients identified as "healthy" patients seem to have different behaviours within the same cluster obtained in each subgroup. In particular, it looks like we could find three different groups of patients that differentiates one another for their functional values throughout the study time period. For this reason, we decided to conduct a brief yet very remarkable analysis, to deeper investigate the characteristics of patients in the "healthy" patients clusters from each subgroup (see Section 5.2.3).

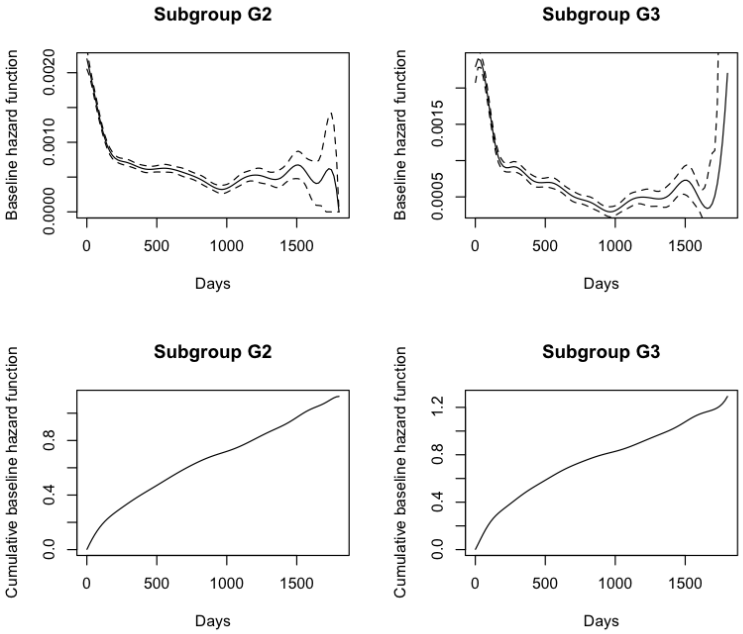


(a) *Baseline hazard functions and corresponding cumulative baseline hazard functions in subgroups G_2 and G_3.*

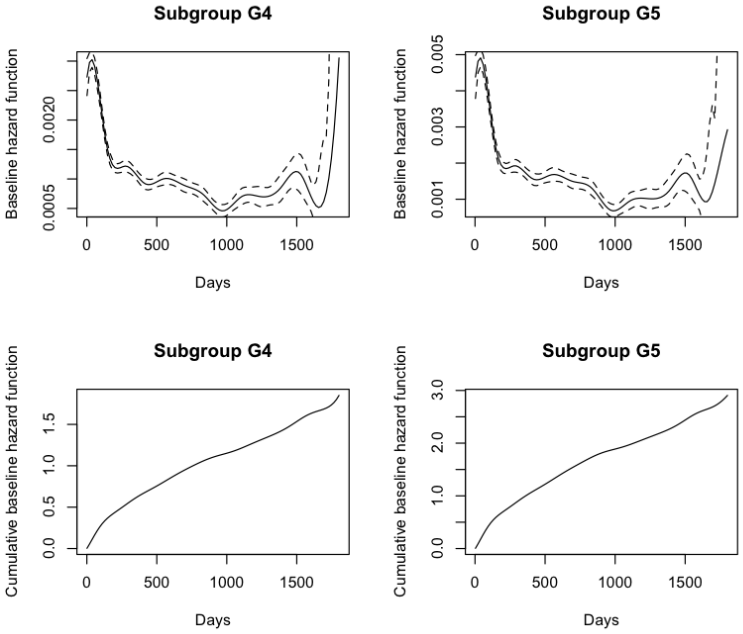


(b) *Baseline hazard functions and corresponding cumulative baseline hazard functions in subgroups G_4 and G_5.*

Figure 5.7: Baseline hazard functions and corresponding cumulative baseline hazard functions for "sick" patients clusters in subgroups G_2, G_3, G_4 and G_5.

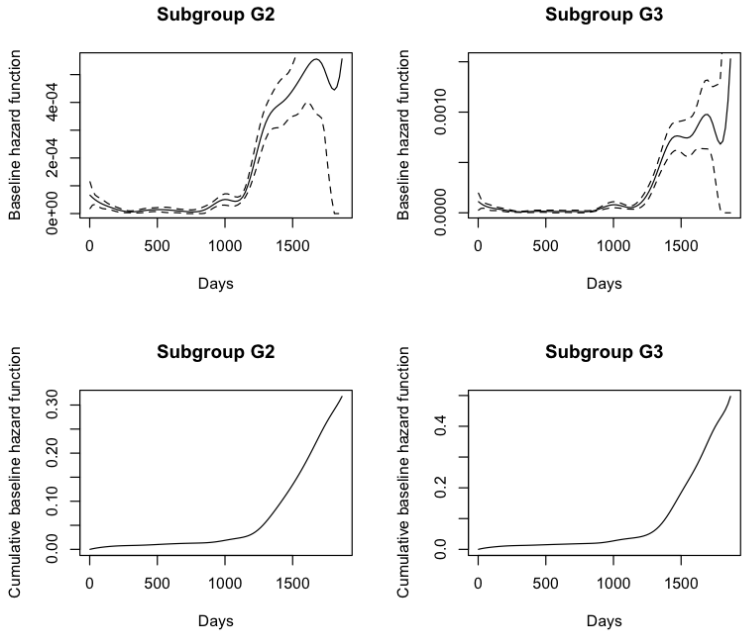


(a) *Baseline hazard functions and corresponding cumulative baseline hazard functions in subgroups G_2 and G_3.*

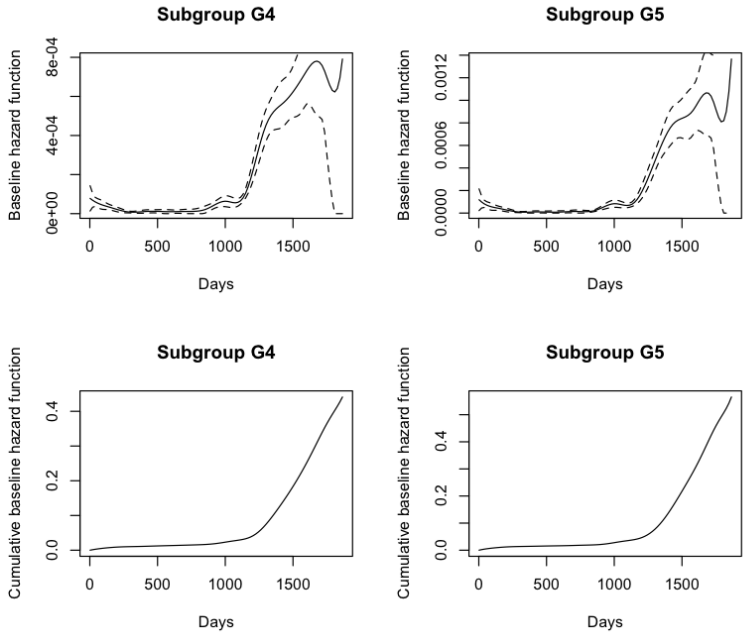


(b) *Baseline hazard functions and corresponding cumulative baseline hazard functions in subgroups G_4 and G_5.*

Figure 5.8: Baseline hazard functions and corresponding cumulative baseline hazard functions for "terminally ill" patients clusters in subgroups G_2, G_3, G_4 and G_5.

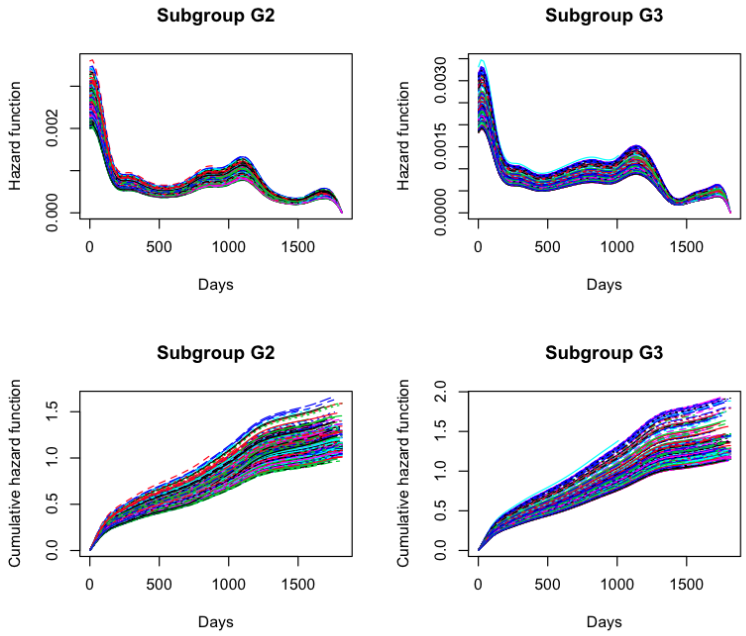


(a) *Baseline hazard functions and corresponding cumulative baseline hazard functions in subgroups G_2 and G_3.*

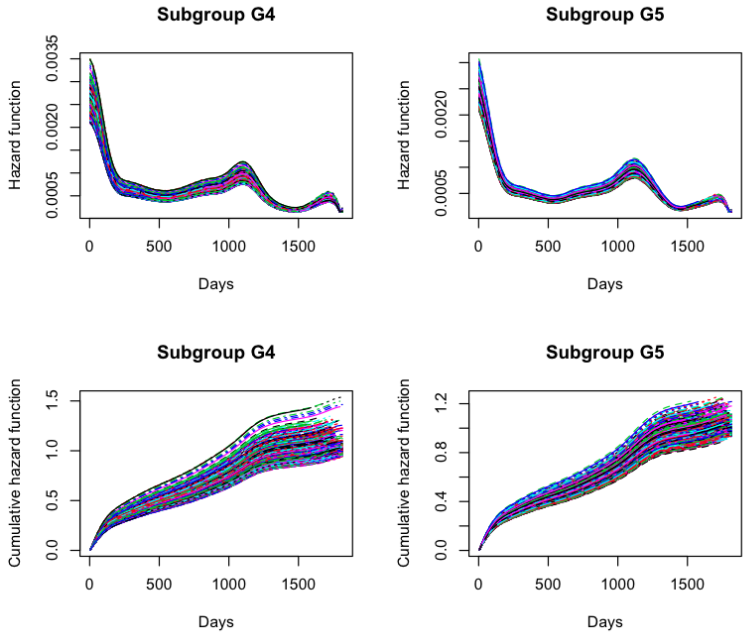


(b) *Baseline hazard functions and corresponding cumulative baseline hazard functions in subgroups G_4 and G_5.*

Figure 5.9: Baseline hazard functions and corresponding cumulative baseline hazard functions for "healthy" patients clusters in subgroups G_2, G_3, G_4 and G_5.

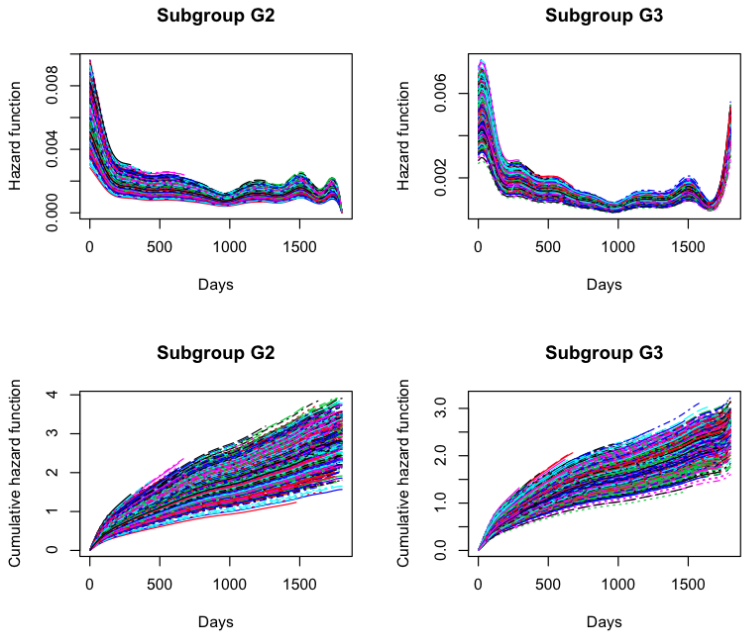


(a) *Patient-specific hazard functions and corresponding cumulative hazard functions in subgroups G_2 and G_3.*

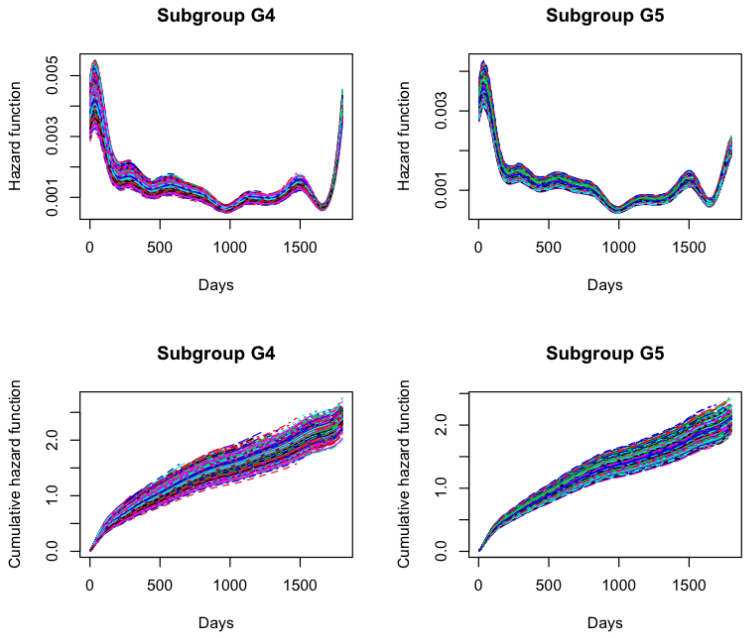


(b) *Patient-specific hazard functions and corresponding cumulative hazard functions in subgroups G_4 and G_5.*

Figure 5.10: Patient-specific hazard functions and corresponding cumulative hazard functions for "sick" patients clusters in subgroups G_2, G_3, G_4 and G_5.

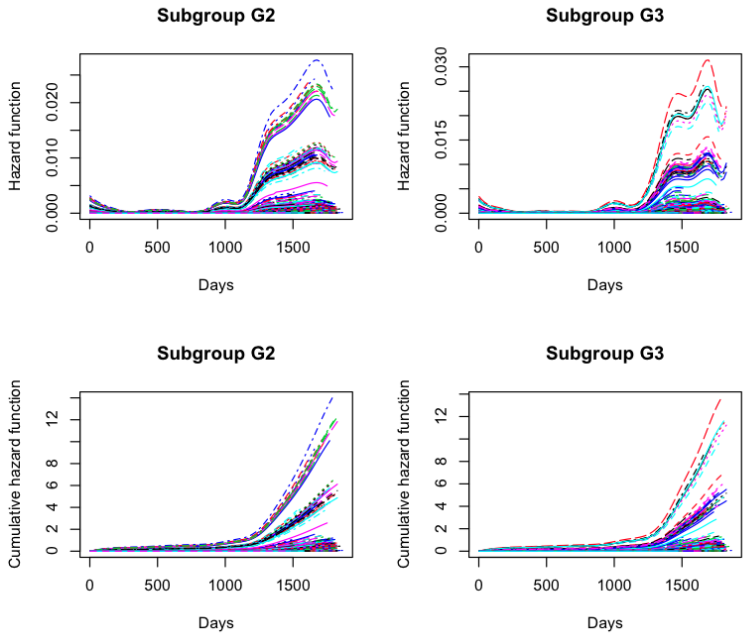


(a) *Patient-specific hazard functions and corresponding cumulative hazard functions in subgroups G_2 and G_3.*

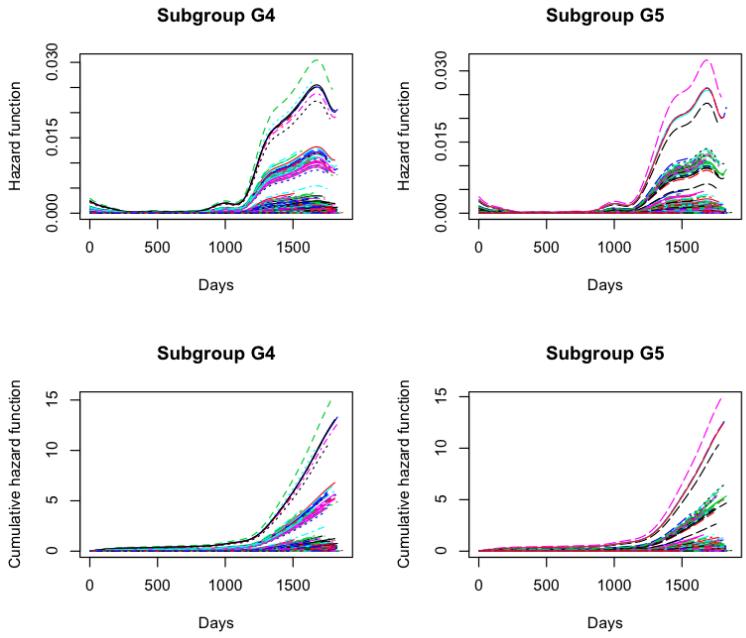


(b) *Patient-specific hazard functions and corresponding cumulative hazard functions in subgroups G_4 and G_5.*

Figure 5.11: Patient-specific hazard functions and corresponding cumulative hazard functions for "terminally ill" patients clusters in subgroups G_2, G_3, G_4 and G_5.



(a) *Patient-specific hazard functions and corresponding cumulative hazard functions in subgroups G_2 and G_3.*



(b) *Patient-specific hazard functions and corresponding cumulative hazard functions in subgroups G_4 and G_5.*

Figure 5.12: Patient-specific hazard functions and corresponding cumulative hazard functions for "healthy" patients clusters in subgroups G_2, G_3, G_4 and G_5.

5.2.3 Analysis of "healthy" patients clusters

As we already pointed out, removing patients dying before their third, fourth, fifth hospitalisation or before the end of the study time period (correspondingly subgroups G_2, G_3, G_4 and G_5), we obtained "healthy" patients clusters with a strong inner characterisation: first of all, these are the clusters made of non-dying patients (with an exception of little importance for subgroup G_2). This means that every time we remove patients dying before their h-th hospitalisation, we are not really modifying the configuration of the "healthy" cluster. Nevertheless, the number of patients in this cluster undergoes slight changes: this is because, as we will discuss in Section 5.3, patients initially labelled as "sick" patients move between clusters when we remove dying patients. This is because the clustering algorithm applied is not always able to distinguish, based on survival times, "sick" patients behaviour from that of "healthy" patients. Luckily, the number of patients with this characteristic does not affect the general behaviour of "healthy" patients cluster, and is for this reason negligible. Moreover, patients in our dataset have been selected because they were diagnosed with heart failure. This means that, in spite of the fact that we labelled these patients as "healthy" patients, they are suffering from this disease. We could say that they are, differently from what we could say for the other clusters, at early stages of their disease, i.e. that their condition has been diagnosed early in terms of its evolution. Over the five years of observation, we are able to analyse and follow the evolution of the disease for these patients thanks to multiple admissions. As said in previous section, observing the behaviour of these patients from a risk of re-hospitalisation point of view, we clearly came to understand that these patients have different disease evolution trends within the same cluster (see Figure 5.12). To deeply study this characteristic of "healthy" clusters, we decided to conduce a cluster analysis over these patients.

We chose to cluster patients within the "healthy" patients groups obtained from subgroups G_2, G_3, G_4 and G_5 because, as already discussed, these subgroups are the ones were the clusters characteristics become evident. We decided to use a hierarchical clustering method to compute and find the clusters within the patients under analysis. In particular, notice that from Figure 5.12, it is evident that we can divide patients into three clusters, based on the shape of the hazard function. Using R functions `dist` and `hclust`, we computed all possible clusters with the considered population, repeating this procedure for each subgroup. The distance matrix is computed as the euclidean distance between hazard functions, as we now explain. Patients in the dataset are observed during different time intervals, as they do not have their first admission in the same day. For this reason,

the computed hazard functions are aligned at the first admission, which is the $t = 0$ time for every patient. Then we will observe the patient, hence compute her/his hazard function, up until the end of the study time period (remember that these patients do not die). As a result, patient-specific hazard functions differ in time length. It follows that, in order to build the distance matrix, we decided to truncate the hazard functions at the time corresponding to the last t where we have a functional value for every patient. The resulting interval of observation of the hazard functions starts at $t = 0$ and ends after $t = 1451$ days, equal to 4 years. Once we truncated the hazard functions, we computed the euclidean distance matrix between functions as follows:

$$d_{ij} = d(h_i; h_j) = \sqrt{\sum_{l=1}^{M} \Big(h_i(t_l) h_j(t_l) \Big)}, \qquad \forall i \neq j, \quad i, j = 1, \dots N_h, \quad (5.1)$$

where h_i and h_j are the hazard functions of patients i and j, t_l is the l-th time point where function `frailtyPenal` evaluated the hazard functions (see Section 4.3), M is the last time point where all the hazard functions are observed (and it is such that $M < 100$ and $t_M = 1451$), N_h is the number of patients in the "healthy" cluster obtained for the h-th subgroup.

Once we computed the distance matrix d_{ij} for each subgroup G_h, we are ready to run the hierarchical clustering method to obtain the desired clusters. The agglomeration method we decided to use is the complete linkage, which computes the distance between a cluster of objects, let's call it (UV), and any other cluster W as:

$$d_{(UV)W} = \max(d_{UW}, d_{VW}). \quad (5.2)$$

We then cut the resulting agglomerative tree in order to obtain three clusters. We repeated this procedure for each "healthy" patients cluster in subgroups G_2, G_3, G_4 and G_5. In Figure 5.13, we show the clustering results for each subgroup. It is an extraordinary result to see that the algorithm is able to find inner clusters within "healthy" patients in each subgroup, selecting three different clusters of patients. In particular, we see from the obtained plots that the algorithm, as expected, merges together into one cluster those functions having the most similar shapes.

At this point of the analysis it becomes interesting to take a look at the obtained clusters' properties among subgroups. Remembering that these are all "healthy" patients, we decided to take into account, in order to understand which patients were assigned to the inner clusters, several indexes:

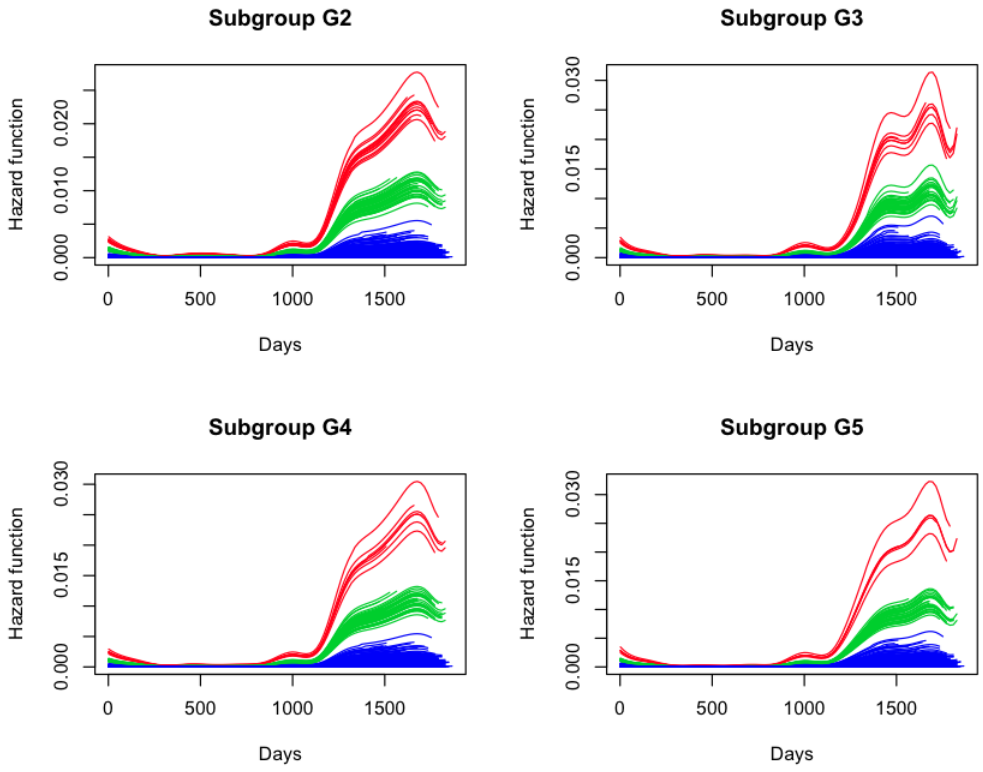


Figure 5.13: Resulting clusters from the hierarchical clustering method used to group "healthy" patients in each subgroup, G_2, G_3, G_4 and G_5.

the mean value of the frailty term estimated when computing the patient-specific hazard functions (see Section 4.3), the mean value of patients' age and the maximum number of hospitalisations for patients in the considered inner clusters. The results are shown in Table 5.4.

First of all, notice that the size of each inner cluster (i.e. clusters labelled as 1, 2 and 3 in Table 5.4) is coherent among all subgroups. The first inner subgroup, namely cluster 1, is the largest cluster, covering up to 98% of the total "healthy" patients group. This inner cluster is, indeed, the one representing the mean behaviour of the considered patients. Secondly, it is interesting to consider the mean value of the frailty term estimated to compute the hazard functions. As we recall from the proposed model (see Section 4.1), the frailty term was introduced to allow the model to evaluate the heterogeneity within the considered group of patients. Here, we can see that the frailty term represents a crucial role in the computation and in the clustering of the resulting patient-specific hazard functions. It permits us, in this considered case, to appreciate the differences between

Subgroup	**G_2**		
Cluster	1	2	3
Cluster size	3880 (98.43%)	47 (1.19%)	15 (0.38%)
Mean frailty	0.23	4.56	5.07
Mean age	71.24 years	73.81 years	77.13 years
Max admissions	3	3	4
Subgroup	**G_3**		
Cluster	1	2	3
Cluster size	3831 (98.71%)	42 (1.08%)	8 (0.21%)
Mean frailty	0.20	4.63	4.94
Mean age	71.24 years	73.26 years	72.13 years
Max admissions	3	3	4
Subgroup	**G_4**		
Cluster	1	2	3
Cluster size	3881 (98.65%)	45 (1.14%)	8 (0.21%)
Mean frailty	0.23	4.59	4.95
Mean age	71.24 years	73.64 years	72.13 years
Max admissions	3	3	4
Subgroup	**G_5**		
Cluster	1	2	3
Cluster size	3849 (98.79%)	41 (1.05%)	6 (0.16%)
Mean frailty	0.21	4.65	4.78
Mean age	71.22 years	73 years	70.5 years
Max admissions	3	3	4

Table 5.4: Table of the properties of clusters obtained from "healthy" patients in each subgroup G_h, for $h = 2, \ldots, 5$.

Subgroup	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{G_4}$	$\mathbf{G_5}$
p-value	4.26×10^{-6}	0.0216	0.0209	0.1801

Table 5.5: Table of the p-values for the Wilcoxon test, conducted to check whether the frailty terms from inner clusters 2 and 3 in each subgroup are equally distributed.

patients and to distinguish them into well separated groups. Moreover, we see that inner clusters 2 and 3 have similar mean values among subgroups, a characteristic that is reinforced by the Wilcoxon test conducted over the two groups of frailty terms to see if they could be considered equally distributed (see Table 5.5). The term that allows to differentiate from patients in the inner clusters 2 and 3 is the patients' age, as we can see from Table 5.4. Moreover, patients in the inner cluster 3 experience more admissions than the other patients, with the consequence of a higher estimated risk of being re-hospitalised. These behaviours can be clearly appreciated in Figure 5.13, where hazard functions of patients in cluster 3 are coloured in red, cluster 2 in green and cluster 1 in blue, for each subgroup.

5.3 Moving among clusters

We now come to the last, but not least, analysis conducted over subgroups G_h, for $h = 1, \ldots, 5$. We discussed, in previous sections, all the properties of clusters obtained each time we removed dying patients and affirmed that these clusters, with minor exceptions, became more and more well characterised, leading to the idea that each patient is reassigned to the same cluster of the successive subgroup. To confirm this idea, it is important to analyse the movement of every patient among clusters.

To conduct this analysis, we took advantage of the ID label assigned to every patient, in order to be able to trace their movements. We then build a matrix with 7 columns and as many rows as patients in the complete dataset, i.e. 13785. We assigned to the first column the IDs of patients, while the remaining columns are for the clusters' labels at each step of the reduction of patients number: second column is for the labels of clusters obtained with the `mixPHM` algorithm over the complete dataset G_{tot}, third column is for the labels of clusters obtained with the same algorithm applied over subgroup G_1, and so on. If a patient dies, i.e. at a certain point she/he is removed from the dataset, in the next subgroups her/his label corresponds to "death". In Figures 5.14, 5.15 and 5.16 we show the paths traced by patients who were initially labelled, i.e. from the complete dataset, as "sick", "terminally ill"

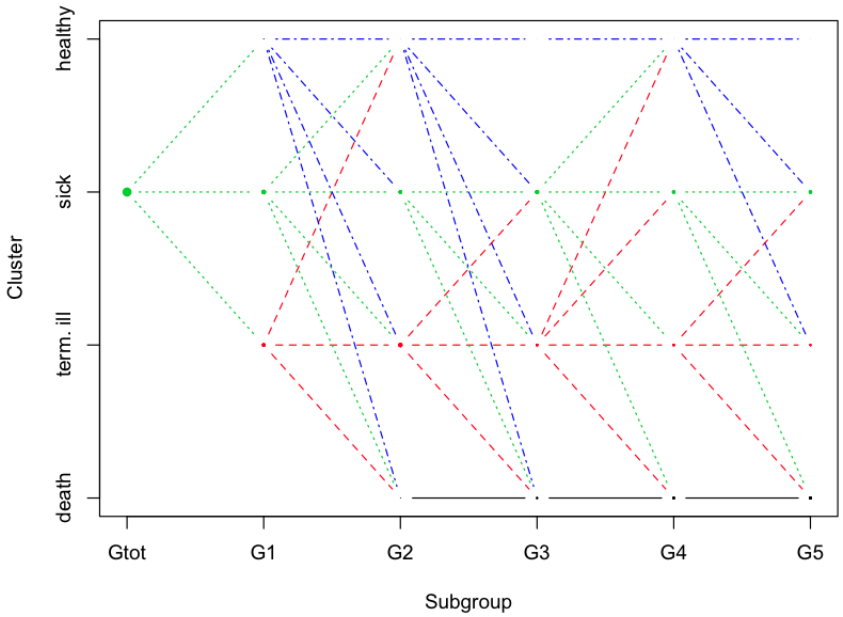


Figure 5.14: Representation of the movement of patients clustered as "sick", starting from the complete dataset, through clusters obtained when running the algorithm 'mixPHM' over subgroups G_1, G_2, G_3, G_4 and G_5.

or "healthy", respectively. In the shown plots, the coloured points have two characteristics, helpful to better interpret the movements of patients: their colour is equal to the colour of the corresponding cluster (blue for "healthy" cluster, green for "sick" cluster, red for "terminally ill" cluster and black for dead patients), and their dimension corresponds to the proportion of patients assigned to the considered cluster. For example, some of the patients in the "healthy" cluster from the complete dataset G_{tot}, are then assigned, when we consider subgroup G_1, to the new "healthy" cluster, while part of them die: the dots shown for subgroup G_1 are as big as the proportion of patients going either to the "healthy" cluster or to the "death" cluster, and, as we already knew, the proportion of dying patients is minimal (equal to 0.28% of the initial "healthy" cluster). This rule applies in every figure.

The first thing that is evident from the shown plots is that, as already discussed, the three clusters have distinctive behaviours: the majority of patients who where at first assigned to the "terminally ill" cluster died and were removed from the dataset, while the ones that survived are reassigned to the new obtained "terminally ill" clusters (notice that, even though there are patients assigned to the "healthy" cluster when analysing subgroup G_1, these are then pushed back to the "terminally ill" cluster). The "healthy"

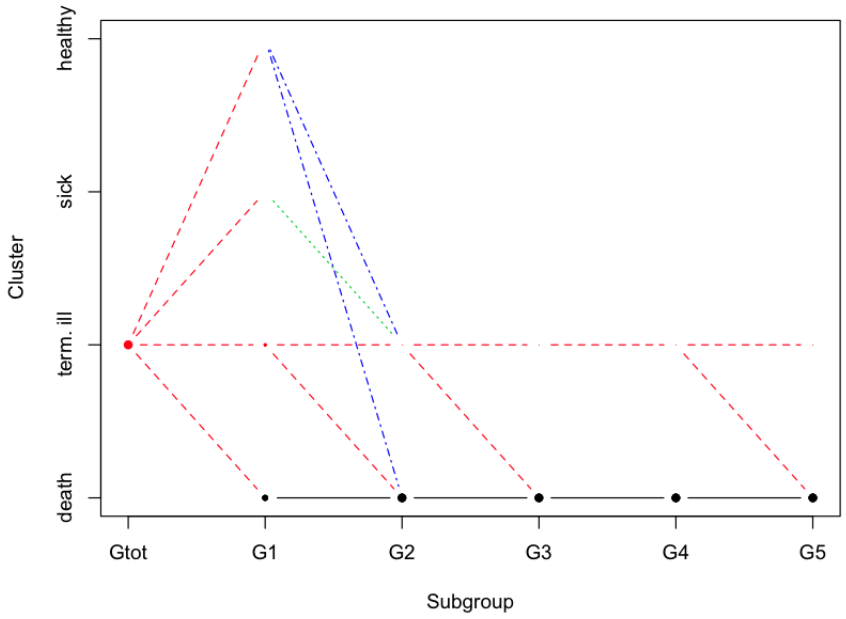


Figure 5.15: Representation of the movement of patients clustered as "terminally ill", starting from the complete dataset, through clusters obtained when running the algorithm 'mixPHM' over subgroups G_1, G_2, G_3, G_4 and G_5.

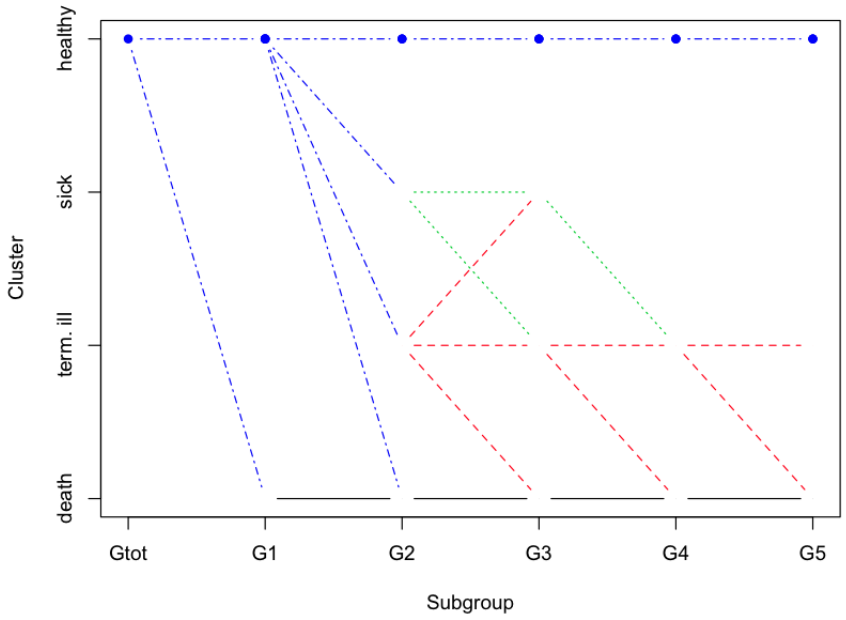


Figure 5.16: Representation of the movement of patients clustered as "healthy", starting from the complete dataset, through clusters obtained when running the algorithm 'mixPHM' over subgroups G_1, G_2, G_3, G_4 and G_5.

patients cluster has a specular behaviour, as we see that, apart from an initial and minor exception, these patients are always classified as "healthy" patients. This result reinforces, once more, the initial idea that clusters of "terminally ill" and "healthy" patients have strong characterisations and are, for this reason, always clearly identified, especially after we removed patients dying before their third hospitalisation (see Section 5.2). Patients in the "sick" cluster show, as throughout all the analyses conducted in previous sections, an *in between* behaviour that leads to more randomised paths, as we can see from Figure 5.14. Every time we remove dying patients, we refine the characteristics of "terminally ill" and "healthy" clusters, but we are not able to clearly identify specific characteristics for the "sick" patients cluster. Nevertheless, we can see from the shown plot that the majority of patients from this cluster moves among "sick", "terminally ill" or "death" clusters, as we can see that the points for the "healthy" patients cluster are essentially of null dimension.

Chapter 6

Conclusions and future developments

In this thesis we analysed data from patients diagnosed with heart failure and chronic heart failure, applying several algorithms in order to investigate on relevant features. Indeed, we found a successful outcome when dividing patients into several groups, whose characteristics were easily recognised by means of a descriptive analysis. Furthermore, we extended the analysis to the functional estimation and reconstruction of patients' hazard functions. As already discussed, the obtained results demonstrate a strong capability of this approach to estimate and recognise the re-hospitalisation process of the considered population. Moreover, it allows a patient-specific analysis that can be useful to detect subjects' behaviour during the time of observation.

When selecting patients to be included in the developed analyses, we decided to consider subjects with a maximum of five hospitalisations over the studied time period. Even though we do not expect great changes in the outcome, it could be of some interest to consider the entire population and apply the studied algorithms to the most complete dataset. Furthermore, a new dataset, gathering a greater number of patients observed over a longer time period, is now available: we will apply the techniques used throughout this thesis over this new population, expecting to obtain more refined estimations thanks to the enlarged number of patients. This newer dataset includes more information: in particular, documentation on drugs taken once dismissed is provided, a data that allows to extend our analyses to a comparison between estimated re-hospitalisation hazards of patients and the kind, dosage and frequency of taken drugs. Additionally, we would like to introduce in our model a stratification of data, distinguishing between women and men. This would go in the direction of epidemiological litera-

ture, which shows an existing difference between women and men diagnosed with heart failure, differences that we also found in our analyses.

At the beginning of the study, we excluded from the considered population those patients who had at least once in their hospitalisations history a circulatory shock. This appeared to be a population with proper features: nevertheless, these are patients diagnosed with chronic heart failure and we will need to study and analyse the re-hospitalisation process of these distinct subjects. Finally, one could discuss over the intra-hospital death index, comparing this phenomenon to deaths outside the hospital. This information could allow for better explanation of groups composition.

Once we refined the estimations obtained in this thesis by means of the proposed developments, the analysis could be extended to the study of the changing points in the re-hospitalisation process of each patient. Being able to understand this specific behaviour could improve the diagnosis of the disease and the ability of a clinician to follow its evolution. Moreover, a better understanding of the disease trend from a general point of view, could allow hospitals to preview the needs of future hospital admissions, in an effort to improve the efficiency of clinical facilities and, consequently, collective welfare.

Appendix A

Proportional Hazards Model

In this appendix we would like to report a brief story of Proportional Hazards Models [5] [6]. These are a class of survival models where the only effect due to the variation of the value of a covariate under investigation, is multiplicative with respect to the hazard rate function of an underlying survival distribution. These models require no specific parametric form for the survival distribution. Sir David Cox in his first work on these models [7], considered first of all the problem of evaluating the relation between the distribution of the time to event data and the variables $z_1, \ldots, z_p$, where we store all the measurements which are available from a patient's history. The rising model from his conjecture is of the form:

$$h(t; \mathbf{z}) = h_0(t) exp(\boldsymbol{\beta}' \mathbf{z}), \tag{A.1}$$

where $\boldsymbol{\beta}$ is the vector of unknown parameters of the regression model and $h_0(t)$ is the resulting hazard function setting $\mathbf{z} = \mathbf{0}$.

The second problem Cox attempted to solve, was the analysis of the obtained models. In his paper, he outlined the necessity to look for the consequences of letting the baseline hazard function $h_0(t)$ be arbitrary, concentrating mostly on the attempt to estimate and compute regression parameters. The $\boldsymbol{\beta}$ parameters vector is for this reason estimated through a maximum likelihood method, computing the partial likelihood rising from this model. This is of the form:

$$L(\beta) = \prod_{i:C_i=1} \frac{\theta_i}{\sum_{j:T_j \geq T_i} \theta_j}, \tag{A.2}$$

where $\theta_i = exp(\mathbf{z}_i \boldsymbol{\beta})$, T_i denotes the observed time (censoring or event time) for subject i and C_i the indicator whether the corresponding time is an event or has been censored. The corresponding log partial likelihood is:

$$\ell(\boldsymbol{\beta}) = \sum_{i:C_i=1} \left(\boldsymbol{\beta}' \mathbf{z}_i - \log \sum_{j:T_j \geq T_i} \theta_j \right). \tag{A.3}$$

Using some maximisation algorithm we are now able to determine the estimate of $\boldsymbol{\beta}$. Notice that all sorts of generalisations are allowed, using time-varying covariates and time-varying coefficients as well. Moreover, if for some reason we can assume that $h_0(t)$ has a particular shape rising from the knowledge of the studied problem nature, then we can replace it with the desired function. An example is that of the Weibull hazard function, which leads to the well known "Weibull proportional hazards model".

It is important to give some kind of physical explanation about the nature of Cox model. As said in his work [7], proportional hazards model is significant to accelerated life testing. Let s be a "stress" variable and suppose that the basic physical process of failure is common for different stress levels. This condition is satisfied when observation units experience only one way of failure, or at least only a predominant one. Of course, the complexity of this approach is to come to know the physical problem in order to establish the correct test conditions and their properties. The hazard function at stress level s, supposing that the failure process is the same for every s, is of the form $g(s)h_0\{g(s)t\}$, where $g(s)$ is a function of s to be determined, such that $g(1) = 1$. If we assume that some ageing process proceeds independently of s, then the model is reduced to $g(s)h_0(t)$. Now, setting $g(s) = s^\beta$ and $z = \log s$, the resulting hazard model is the same as in equation (A.1).

Finally, for the purpose of research and personal knowledge, we transcribed a passage of the interview to Sir Cox by Nancy Reid [27], where he describes the background leading to his work and his Proportional Hazards Model.

> *Quite a few people - I think particularly of Peter Armitage and Ed Gehan and Marvin Zelen, and I think there were others - said they were getting a certain kind of data, censored survival data, with a lot of explanatory variables. Nobody knew quite how to handle this sort of data in a reasonably general way, and there seemed to be dissatisfaction with assuming an underlying exponential distribution or Weibull distribution modified by some factor. It seemed that something slightly more general was called for. Well, in the light of all sorts of things I'd done in stochastic processes it's entirely natural to approach this in terms of hazard.*

So the specification of some basic function of the underlying time scale, multiplied by a factor, that's sort of immediate and obvious really. I don't know whether it's new to that paper, I think probably it is, but anyway it's sort of immediate.

Then the question was how to actually do the statistical analysis. I wrote down the full likelihood function and was horrified at it because it's got exponentials of integrals of products of all sorts of things, unknown functions and so forth. I was stuck there for quite a long time - I would think the best part of five years or maybe even longer. Then suddenly I thought that the obvious thing to do was to concentrate on the part of the likelihood that actually gave you the information about the regression coefficients that you were interested in. It was absolutely obvious how to do that, and so just write down the answer. It occurred to me while I had a high temperature and was in bed with flu. It suddenly struck me that you could do this, and then when I felt better I tried to recover the argument and couldn't. But I was so convinced that when I was ill I had done this, that I tried again and then I saw what it was that I'd done. [...]

So that's the essence of it. It didn't come from one particular application, but it came from perceiving, on the advice of others, that in medical statistics people were getting a certain kind of data that they didn't know how to analyse. And I think, though it's a long time ago and I don't remember too clearly, I could conceive that in industrial reliability and perhaps other fields essentially the same problems were arising.

Appendix B

"Whatever works": The code

There is nowadays a strictly connection between statistical analysis and computer programming. This thesis project was mostly concentrated on the analysis element of this connection, although behind all the presented results there was a necessary large use of the computer power. For this reason, we think it could be of some interest to present here the major aspects of the code we wrote to produce the discussed analyses.

We herein also would like to quote a picture that inspired us while writing the below code, for the quoted sentence apply perfectly to the love and hate relationship we had during these last months with the computer:

> *Love, despite what they tell you, does not conquer all, nor does it even usually last. In the end the romantic aspirations of our youth are reduced to, whatever works.*
>
> *Whatever Works, 2009, Woody Allen.*

```
1  ###################################
2  ### Creating Proper Matrices ###
3  ###################################
4
5  # The Dataset
6  dati <- read.table('dati2006_over18dateOKlosOK.txt')
7  IDs <- unique(dati$ID) # Patients IDs
8
9  #### Matrix for the mixPHM package ####
10
11 dateADM <- as.Date(dati$dateADM)
12 dateDEATH <- as.Date(dati$dateDEATH)
13 dateEND <- as.Date("2010-12-31")
14
15 ## Maximum columns number
16 maxcol <- max(dati$adm_number)
17
18 #Initialising the matrix
19 times_mixPHM <- matrix(NA, nrow = length(IDs), ncol = maxcol)
20 for(i in 1:length(IDs))
```

```
{
  #Patient i admission dates
  paz_times <- dateADM[which(dati$ID==IDs[i])]
  for(j in 2:count_col[i])
  {
    #hospitalisation length= number of days between two consecutive
        admissions
    times_mixPHM[i,j-1] <- as.numeric(difftime(paz_times[j],
                                               paz_times[j-1],
                                               units="days"))
  }

  #last hospitalisation (censored or death)
  if((dati$DEATH_ind[which(dati$ID==IDs[i])])[1]==0) # Censored
  {
    times_mixPHM[i,j] <- as.numeric(difftime(dateEND,
                                             paz_times[j],
                                             units="days"))
  }else # Death
  {
    times_mixPHM[i,j] <- as.numeric(difftime(dateDEATH[which(dati$ID==IDs[i
        ])][1],
                                             paz_times[j],
                                             units="days"))
  }
}

times_mixPHM <- data.frame(IDs, times_mixPHM)

#### removing patients who had at least one shock.
#### removing patients who have more than 5 amds.
shock <- NULL
count_col <- NULL
for(i in 1:length(IDs))
{
  shock <- c(shock,ifelse(sum(dati$SHOCK[which(dati$ID==IDs[i])])>0, 1, 0))
  count_col <- c(count_col, dim(dati[which(dati$ID==IDs[i]),])[1])
}

n <- 5
elimino_n <- which(count_col > n)

# Selecting patients with a maximum of 5 hospitalisations
times5_mixPHM <- as.matrix(times_mixPHM[-elimino_n, 1:(n+1)])

shock <- shock[-elimino_n]
elimino_shock <- which(shock==1)
# Selecting patients with no shock
times5_mixPHM <- as.matrix(times5_mixPHM[-elimino_shock,])

write.table(times5_mixPHM, "times5_mixPHM.txt")


#### Matrix for the frailtypack package ####
id.good <- NULL
t.start <- NULL
t.stop <- NULL
time <- NULL
event <- NULL
enum <- NULL
death <- NULL
age <- NULL
```

```
shock <- NULL
cabg <- NULL
ptca <- NULL
icd <- NULL
stent <- NULL
for(i in 1:length(IDs))
{
  dati.id <- dati[which(dati$ID==IDs[i]),c(1:9,15:19)] #patient i
  dim.id <- dim(dati.id)[1]
  if(dim.id<=5) # patients with a maximum of 5 adms
  {
    id.good <- c(id.good, rep(dati.id[1,1], times=dim.id)) # ID
    enum <- c(enum, c(1:dim.id)) # event number
    has.any <- colSums(dati.id[,10:14]) # accounting for procedures
    if(has.any[1]>0) { shock <- c(shock, rep(1,times=dim.id))
                       }else{ shock <- c(shock, rep(0,times=dim.id)) }
    if(has.any[2]>0) {cabg <- c(cabg, rep(1,times=dim.id))
                       }else{ cabg <- c(cabg, rep(0,times=dim.id)) }
    if(has.any[3]>0) {ptca <- c(ptca, rep(1,times=dim.id))
                       }else{ ptca <- c(ptca, rep(0,times=dim.id)) }
    if(has.any[4]>0) {icd <- c(icd, rep(1,times=dim.id))
                       }else{ icd <- c(icd, rep(0,times=dim.id)) }
    if(has.any[5]>0) {stent <- c(stent, rep(1, times=dim.id))
                       }else{ stent <- c(stent, rep(0,times=dim.id))}
    age <- c(age, rep(dati.id[1,2], times=dim.id)) # Age at first event
    if(dim.id==1) # only 1 event
    {
      if(dati.id[1,9]==1) # death
      {
        time.id <- as.numeric(difftime(as.Date(dati.id[1,8]),
                                       as.Date(dati.id[1,6])))
        death.id <- 1
        event.id <- 1 # Not censured event
      }else # end of the study time
      {
        time.id <- as.numeric(difftime(as.Date("2010-12-31"),
                                       as.Date(dati.id[1,6])))
        death.id <- 0
        event.id <- 0 # Censured event
      }
    }else # more than 1 event
    {
      time.id <- NULL
      event.id <- NULL
      death.id <- NULL
      for(j in 2:(dim.id)) # days between two consecutive admissions
      {
        time.id <- c(time.id, as.numeric(difftime(as.Date(dati.id[j,6]),
                                                  as.Date(dati.id[j-1,6]))))
        event.id <- c(event.id, 1) # Not censured events
        death.id <- c(death.id, 0)
      }
      if(dati.id[1,9]==1) # death
      {
        time.id <- c(time.id, as.numeric(difftime(as.Date(dati.id[1,8]),
                                                  as.Date(dati.id[dim.id,6]
                                                    ))))
        event.id <- c(event.id, 1) # Not censured event
        death.id <- c(death.id, 1)
      }else # end of the study time
      {
```

```
141         time.id <- c(time.id, as.numeric(difftime(as.Date("2010-12-31"),
142                                                   as.Date(dati.id[dim.id,6]
                                                      ))))
143         event.id <- c(event.id, 0) # Censured event
144         death.id <- c(death.id, 0)
145       }
146
147     }
148     t.start.id <- 0 # starting time of event
149     t.stop.id <- NULL # ending time of event
150     for(j in 1:(length(time.id)))
151     {
152       t.stop.id <- c(t.stop.id, time.id[j]+t.start.id[j])
153       t.start.id <- c(t.start.id, t.stop.id[j])
154     }
155     t.start <- c(t.start, t.start.id[1:dim.id])
156     t.stop <- c(t.stop, t.stop.id)
157     time <- c(time, time.id)
158     event <- c(event, event.id)
159     death <- c(death, death.id)
160   }
161 }
162
163 #building up the matrix
164 mat <- cbind(id.good,enum,t.start,t.stop,time,
165              event,death,age,shock,cabg,ptca,icd,stent)
166 mat <- as.data.frame(mat)
167
168 #removing patients with at least one shock event
169 mat <- mat[-which(mat$shock==1),]
170
171 groups <- read.table("Weibull_Gruppi.txt")[,1]
172 group <- NULL
173 for(i in 1:length(unique(mat$id.good)))
174 {
175   g <- groups[i]
176   n <- length(which(mat$id.good==unique(mat$id.good)[i]))
177   group <- c(group, rep(g,n))
178 }
179
180 category <- NULL #collapsing procedure variables in one.
181 for(i in 1:dim(mat)[1])
182 {
183   if(mat[i,11]+mat[i,12]+mat[i,14]>=1)
184   {
185     if(mat[i,13]>=1)
186     {
187       category <- c(category, "All")
188     }else{
189       category <- c(category, "Three")
190     }
191   }else{
192     if(mat[i,13]>=1)
193     {
194       category <- c(category, "Icd")
195     }else{
196       category <- c(category, "None")
197     }
198   }
199 }
200
201 age_c <- NULL # Age at each event
```

```
for(i in 1:length(unique(mat$id.good)))
{
  eta <- dati$age[which(dati$ID==unique(mat$id.good)[i])]
  age_c <- c(age_c,eta)
}
age_c <- age_c/100 #dividing for computational porpuses

mat <- cbind(mat,group,category,age_c)
mat$age <- mat$age/100 #dividing for computational porpuses

write.table(mat, "mat_time_noshock.txt")


#############################################################
###                      mixPHM code                      ###
###                                                       ###
###  Note: We will herein show how to use it              ###
###        and then conduct further analyses              ###
###        only for the first group.                      ###
###        To the others it applies the same way.         ###
#############################################################
library('mixPHM') # version 0.7.0

# requesting 3 groups and weibull-mixture model
weibullK <- phmclust(times5_mixPHM[,-1],
                     K=3,
                     method="separate",
                     EMoption="randomization")

write.table(weibullK$group, "Weibull_Gruppi.txt")

groups <- read.table("Weibull_Gruppi.txt")[,1]
gruppo1 <- which(groups==1)
gruppo2 <- which(groups==2)
gruppo3 <- which(groups==3)

#### Cluster 1 ####
death_cl1 <- death01[gruppo1]
count_cl1 <- count_col[gruppo1]
length(death_cl1)  # sample size
sum(death_cl1)
sum(death_cl1)/length(death_cl1)*100  # mortality rate

# number of patients having a maximum number of adms
length(which(count_cl1==1))
length(which(count_cl1==2))
length(which(count_cl1==3))
length(which(count_cl1==4))
length(which(count_cl1==5))

# number of patients dying at their h-th adm
length(which(death_cl1[which(count_cl1==1)]==1))
length(which(death_cl1[which(count_cl1==2)]==1))
length(which(death_cl1[which(count_cl1==3)]==1))
length(which(death_cl1[which(count_cl1==4)]==1))
length(which(death_cl1[which(count_cl1==5)]==1))

# age distribution
age_cl1 <- age[gruppo1]
minmax0 <- c(min(age_cl1[which(death_cl1==0)]),
             max(age_cl1[which(death_cl1==0)]))
minmax1 <- c(min(age_cl1[which(death_cl1==1)]),
```

```
264                max(age_cl1[which(death_cl1==1)]))
265 minmax1h <- c(min(age_cl1[which(count_cl1==1)]),
266                max(age_cl1[which(count_cl1==1)]))
267 minmax2h <- c(min(age_cl1[which(count_cl1==2)]),
268                max(age_cl1[which(count_cl1==2)]))
269 minmax3h <- c(min(age_cl1[which(count_cl1==3)]),
270                max(age_cl1[which(count_cl1==3)]))
271 minmax4h <- c(min(age_cl1[which(count_cl1==4)]),
272                max(age_cl1[which(count_cl1==4)]))
273 minmax5h <- c(min(age_cl1[which(count_cl1==5)]),
274                max(age_cl1[which(count_cl1==5)]))
275
276 mean(age_cl1)
277 sd(age_cl1)
278
279 # Sex distribution
280 sex_cl1 <- sex[gruppo1]
281 length(which(sex_cl1==1))
282 length(which(sex_cl1[which(death_cl1==1)]==1))
283 length(which(sex_cl1[which(death_cl1==1)]==2))
284 length(which(sex_cl1[which(count_cl1==1)]==1))
285 length(which(sex_cl1[which(count_cl1==1)]==2))
286 length(which(sex_cl1[which(count_cl1==2)]==1))
287 length(which(sex_cl1[which(count_cl1==2)]==2))
288 length(which(sex_cl1[which(count_cl1==3)]==1))
289 length(which(sex_cl1[which(count_cl1==3)]==2))
290 length(which(sex_cl1[which(count_cl1==4)]==1))
291 length(which(sex_cl1[which(count_cl1==4)]==2))
292 length(which(sex_cl1[which(count_cl1==5)]==1))
293 length(which(sex_cl1[which(count_cl1==5)]==2))
294
295 ##############################################
296 ###           frailtypack code             ###
297 ###                                        ###
298 ###  Note: We will herein show how         ###
299 ###        to use it only for the          ###
300 ###        first group. To the others      ###
301 ###        it applies the same way.        ###
302 ##############################################
303
304 times1 <- mat[which(mat$group==1),]
305 times2 <- mat[which(mat$group==2),]
306 times3 <- mat[which(mat$group==3),]
307
308 ### Cluster 1 ###
309 hazard.gruppo1 <- frailtyPenal( Surv(time, event) ~ cluster(id.good) +
310                                                     as.factor(category) +
311                                                     age_c,
312                                 data=times1,
313                                 Frailty=TRUE,
314                                 RandDist="LogN",
315                                 n.knots=14,
316                                 kappa1=100000)
317 hazard.gruppo1 #summary
318 # hazard reconstruction
319 reg <- hazard.gruppo1$linear.pred
320 lam0 <- hazard.gruppo1$lam[,1]
321 n <- hazard.gruppo1$groups
322 t <- hazard.gruppo1$x1
323 ids <- unique(times1$id.good[])
324 #patients on the rows
325 #time points on the columns
```

```
lam1 <- matrix(nrow=n, ncol=length(t))
for(i in 1:n)
{
  idx <- which(times1$id.good==ids[i])
  k <- 1
  for(j in 1:length(idx))
  {
    while(k<=length(t) &&
          times1$t.start[idx][j]<=t[k] &&
          t[k]<times1$t.stop[idx][j])
    {
      lam1[i,k] <- lam0[k] * exp(reg[idx][j])
      # lambda0(t) * exp(frailty_i + beta*X_i(t))
      k <- k+1
    }
  }
}
quartz()
matplot(t, lam1, type="l")

write.table(lam0, "Gruppo1_lambda0.txt")
write.table(t,"Gruppo1_timevector.txt")
write.table(lam1,"Gruppo1_lambdapiccola.txt")

######## Cumulative hazard: trapezoidal rule #########
Lam1 <- lam1[,1]
Lam0 <- lam0[1]
new <- (lam1[,1]+lam1[,1+1])*(t[1+1]-t[1])
new0 <- (lam0[1]+lam0[1+1])*(t[1+1]-t[1])
Lam1 <- cbind(Lam1, Lam1+new/2)
Lam0 <- c(Lam0, Lam0+new0/2)
for(i in 2:99)
{
  new <- (lam1[,i]+lam1[,i+1])*(t[i+1]-t[i])
  Lam1 <- cbind(Lam1,new/2 + Lam1[,i])
  new0 <- (lam0[i]+lam0[i+1])*(t[i+1]-t[i])
  Lam0 <- c(Lam0,new0/2 + Lam0[i])
}

quartz()
matplot(t,Lam1,type="l")
write.table(Lam1,"Gruppo1_lambdacumulata.txt")
write.table(Lam0,"Gruppo1_lambda0cumulata.txt")
```

Bibliography

[1] A. Araùjo and L. Meira-Machado. *smoothHR: Smooth Hazard Ratio Curves taking a Reference Value.* R package version 1.0. 2013.

[2] A. Baddeley and R. Turner. "Spatstat: an R package for analyzing spatial point patterns." In: *Journal of Statistical Software* 12.6 (2005), pp. 1–42.

[3] S. Baraldo et al. "Outcome prediction for heart failure telemonitoring via generalized linear models with functional covariates." In: *Scandinavian Journal of Statistics: Theory and Applications* 40.3 (Sept. 2013), pp. 403–416.

[4] R. E. Beard. "Appendix: Note on some mathematical mortality models." In: *Ciba Foundation Colloquia on Aging: The Lifespan of Animals.* 5 (1959). Ed. by G. E. W. Wolstenholme and M. O'Connor, pp. 802–811.

[5] N. E. Breslow. "Analysis of Survival Data under the Proportional Hazards Model." In: *International Statistical Review / Revue Internationale de Statistique* 43.1 (Apr. 1975), pp. 45–57.

[6] Wikipedia contributors. *Proportional hazards model.* 2014. URL: `http://en.wikipedia.org/w/index.php?title=Proportional_hazards_model&oldid=594336292`.

[7] D. R. Cox. "Regression Models and Life-Tables." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–220.

[8] L. Duchateau and P. Janssen. *The Frailty Model.* Statistics for Biology and Health. Springer, 2008.

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

[10] J. He et al. "Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study." In: *Archives of Internal Medicine* 161.7 (2001), pp. 996–1002.

[11] P. Hougaard. *Analysis of Multivariate Survival Data.* Statistics for Biology and Health. Springer, 2000.

[12] AHRQ Quality Indicators. *Guide to Inpatient Quality Indicators: Quality of Care in Hospitals - Volume, Mortality, and Utilization.* Version 3.1. Department of Health, Human Services Agency for Healthcare Research, and Quality, 2007.

[13] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Sixth Edition. Pearson International Edition, 2007.

[14] E. L. Kaplan and P. Meier. "Nonparametric Estimation from Incomplete Observations." In: *Journal of the American Statistical Association* 53.282 (June 1958), pp. 457–481.

[15] H. M. Krumholz et al. "Predictors of readmission among elderly survivors of admission with heart failure." In: *American Heart Journal* 139.1 (Jan. 2000), pp. 72–77.

[16] P. Mair and M. Hudec. *mixPHM: Mixtures of proportional hazard models.* R package version 0.7.0. 2008.

[17] P. Mair and M. Hudec. "Multivariate Weibull mixtures with proportional hazard restrictions for dwell-time-based session clustering with incomplete data." In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 58.5 (Nov. 2009), pp. 619–639.

[18] V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* English. Eamon Dolan/Houghton Mifflin Harcourt, Mar. 2013.

[19] M. Munda, F. Rotolo, and C. Legrand. *parfm: Parametric Frailty Models.* R package version 2.5.3. 2013.

[20] M. Munda, F. Rotolo, and C. Legrand. "parfm: Parametric Frailty Models in R.". In: *Journal of Statistical Software* 51.11 (2012), pp. 1–20.

[21] M. Packer. "Pathophysiology of chronic heart failure." In: *Lancet* 340 (1992), pp. 88–92.

[22] J. T. Parissis, L. Mantziari, N. Kaldoglou, et al. "Gender-related differences in patients with acute heart failure: Management and predictors of in-hospital mortality." In: *International Journal of Cardiology* 168.1 (Sept. 2013), pp. 185–189.

[23] E. A. Peña and M. Hollander. "Mathematical Reliability: An Expository Perspective." In: Kluwer Academic Publishers, 2004. Chap. 6:

Models for Recurrent Events in Reliability and Survival Analysis. Pp. 105–123.

[24] G. C. Pope et al. "Evaluation of the CMS-HCC Risk Adjustment Model. Final Report." In: *RTI International* (2011).

[25] G. C. Pope et al. "Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model." In: *Health Care Financing Review* 25 (2004), pp. 119–141.

[26] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2012.

[27] N. Reid. "A Conversation with Sir David Cox." In: *Statistical Science* 9.3 (Aug. 1994), pp. 439–455.

[28] V. Rondeau, Y. Mazroui, and J. R. Gonzalez. "Frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametric estimation." In: *Journal of Statistical Software* 47 (2012), pp. 1–28.

[29] V. Rondeau, Y. Mazroui, and J. R. Gonzalez. *frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation.* R package version 2.5. 2012.

[30] T. M. Therneau. *A package for survival analysis in S.*. R package version 2.37-4. 2013.

[31] T. M. Therneau and P. M. Grambsch. *Modelling Survival Data: Extending the Cox Model.* Statistics for Biology and Health. Springer, 2000.