POLITECNICO DI MILANO
School of Industrial Engineering and Information
Master's Degree in Mathematical Engineering

Applications of nonparametric frailty models for the analysis of long term survival in Heart Failure patients

Supervisor: Prof. Francesca IEVA
Co-supervisor: Dott. Francesca GASPERONI

Candidate: Michela AGOSTI
Personal ID: 862944

Academic Year 2016-2017

A mia mamma
e a mio papà...

Contents

List of Figures

List of Tables

Abstract

Heart Failure (HF) is one of most common disease in our society and one of the most important cause of hospitalisation in people over 65. When dealing with patients affected by chronic disease (like HF), the matter of predicting readmissions is a real challenge for hospitals. Finding which patients features determine a higher incidence of readmission can help to improve therapies and to target interventions.
Thanks to the large amount of data collected by Trieste hospital, we will try to understand which factors affect the admission to hospital, discharge from hospital and death of HF patients, through the implementation of multi-state models. Secondly, thanks to the data collected by Friuli Venezia Giulia hospitals, we will analyze the homogeneity of the hospital treatment in the different residence districts. In order to do this, we will apply a nonparametric and discrete frailty Cox model that, thanks to the discrete distributed frailty, allow us to detect if a possible clustering structure can be found among the residence districts. Moreover, we will analyze the homogeneity of the regional cohort and we will evaluate, through the k-means algorithm, the presence of a clustering structure. Once identified the clusters we will analyze them in order to find the features that best characterize them.

Keywords: Heart Failure, multi-state model, nonparametric frailty Cox model

Sommario

Lo scompenso cardiaco è una delle malattie più comuni nella nostra società e una delle più importanti cause di ospedalizzazione nelle persone sopra i 65 anni. Quando si ha a che fare con pazienti affetti da malattie croniche (come lo scompenso cardiaco), prevedere le riammissioni è una vera sfida per gli ospedali. Trovare quali caratteristiche dei pazienti determinano una maggiore probabilità di riammissione può, infatti, aiutare a migliorare le terapie e indirizzare gli interventi.
Grazie alla grande quantità di dati raccolti dall'ospedale di Trieste, cercheremo di capire quali fattori influenzano l'ammissione in ospedale, la dimissione dall'ospedale e la morte dei pazienti con scompenso cardiaco, attraverso l'implementazione di modelli multi-stato. Successivamente, grazie ai dati raccolti dagli ospedali del Friuli Venezia Giulia, analizzeremo l'omogeneità del trattamento ospedaliero nei diversi distretti di residenza. Per fare ciò implementeremo un modello di tipo Cox con frailty non parametrica e discreta, che consentirà di capire se è possibile trovare una struttura di clustering tra i distretti di residenza. Inoltre, analizzeremo l'omogeneità della coorte regionale e valuteremo, attraverso l'algoritmo k-means, l'esistenza di una struttura di clustering. Una volta identificati i cluster, li analizzeremo per trovare gli aspetti che meglio li caratterizzano.

Parole chiave: scompenso cardiaco, modello multi-stato, modello Cox con frailty non parametrica

Introduction

Heart Failure (HF) is a chronic disease that occurs when the heart fails to pump sufficiently to maintain blood flow at the right pressure for human needs. It may be caused by many conditions that lead damage to the heart muscle: coronary artery disease, high blood pressure, heart muscle weakness, heart rhythm disturbance, damage with heart's valves or a combination of all these.

Nowadays, HF is one of most common disease in our society, due to many causes, for example population ageing. To understand the relevance of this disease, we just point out that HF is one of the most important cause of hospitalisation in people over 65 in Italy [18].

When dealing with patients affected by chronic disease (like HF), the matter of predicting readmissions is a real challenge for hospitals, mainly for two reasons. The first one is concerned with the high costs of hospitalization, so, discovering the reasons of readmission may lead to improve hospital care and, consequently, to save money. Much more important could be the second reason: to find which features in patients determine a higher incidence of readmission, in order to improve therapies and to target interventions. This is twice as useful, as for it takes benefits to patients and to hospitals as well. Evaluating hospital readmissions and linked quantities for any kind of chronic disease is one of the aims of the healthcare research, thanks to the large amount of data collected by hospitals.

In particular, this is the aim of the first part of this thesis: analyzing the dataset whose informations are collected in the Trieste area, we will try to understand which factors influence the admission, discharge and death dinamic of HF patients, through the implementation of musti-state models.

Instead, in the second part of this thesis, exploring the dataset whose informations are collected in all the Friuli Venezia Giulia region, we will analyze the impact of patients risk profiles and geographical residential effects on the survival of HF patients. To this aim, we will apply Cox models with nonparametric and discrete frailty in order to understand if the different residence districts have some measurable influence on the patients survival.

The thesis is structured as follows:

- In Chapter 1 we will describe the main theoretical features of survival analysis and of multi-state models.
- In Chapter 2 we will discuss the main aspects of the frailty Cox models focusing on a specific proposal based on nonparametric discrete frailty [14]. An important advantage of this model is that, through the discrete frailty, it is possible to build a probabilistique clustering tecnique. We will describe the model and the tailored Expectation-Maximization algorithm used to estimate the model parameters.

Once described the necessary theory, we will present the results achieved from the analysis of two different dataset: the Trieste and the Friuli Venezia Giulia one.

- In Chapter 3 we will analyze the Trieste dataset. After a descriptive analysis of the cohort considered, we will implement a multi-state model in order to understand which factors most influence the admission, discharge and death dinamic.
- In Chapter 4 we will analyze the Friuli Venezia Giulia dataset. Firstly, through the implementation of the nonparametric discrete frailty Cox model, we will discuss the residence districts homogeneity. Secondly, through the implementation of different Cox models and through the application of the k-means algorithm, we will discuss the cohort homogeneity. Finally, we will perform a multi-state modelling of the data in order to see if the results are coherent with the ones obtained with the smaller dataset, the Trieste one.
- Finally, in Chapter 5, we will conclude with a summary of the results obtained and with few proposal for future developments.

All the analysis are carried out with R [25].

Chapter 1

Multi-state models

In this chapter we will introduce survival analysis and multi-state models as statistical methodologies to deal with time to event data.
After a brief introduction of basic notions in Section 1.1, Section 1.2 will report the main concepts of survival analysis, where the time occured between a starting event and a final event of interest is studied.
In Section 1.3, the concepts of Section 1.2 will be generalized to the multi-state models setting, where the time occured between several events of interest is studied.

The concepts explained in this theoretical chapter will be put into practice in Chapter 3, where multi-state models will be applied to a real dataset arising from healthcare context.

1.1 Introduction

Survival and event history analysis [1] are statistical methodologies used in many different settings where the occurrence of events is studied. By events we mean occurrences in the lives of individuals that are of interest in scientific studies, in medicine, demography, biology, sociology, econometrics, etc. The main objective is trying to understand their cause or establish risk factors which may affect their occurrence.

In classical survival analysis we focus on a single event of interest for each individual, describing its occurrence by means of survival curves and

hazard rates (see Section 1.2 for details) and analyzing the dependence on covariates by means of regression models.

When an individual has the possibility to experience several different events we talk about event histories.

1.2 Survival analysis

We start by considering classical survival analysis, which focuses on the time elapsed from a starting event to an event, or endpoint, of interest. Some examples may be:

- time from birth to death;
- time from disease onset to death or relapse;
- time from birth to disease diagnosis.

The time between the starting event and the event of interest is defined as *survival time*, even when death is not the final state. On the other hand, the events definingh such times are called *states*.

Figure 1.1 shows a sketch of a survival model, reported in [4]. The blocks, named *Alive* and *Dead*, represent the states. The arrow represents the transition between states. $\alpha(t)$ is defined rate of transition.

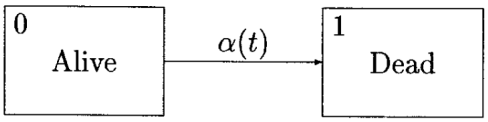


Figure 1.1: Example of survival scheme.

1.2.1 Right censoring

Time to event data are often only partially observed. The event of interest can occur for some individuals but not for the others. In this latter case we could know that the event is not occurred up to a certain time, without

knowing when and if it will happen. That is called *right censoring*. Alternatively, we could know that the event is occurred before a certain time, without knowing the exact time in which it occurred. That is called *left censoring*.

The most common case is that of right censored data. For each patient i, let T_i^* be the non negative random variable denoting the time at which the event occurs and C_i be the random variable denoting the time at which censoring happens. What we observe is the *failure time* T_i, that is the event time T_i^* or the censoring time C_i, whichever is smaller.

$$T_i = min(T_i^*, C_i). \tag{1.1}$$

In addition, through an indicator random variable δ_i, we can get the information on whether T_i is an event time or a censoring time.

$$\delta_i = \begin{cases} 1 & \text{if } T_i^* \leq C_i \\ 0 & \text{if } T_i^* > C_i \end{cases} \tag{1.2}$$

Hence the observed data consist of pairs (T_i, δ_i) for each patient i.

An illustration of how censored survival times may arise is given in Figure 1.2. The figure illustrates a hypothetical clinical study where 10 patients are observed over a time period to see whether some specific event occurs. Observations are shown as they occur in calendar time. The patients enter the study at different times and, then, they are followed until the event occurs or until the closure of the study, after 10 years.

1.2.2 Survival function and hazard rate function

There are two basic concepts that pervade the whole theory of survival analysis, namely the *survival function* and the *hazard rate*.

At time zero a set of individuals waits for an event that might happen. The observation for a given individual consists of a random variable, say T, representing the time from a given origin to the occurrence of the event.
The distribution of T is characterized by the *probability distribution function*:

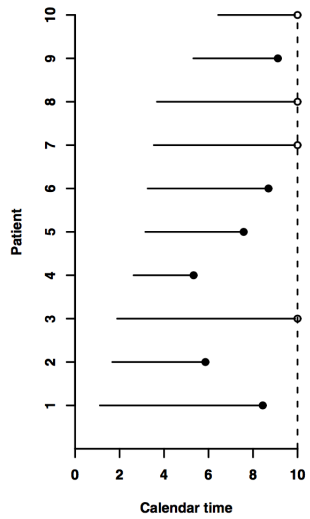


Figure 1.2: Example of right censored data.

$$F(t) = P(T \leq t), \tag{1.3}$$

or, equivalently, by the *survival distribution function*, defined as:

$$S(t) = 1 - F(t) = P(T > t), \tag{1.4}$$

The survival function gives the probability that the event of interest has not happened by time t.
The survival function is a non-increasing function that often tends to zero as t increases. However, since we use the terms *survival time* and *survival function* also for events that do not necessarily happen to all individuals, the survival function could also tend to a positive value as t goes to infinity.

The random variable T could be a discrete or an absolutely continuous variable. In this latter case T has a probability density and, consequently, by means of a conditional probability, we can define the hazard rate $\alpha(t)$. The hazard rate represents the istantaneous probability to pass from the starting state to the final state. If we look at those individuals who have not yet experienced the event of interest by time t and consider the probability of experiencing the event in the small time interval $[t, t + dt)$, then this probability equals $\alpha(t)dt$. To be more precise, the hazard rate is defined as

a limit in the following way:

$$\alpha(t) = \lim_{\Delta(t)\to 0} \frac{P(T \leq t + \Delta(t)|T \geq t)}{\Delta(t)}. \tag{1.5}$$

Notice that while the survival curve is a function that starts at 1 and declines over time, the hazard rate can be essentially any non negative function.
In Figure 1.3 we can find an example of hazard rate and one of survival function. The left-hand panel shows a typical hazard rate, reaching a maximum and then declining. The right-hand panel shows a survival curve.

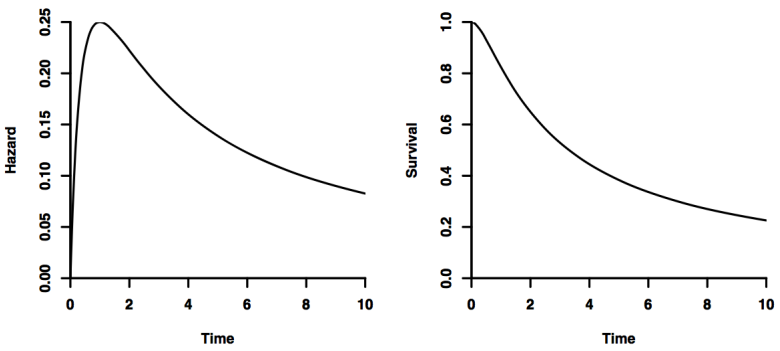


Figure 1.3: Examples of hazard rate (left panel) and survival function (right panel).

Starting from the hazard rate function $\alpha(t)$ and integrating it over the time variable t, the *cumulative hazard function* can be obtained:

$$A(t) = \int_0^t \alpha(s)ds. \tag{1.6}$$

There is an important relation between the survival function and the cumulative hazard function:

$$S(t) = exp(-A(t)). \tag{1.7}$$

1.2.3 Kaplan-Meier estimator

To estimate the survival function from a sample of survival data, the Kaplan-Meier estimator [1] is the most popular method.

In order to estimate the survival functon, we consider a sample of n individuals from the population. $N(t)$ is the variable that counts the number of occurrences of the event in $[0, t]$, while $Y(t)$ is the number of individuals at risk "just before" time t. We write $T_1 < T_2 < ...$ for the ordered times when an occurrence of the event is observed, that is, for the jump times of N.

To give an intuitive justification of the Kaplan-Meier estimator, we partition the time interval $[0, t]$ into a number of small time intervals $0 = t_0 < t_1 < ... < t_K = t$ and use the multiplication rule for conditional probabilities to write:

$$S(t) = \prod_{k=1}^{K} S(t_k|t_{k-1}) \tag{1.8}$$

where

$$S(v|u) = \frac{S(v)}{S(u)}, \text{ for } v > u,$$

is the conditional probability that the event will occur later than time v given that it has not yet occurred by time u.

We assume that there are no tied event times. This assumption is reasonable since we may make each time interval so small that it contains at most one observed event, and that all censorings occur at the right- hand endpoint of an interval.

Then, if no event is observed in $(t_k - 1, t_k]$, we estimate $S(t_k|t_k - 1)$ with 1, otherwise, if an event is observed at time $T_j \in (t_k - 1, t_k]$, the natural estimate of $S(t_k|t_k - 1)$ is

$$1 - \frac{1}{Y(t_k - 1)} = 1 - \frac{1}{Y(T_j)}.$$

Inserting these estimates into (1.8), we obtain:

$$\hat{S}(t) = \prod_{T_j \leq t} 1 - \frac{1}{Y(T_j)} \tag{1.9}$$

which is the Kaplan-Meier estimator.

In Figure 1.4, an example of Kaplan-Meier estimates from [1] is reported. In order to be able to correctly interpret the figure, we underline that on the x-axis there is the time of the study, that could be expressed in hours, days, months or years, like in this case, while on the y-axis there is the probability of survival.

In this example we are interest in the time between the first and second births of a woman (by the same father), and how this is affected if the first child dies within one year of his birth.
From the survival curves we see, for example, that it takes less than two years before 50% of the women who lost their first child will have another one, while it takes about four years before 50% of the women who do not experience this traumatic event will have another one. We note that the survival curves give a very clear picture of the differences between the two group of women.

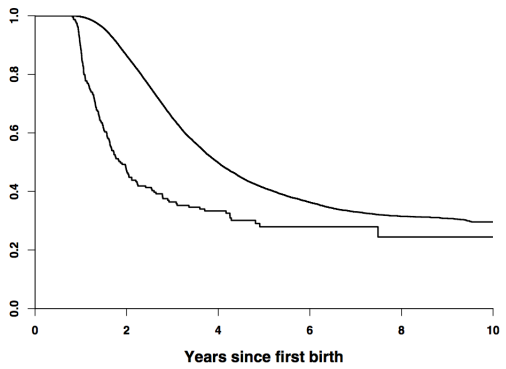


Figure 1.4: Kaplan-Meier estimates for the time between first and second birth of a woman (by the same father), and how this is affected if the first child dies within one year of its birth. Upper curve: first child survived one year; lower curve: first child died within one year.

In this section we explained the main concepts of the survival analysis. These will be used to deal, in Section 1.3, with the multi-state models.

1.3 Multi-state models

In classical survival analysis we focus on the time to the occurrence of a single event for each individual. This may, however, be too simplistic in a number of situations. Sometimes more than one type of event is of interest. Such situations may conveniently be described by *multi-state models* (see among others [4], [2], [15]), where the number of *states* could be any possible finite integer number.

Multi-state models are models for time-to-event data in which all individuals start in one state and, after having transitated through different states, possibly more than once, end up in a final state.
Theoretically, the transition process ends up when it reaches an absorbing state, like *Dead* state in Figure 1.5. However, often, truncation does not allow the process to reach this equilibrium state. In this case every state could be a final state.

Multi-state models well describe the development of longitudinal failure time data; for this reason they are frequently used in medicine, especially in chronic diseases, where the states can be used to describe the patients's conditions over time.

Graphically, multi-state models may be illustrated using diagrams with boxes representing the states and with arrows between the states representing the possible transitions.
An example of a multi-state model, the illness-death model, reported in [4], is given in Figure 1.5.

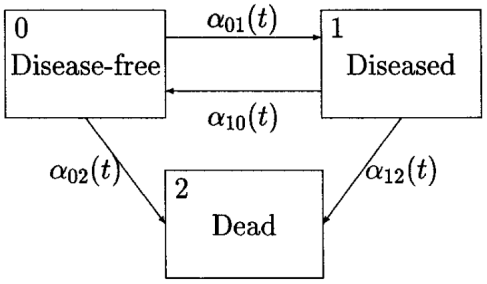


Figure 1.5: Example of *illness-death model* with recovery.

1.3.1 Notation

Hereafter we report some notations, generalizing the one previously introduced in Section 1.2.

For every individuals i, let $X_i(t)$, $t \in T = [0, \tau]$, be a random variable denoting the state occupied by individual i at time t. The possible states are the ones in the state space $S = \{1, ..., r\}$.
The process has initial distribution

$$\pi_h(0) = P(X(0) = h). \tag{1.10}$$

The *transition probability* from the state h to the state j is:

$$P_{h,j}(s,t) = P(X(t) = j | X(s) = h) \tag{1.11}$$

with $h, j \in S, s, t \in T$.
The *transition intensity*, or *hazard rate*, expressed as

$$\alpha_{h,j}(t) = \lim_{\Delta(t) \to 0} \frac{P_{h,j}(t, t + \Delta(t))}{\Delta(t)} \tag{1.12}$$

with $h, j \in S, t \in T$, is the istantaneous probability for the transition from state h to state j to happen.

If $\alpha_{h,j}(t)$ only depends on the history via the state $h = X(t)$ occupied at t then it is said that the process is Markovian: future evolution depends on the current state and time, but not on the whole history of the process.

Also with multi-state models, integrating the hazard rate $\alpha_{h,j}(t)$ we can obtain the *cumulative hazard rate*:

$$A_{h,j}(t) = \int_0^t \alpha_{h,j}(s) ds, \tag{1.13}$$

where $h, j \in S, t \in T$.

1.3.2 Cox regression models

A peculiarity of Multi-State models lies in the possibility of introducing the role of covariates in the transition intensities, that hence may depend on time t and on a set of individual variables $z_1(t), ..., z_p(t)$ as well, where p is the number of covariates.
When covariates are introduced we talk about regression models.

A big family of regression models is the *relative risk regression models*, which have the following general shape:

$$\alpha(t) = \alpha_0(t)\, r(\boldsymbol{\beta}, \boldsymbol{z}_i(t)) \tag{1.14}$$

where $r(\boldsymbol{\beta}, \boldsymbol{z}_i(t))$ is a relative risk function, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a vector of regression coefficients describing the effect of the covariates, and $h_0(t)$ is a baseline hazard rate.

A tipical regression model is the *Cox regression model* for censored survival data ([3], [26]), that is obtained when $r(\boldsymbol{\beta}, \boldsymbol{z}_i(t))$ is equal to $exp\{\boldsymbol{\beta}^T \boldsymbol{z}_i(t)\}$. The Cox regression model specifies that covariates have a proportional effect on hazard function of the life-time distribution of an individual.
The transition intensity for patient i at observation time t take the following form:

$$\alpha(t) = \alpha_0(t) exp(\beta_1 z_1 + ... + \beta_p z_p) \tag{1.15}$$

where β is a p-vector of unknown regression coefficients, $\alpha_0(t)$ is the baseline hazard and $exp(\beta_1 z_1 + ... + \beta_p z_p)$ is the hazard ratio. The objective is estimating $\boldsymbol{\beta}$ and α_0.

Frailty models

Cox models, through the inclusion of the covariates, can explain the variability among statistical units. However, in certain situations, in spite of the introduction of covariates, a certain residual variability remains unexpressed. To take into account the effects of this unobserved or unobservable heterogeneity, *frailty models* are used.
In frailty models a variable w, said *frailty* term, is introduced. It acts as a factor on the hazard function that hence, conditionally on the frailty, can be written as:

$$\alpha(t|w) = w \cdot \alpha(t). \tag{1.16}$$

The idea is to suppose that different patients have different frailties and patients more "frail" tend to experience the event of interest earlier than those who are less frail.

The description of these model will be the subject of Chapter 2.

Chapter 2

Nonparametric frailty Cox models

"Individuals differ" has been written by Aalen et al. in [1]. This very simple statement represents the leading idea of the frailty models. Indeed, it is common knowledge that the same therapy can lead to different results for different people, and that some people, despite similar conditions, die before others or survive longer.
This variability is not only recorded in medical field, but also in the wide contest of biology, economy and technology.
One aim of the survival analysis is to build models that are able to justify this variability among statistical units. However, there is a part of variability which cannot be explained by the covariates. Sometimes, not only a residual error term is present but also an overdispersion due to the grouped nature of data. In some situations this term is large and not negligible, but it can be taken into account despite its unobservability. We usually refer to unobserved heterogeneity that can be accounted for into the model as *frailty*. This term highlights that some people are more frail than others and that the event of interest is more likely to happen for them.

In Section 2.1 we will introduce the Cox proportional hazard frailty model. In Section 2.2 we will focus on Cox model with a non parametric frailty and we will present a tailored Expectation Maximization algorithm to estimate the parameters of this model.

2.1 Cox proportional hazard frailty model

The term "frailty theory" has been primarily associated with one particular mathematical formulation of frailty, the proportional frailty model [1], where we assume that the hazard rate of an individual is given as the product of an individual specific quantity w and a basic rate $\alpha(t)$:

$$\alpha(t|w) = w \cdot \alpha(t). \tag{2.1}$$

In case of Cox proportional frailty model, the hazard rate expression becomes:

$$\alpha(t|w) = w \cdot \alpha_0(t) exp(\boldsymbol{\beta}^T \boldsymbol{x}), \tag{2.2}$$

where w is the frailty term, $\alpha_0(t)$ is the baseline hazard function, $\boldsymbol{x}$ is the vector of covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients.
If the random component w is higher than 1, the subject has an higher risk, on the contrary, if w is less than 1, he is exposed to lower risk.

In general the heterogeneity is subject-specific, in this case the frailty model is known as *univariate frailty model*. The hazard rate for subject i then becomes:

$$\alpha_i(t|w_i) = w_i \cdot \alpha_0(t) exp(\boldsymbol{\beta}^T \boldsymbol{x_i}) \quad \forall i \in 1, ..., n, \tag{2.3}$$

where n is the total number of subjects in the study.
The univariate frailty model is a particular case of *shared frailty model*, where some subgroups, called *clusters*, are recognized among the population studied. Subjects of the same group have common unobserved risk factors. The hazard rate for subject $i = 1, ..., n$ in group $j = 1, ...K$ is:

$$\alpha_{ij}(t|w_j) = w_j \cdot \alpha_0(t) exp(\boldsymbol{\beta}^T \boldsymbol{x_i}). \tag{2.4}$$

A keypoint in the definition of the frailty model (2.3) and (2.4) is the choice of the distribution of w. The usual choice is the parametric one, in particular the most common frailty distributions are Gamma, Lognormal, Inverse Normal and Positive stable [16]. An advantage of the parametric choice is that only few parameters have to be taken into account. Moreover, due to the availability of different software, a parametric distribution can be advantageous also from a computational point of view.
However, due to the potential misspecification of the parametric form, since

the frailty distribution has to be chosen a priori and sometimes no prior informations are available, a nonparametric frailty distribution, with its good level of flexibility, is desiderable. In fact, in Section 2.2 we will present an extension of the shared frailty Cox model for hierarchical time-to-event data proposed in [14], in which a nonparametric frailty is included.

Another interesting aspect of the nonparametric approac is that, considering a discrete distribution, it is possible to build a probabilistic clustering technique. In fact, since the frailty values are extrected from a discrete distribution with finite support, subjects with the same frailty can be grouped in a unique cluster.

2.2 Cox model with a nonparametric frailty

In this section we will present the Cox model with a nonparametric and discrete frailty (npdf Cox) described in [14]. Firstly we will describe the main concepts and the main variables of this new model. In Section 2.2.1 we will report the tailored Expectation-Maximation algorithm used to estimate the model parameters, while in Section 2.2.2 we will describe how the parameters standard errors can be computed. Finally, in Section 2.2.3 we will describe how to compute model selection and hence how to estimate the correct number of latent populations.

In order to implement the npdf Cox model, we consider a random sample where each statistical unit i, $i = 1, ..., n$ belongs to one group j, $j = 1, .., J$. We define:

- T^*_{ij} the survival time,
- C_{ij} the censoring time,
- $T_{ij} = min(T^*_{ij}, C_{ij})$,
- $\delta_{ij} = \mathbb{1}_{T^*_{ij} \leq C_{ij}}$,
- $\boldsymbol{X}_{ij}$=$(X_{ij,1}, ..., X_{ij,p})^T$ the vector of covariates,
- $\boldsymbol{w}$ the vector of shared frailties.

According to [14], the non parametric frailty term is modelled through a random variable with discrete distribution with an unknown number of point in the support.
We assume that each group j can belong to one latent population $k = 1, ..., K$ with probability π_k, $\mathbb{P}(w = w_k) = \pi_k$, hence, $w_1, ..., w_K$ are the support points of w. z_{jk} is an auxiliary random variable that is equal to 1 if the j-th group belongs to the k-th latent population, hence $z_{jk} \overset{i.i.d.}{\sim} Be(\pi_k)$. In other words, the prior probability that the j-th group belongs to the k-th latent population, hence that $z_{jk} = 1$, equals π_k.
The hazard function for subject i in group j is then:

$$\alpha(t; \boldsymbol{X}_{ij}, w_k, z_{jk}) = \prod_{k=1}^{K} [\alpha_0(t)\ w_k\ exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta})]^{z_{jk}}, \tag{2.5}$$

where w_k is the frailty term shared among groups of the same latent population k.
For each subject i in group j the observable and "incomplete" data are $\boldsymbol{Y}_{ij} = \{T_{ij}, \delta_{ij}, \boldsymbol{X}_{ij}\}$, while the "complete" data, that are needed to write the log-likelihood of the model, are $\{T_{ij}, \delta_{ij}, \boldsymbol{X}_{ij}, w_k, z_{jk}\}$.
We assume that censoring is noninformative, hence T_{ij}^* and C_{ij} are conditionally independent given $\boldsymbol{X}_{ij}, w_k, z_{jk}$.

If we denote with $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{w}, \lambda_0(t), \boldsymbol{\beta})$ the vector of parameters to be estimated, the full likelihood of the model can be explicitely written as:

$$L_{full}(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{z}) = \prod_{k=1}^{K} \prod_{j=1}^{J}\ \pi_k^{z_{jk}} \cdot L_{full}^{jk}(\boldsymbol{\theta}; \boldsymbol{Y}_j|\boldsymbol{z}) \tag{2.6}$$

where

$$L_{full}^{jk}(\boldsymbol{\theta}; \boldsymbol{Y}_j|\boldsymbol{z}) = \prod_{i=1}^{n_j} \{[\alpha_0(t_{ij})\ w_k\ exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta})]^{\delta_{ij}} \cdot exp[-A_0(t_{ij})\ w_k\ exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta})]\}^{z_{jk}} \tag{2.7}$$

and $A_0(t) = \int_0^t \alpha_0(s)ds$ is the cumulative baseline hazard function.
The number of latent populations K can be considered as an unknown parameter, the relative hazard between two individuals with the same covariate values but from different latent population k and $\tilde{k}$ can be described by the frailty ratio $\frac{w_k}{w_{\tilde{k}}}$. This parameter is evaluated in order to define the effect of different groups.

2.2.1 Tailored Expectation-Maximization (EM) algorithm

In order to estimate $\boldsymbol{\theta}$ for a given K we report an EM algorithm proposed in [14]. The algorithm iterates between two steps, Expectation and Maximization, and is guaranteed to converge to a stationary point, under regularity conditions [7], [10], [22], [27].

E-step

The Expectation step consist of computing the expectation over $\boldsymbol{z}$ of the full log-likelihood, given the current values of parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\pi}}, \hat{\alpha}_0(t), \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{w}})$:

$$Q(\boldsymbol{\theta}) = E_{\boldsymbol{z}|\hat{\boldsymbol{\theta}}}[l_{full}(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{z})] \tag{2.8}$$

The full log-likelihood can be decomposed into two parts, one (2.9) depending on $\boldsymbol{\pi}$ and one (2.10) depending on $\alpha_0(t), \boldsymbol{\beta}, \boldsymbol{w}$.

$$l_{full,1}(\boldsymbol{\pi}; \boldsymbol{Y}|\boldsymbol{z}) = \sum_{k=1}^{K}\sum_{j=1}^{J} z_{jk} \cdot log(\pi_k). \tag{2.9}$$

$$\begin{aligned} l_{full,2}(\alpha_0(t), \boldsymbol{\beta}, \boldsymbol{w}; \boldsymbol{Y}|\boldsymbol{z}) = &\sum_{k=1}^{K}\sum_{j=1}^{J} z_{jk} \cdot \sum_{i=1}^{n_j} \delta_{ij}[log(\alpha_0(t_{ij})) + log(w_k) + \boldsymbol{X}_{ij}^T\boldsymbol{\beta}] - \\ &A_0(t_{ij}) w_k exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}). \end{aligned} \tag{2.10}$$

Hence the (2.8) becomes:

$$Q(\boldsymbol{\theta}) = E_{\boldsymbol{z}|\hat{\boldsymbol{\theta}}}[l_{full}(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{z})] = E_{\boldsymbol{z}|\hat{\boldsymbol{\theta}}}[l_{full,1}(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{z})] + E_{\boldsymbol{z}|\hat{\boldsymbol{\theta}}}[l_{full,2}(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{z})]. \tag{2.11}$$

The (2.11) can be reduced to the computation of $E[z_{jk}|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}]$, which can be derived in closed form using Bayes' theorem:

$$E[z_{jk}|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}] = \frac{\pi_k \cdot exp\{\sum_{i=1}^{n_j} \delta_{ij} \cdot log(w_k) - A_0(t_{ij}) w_k exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta})\}}{\sum_{r\in\{1:K\}} \pi_r \cdot exp\{\sum_{i=1}^{n_j} \delta_{ij} \cdot log(w_r) - A_0(t_{ij}) w_r exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta})\}}. \tag{2.12}$$

For simplicity we will write $a_{jk} = E[z_{jk}|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}]$.

M-step

The Maximization step consists of maximizing $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. As we can see from (2.11), in order to maximize $Q(\boldsymbol{\theta})$, we can maximize $Q_1(\boldsymbol{\pi}) := E_{\boldsymbol{z}|\hat{\boldsymbol{\theta}}}[l_{full,1}|\hat{\boldsymbol{\theta}}]$ with respect to $\boldsymbol{\pi}$ and $Q_2(\alpha_0(t), \boldsymbol{\beta}, \boldsymbol{w}) := E_{\boldsymbol{z}|\hat{\boldsymbol{\theta}}}[l_{full,2}|\hat{\boldsymbol{\theta}}]$ with respect to $\alpha_0(t), \boldsymbol{\beta}, \boldsymbol{w}$ separately.

The maximization of $Q_1(\boldsymbol{\pi})$ is a constrained optimization problem and can be solved by applying the Lagrange multipliers technique, getting:

$$\hat{\pi}_k = \frac{1}{J} \sum_{j=1}^{J} \alpha_{jk}. \tag{2.13}$$

Adapting a profile log-likelihood approach for the estimation of the shared parametric frailty Cox model [20], $\boldsymbol{w}$ can be estimated fixing α_0 and $\boldsymbol{\beta}$. The resulting estimate is:

$$\hat{w}_k = \frac{\sum_{j=1}^{J} \alpha_{jk} \sum_{i=1}^{n_j} \delta_{ij}}{\sum_{j=1}^{J} \alpha_{jk} \sum_{i=1}^{n_j} \{A_0(t_{ij}) \cdot exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta})\}}. \tag{2.14}$$

By substituting these estimates in Q_2 we obtain:

$$\begin{aligned} Q_2(\alpha_0(t), \boldsymbol{\beta}, \hat{\boldsymbol{w}}) = & \sum_{k=1}^{K} \sum_{j=1}^{J} \alpha_{jk} \cdot \sum_{i=1}^{n_j} \delta_{ij} [log(\alpha_0(t_{ij})) + log(\hat{w}_k) + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}] - \\ & A_0(t_{ij}) \hat{w}_k exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}). \end{aligned} \tag{2.15}$$

With arguments similar to the ones used in [20], it is possible to show that the estimate of the cumulative baseline that maximize (2.15) is:

$$\hat{A}_0(t_{ij}) = \sum_{(fg):t_{fg} \leq t_{ij}} \frac{d_{fg}}{\sum_{rs \in R(t_{fg})} (\sum_{k=1}^{K} \alpha_{sk} \hat{w}_k) exp(\boldsymbol{X}_{rs}^T \boldsymbol{\beta})}, \tag{2.16}$$

where d_{fg} is the total number of events happening at time t_{fg} and $R(t_{fg})$ represents the set of subjects who are at risk at time t_{fg}, which is the event time of subject f in cluster g.
Including (2.16) in $Q_2(\alpha_0(t), \boldsymbol{\beta}, \boldsymbol{w})$ we obtain the profile log-likelihood as a function of only $\boldsymbol{\beta}$:

$$l_{profile}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \delta_{ij} \Big[\boldsymbol{X}_{ij}^T \boldsymbol{\beta} - log \sum_{rs \in R(t_{ij})} \Big(\sum_{k=1}^{K} \alpha_{sk} \hat{w}_k \Big) exp(\boldsymbol{X}_{rs}^T \boldsymbol{\beta}) \Big]. \tag{2.17}$$

Since (2.17) is of the form of the usual partial log-likelihood in the Cox model with known offsets, a standard software like R [25], in particular the function *coxph* of the package *survival*, can be used to obtain the maximal $\hat{\boldsymbol{\beta}}$.

2.2.2 Estimation of the Standard Errors

In case of the Cox model with shared frailty terms, it is not possible to compute the variance-covariance matrix directly from the marginal log-likelyhood, but it is possible to derive it from the observed information matrix, $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$.
The observed information matrix can be written as:

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \tag{2.18}$$

where $l(\boldsymbol{\theta})$ is the observable log-likelihood, obtained by integrating the full log-likelihood over $\boldsymbol{z}$:

$$\begin{aligned} l(\boldsymbol{\theta}) = &\sum_{j=1}^{J}\sum_{i=1}^{n_j} \delta_{ij} log(\lambda_0(t_{ij}) exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta})) + \\ &log\Big(\sum_{k=1}^{K} \pi_k w_k^{D_j} \cdot exp \sum_{i=1}^{n_j} \Big[-A_0(t_{ij}) w_k exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta})\Big]\Big), \end{aligned} \tag{2.19}$$

where D_j is the total number of events in cluster j, $D_j = \sum_{i=1}^{n_j} \delta_{ij}$.
A more computationally convenient approximation of the information matrix is proposed by Louis in [23]:

$$\boldsymbol{I}^j = \mathbb{E}[B_j] - \mathbb{E}[S_j S_j^T] + S_j^* S_j^{*T}, \tag{2.20}$$

where S and S^* are the gradient vectors of the full log-likelihood and the observable log-likelihood respectively, while B is the negative second derivative matrix of the full log-likelihood.

The standard errors related to the ratios of frailties can be estimated through the following formula:

$$Var\Big(\frac{\hat{w}_k}{\hat{w}_1}\Big) = \Big(\frac{\mu_{\hat{w}_k}}{\mu_{\hat{w}_1}}\Big)^2 \cdot \Big[\frac{\sigma^2_{\hat{w}_1}}{\mu^2_{\hat{w}_1}} + \frac{\sigma^2_{\hat{w}_k}}{\mu^2_{\hat{w}_k}} - \frac{2Cov(\hat{w}_1, \hat{w}_k)}{\mu_{\hat{w}_1}\mu_{\hat{w}_k}}\Big]. \tag{2.21}$$

2.2.3 Selection of the number of latent populations

As previously seen, an advantage of a nonparametric and dicrete distributed frailty is the possibility to build a probabilistic clustering technique. In fact, subject with the same frailty estimate can be considered as an unique latent population. In this section we will discuss how to select the correct number of latent population K.
Since it is impossible to estimate K using a log-likelihood maximization argument [11], we estimate $\boldsymbol{\theta}$ for each potential K and, then, we establish a model selection criterion. *Akaike's information criterion* (AIC), *Bayesian information criterion* (BIC) and Laird [22] are used to this aim.

AIC and BIC are founded on information theory: they offer an estimate of the relative information lost when a given model is used to represent the process that generated the data. In doing so, they deal with the trade-off between the goodness of fit of the model and the simplicity of the model. Indeed, when fitting models, it is possible to increase the likelihood by adding parameters, but this may also lead to overfitting. Both BIC and AIC attempt to solve this problem introducing a penalty term for the number of parameters in the model. The penalty term is different in BIC and in AIC. Denoting with l the maximum value of the likelihood function, with p the number of parameters and with n the number of observations, they can be written as:

$$AIC = -2log(l) + 2p \tag{2.22}$$

$$BIC = -2log(l) + p\ log(n) \tag{2.23}$$

They do not give indications about the quality of a model in absolute terms, but just evaluate the relative gain/loss passing from a model to another one. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC/BIC value. On the other hand, Laird proposes to choose the maximum (minimum) number of clusters, when the number of clusters increases (decreases) in the algorithm, such that in each population at least one member can be found. This usually leads to the choice of more complex models than the ones selected by AIC and BIC.

Chapter 3

Analysis of Trieste dataset

In this chapter we will report the analysis we made on the Trieste dataset, in particular we will present an application of the multi-state model described in Chapter 1.
This study is inspired by the work presented in [13].

In Section 3.1 we will introduce the dataset and in Section 3.2 we will report the relative descriptive analysis. In Section 3.3 we will perform a multi state modelling of such data. In Section 3.4 we will report the survival curves estimates obtained with the Kaplan-Meyer estimator.

3.1 Presentation of dataset from Trieste

In this section we will introduce the dataset, we will describe the preprocessing of data carried out on the dataset and we will explain the variables we decided to exctract from the dataset and that we considered for the implementation of the multi-state model.

The original dataset is composed by informations about 10,287 patients, identified by an univocal anonymous personal code, hospitalized with Heart Failure (HF) in the Trieste area.
The cohort we consider is composed of patients hospitalized between 2009 and 2016. 2017 is considered as a follow-up year. The five-year period from 2004 to 2008 is used for the calculation of significant clinical quantities.

Each row of the dataset refers to a specific event.
Possible events are:

- hospitalization for HF;
- hospitalization for any cause;
- Intermediate Care Unit admission (ICU);
- Integrated Home Care (IHC) activation;
- passage in Emergency Room (ER).

Several patient specific informations are recorded for each event: gender, age, length of stay, department of admission, presence of cardiological evaluation before hospitalization, laboratory tests, comorbidities.

These informations arise from the linkage between administrative data and *Cardionet*® clinical registry. This choice is giustified by the fact that the administrative data has been acquiring an important role over the years, providing useful information about the patient's status and pattern of care. An attempt of administrative data utilization in healthcare practice is described in [5], in [17] and in [24], where applications of multi-state models to HF patients's clinical evolution are presented.

For the following analysis we will focus only on hospital admissions, reshaping the dataset with the aim of fitting a multistate model as described in Chapter 1.
In what follows, we describe the preprocessing of data and the variable extracted.

3.1.1 Preprocessing

The original dataset is composed by 10,287 patients hospitalized in Trieste.
For the uncorrect registration of some dates we decide to not include in the study 3 patients (0.03% of the total population).
Since we decide to focus only on patients whose events last more than one

Starting: 10287 patients from Trieste			
Cause of removal	# removed	% removed	Patients left
Uncorrect dates	4	0.04%	10,283
One-day events	36	0.35%	10,247
Begin isn't a hospitalization	5	0.04%	10,242
Follow up	712	6.95%	9,530
Cohort 2009-2016 selection	2,479	26.01%	7,051
Only hospitalization events	93	3.75%	6,958

Table 3.1: Preprocessing steps: cause, number of patients removed, percentage of patients removed and number of patients left, at every step.

day, we don't include others 36 patients (0.35%).
5 more patients (0.04%) are excluded due to the fact that their clinical history deos not start with a hospitalization, as request by the clinical protocol under study.
In order to have at least one year of follow up for each considered patient, we do not consider the 712 patients (6.95%) whose index admission is dated after January 1st 2016.
Focusing the attention only on events regarding the cohort 2009-2016, 2,497 (26.01%) more patients are excluded.
Moreover, 93 more patients (3.75%) are excluded due to the fact that, in our study, we want to consider only hospitalizations as possible events.

In Table 3.1 all the preprocess steps are reported. For every step we can find the cause of the negletion, the number (%) of patients not considered and the number of patients left.
We can notice that after the preprocessing our dataset is composed by 6,958 patients.

3.1.2 Variables description

The variables used for the analysis can be divided in two groups: variables that do not depend on the hospitalizations (group 1) and variables which depend on hospitalizations (group 2).

Variables in group 1 are:

- **Sex**: 1 for male, 0 for female;
- **Death index**: 1 if patient die before the end of the study, 0 otherwise;
- **Worsening index**: 1 if patient have an hospitalization for HF in the five years preceding the index admission (worsening patient), 0 otherwise (*de novo* patient).

Variables in group 2 are:

- **Age** [years]: time difference between the admission date of the considered event and the date of birth;
- **Charlson index**: index of comorbidity;
- **Pre hospitalization cardiological evaluation**: 1 if the patient has an hospitalization in cardiology before the considered event, 0 otherwise;
- **Admission in Cardiological Ward (CW)**: 1 if patient is admitted in a cardiological ward, 0 otherwise;
- **ICU/IHC index**: 1 if patient have at least one events of type ICU or IHC before the considered hospitalization, 0 otherwise;
- **In-hospital death index**: 1 if patient die in the considered hospitalization, 0 otherwise.

Variables like sex, death index, in-hospital death, pre hospitalization cardiological evaluation and CW are present in the original dataset, so we simply exctract them. Variables like worsening index, Charlson index and ICU/IHC index has to be computed from other variables present in the original dataset.

In particular, for the creation of the worsening index, we compare for each patient the first event after 2009 (reference event), that could be hospitalization for HF (as in 72.06% of cases) or for any cause (as in 27.94% of cases), with all the existing events before it (if present): if an event regarding hospitalization for HF is present and the time difference between it and the reference event is less than five years, we classify the patient as worsening.

Charlson index is calculated using hospital diagnosis occured within five years prior to the first admission (like myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic obstructive pulmonary disease, rheumatic disease, peptic ulcer disease, mild liver disease, diabetes with and without complications, hemiplegia, renal disease, liver disease, cancer and AIDS) as suggested in [13] and refined with laboratory data and diagnosis recorded at the first admission, as indicated in [12].

Instead, for the creation of the ICU/IHC variable, we activate the index from the first ICU/IHC occurence until the last event recorded. We poin out that, in contrast with our decision of keeping only patients with hospitalization as first event, this index could be active also in first events. Indeed, we can note that after the elimination of the events preceding 2009, some patients have an ICU or a IHC as starting event. We decide to consider these starting events for the computation of the relative index and only after to eliminate all the ICU and IHC events.

3.1.3 Dataset transformation in long format

Once the selection of features and statistical unit is complete, we have to modify the dataset in order to make it manageable by the algorithms of the survival package in R [25], so to build the multi-state model.

In Table 3.2 we can see our data format after the preprocessing described in Section 3.1.1.
For a better visualization, we reported events regarding only one patient, identified by *ID* equal to 1, and we neglected the covariates.
Hospitalization events are identified by *state* number equal to 0 (hospitalization for HF) or 1 (hospitalization for any cause). We can see that there is one row for every single patient's hospitalization.

The first transformation step in order to obtain the long format is reported in Table 3.3.
First of all we have to consider every hospitalization as made by two different and separate events, since we want to distinguish between admission and

ID	state	adm_num	dateADM	dateDIS	dateOUT	DEATH_ind
1	0	1	2009-12-31	2010-01-09	2017-12-31	0
1	1	2	2015-05-28	2015-05-30	2017-12-31	0
1	1	3	2016-03-09	2016-03-16	2017-12-31	0

Table 3.2: Data format of the original dataset: one row for every single event.

discharge.
Consequentely, we create the *STATUSdef* column reported in Table 3.3, where the number of the hospitalization is reported togheter with the distinction if the event is an admission or a discharge.
DateADM and *dateDIS* columns in Table 3.2 are grouped in *dates* column in Table 3.3.
The *eventTIME* column is created. This variable indicates, for every patient, the days spent between the entrance in the study (indicated always with time equal to 0) and the corresponding event.

ID	dates	eventTIME	STATUSletter	STATUSnum	STATUSdef
1	2009-12-31	0	IN	1	1 IN
1	2010-01-09	9	OUT	1	1 OUT
1	2015-05-28	1974	IN	2	2 IN
1	2015-05-30	1976	OUT	2	2 OUT
1	2016-03-09	2260	IN	3	3 IN
1	2016-03-16	2267	OUT	3	3 OUT
1	2018-01-01	2923	OUT	3	3 OUT

Table 3.3: Data format after the first transformation step: one row for every different hospital admission or discharge.

After this first propedeutic transformation, we have to obtain a long format dataset, attackable by the function *Surv*. In order to obtain the long format, we have to consider all the possible transitions and status, so we create the *from* and *to* columns as in Table 3.4.
In Table 3.5 we report the correspondence between the states and the number with which they are indicated.
In the example reported in Table 3.4 we can see that the patient, with ID equal to 1, starts in state 1 (1IN) and has two possibilities: to pass to state 2 (1OUT) or to state 12 (D). The transition that actually takes place is

indicated with 1 in *status* variable. The time needed for the transition to happen is reported in column *time*, togheter with the transition initial time (*Tstart*) and the transition final time (*Tstop*).
The format in Table 3.4 is the long format needed by the function *Surv*.

ID	from	to	Tstart	Tstop	time	status
1	1	2	0	9	9	1
1	1	12	0	9	9	0
1	2	3	9	1974	1965	1
1	2	12	9	1974	1965	0
1	3	4	1974	1976	2	1
1	3	12	1974	1976	2	0
1	4	5	1976	2260	284	1
1	4	12	1976	2260	284	0
1	5	6	2260	2267	7	1
1	5	12	2260	2267	7	0
1	6	7	2267	2923	656	0
1	6	12	2267	2923	656	0

Table 3.4: Final dataset in long format: one row for every possible transition.

State number	State kind
1	1IN
2	1OUT
3	2IN
4	2OUT
5	3IN
6	3OUT
7	4IN
8	4OUT
9	5IN
10	5OUT
11	6+
12	D

Table 3.5: Correspondence between state number and state kind.

3.2 Descriptive analysis

From now on we focus the attention on two different dataset: in the first one (Overall data-OD) all patients are present, in the second one (DeNovo-DN) only *de novo* patients are considered. This latter distinction is intended to allow the study of incident cases only, which represent the 72.67% of the total.

3.2.1 Overall Data-OD

We start analysing and describing the characteristics of the OD population, whose summary is reported in Table 3.6.

The patients considered are 6,958. Among these 3,760 (54%) are females. 1,901 patients (27% of the total) had a HF hospitalization in the five years preceding the index admission (worsening patients).
The mean age at first hospitalization is 80.88 years, the corresponding standard deviation is 10.31.
The mean length of stay (LOS) of all the hospitalizations is 11.84 days.
Over the observing period 4,533 deaths (65%) are recorded; 2,788 patients (61% of the death) die during a hospitalization. On average in-hospital death occurs after 4 hospedalizations and last hospitalization has a mean length of 11.84 days (sd=15.69).

Focusing on indices at first hospitalizations, there is a prevalence of non-cardiac comorbidities in patient's background (26% has pulmonary disease, 10% cancer, 29% diabetes and 59% renal disease).
The mean Charlson index of 2.44, with a standard deviation of 2.19, witnesses a high level of comorbidity burden. The 82% of patients has this index grater that zero and the 66% has at least one morbidity. The rate of admission in CW is 19%. Finally, 36% of patients has a cardiological pre hospitalization visit.

Table 3.6: Summary of the overall population features.

General characteristics		
Number of patients	6,958	
Male	3,198	45.96%
DeNovo patients	5,057	72.67%
Deaths	4,533	65.14%
In hospital death	2,788	61.5%
Mean number of ospedalizations before death	4.22	
Length of last ospedalization (in hospital death)	m=11.58	sd=4.10
LOS	m=11.84	sd=15.69
Values refered to the first event		
Age	m=80.88	sd=10.31
Male age	m=77.46	sd=10.68
Female age	m=83.78	sd=9.03
Pulmonary disease	1,862	26.76%
Cancer	764	10.98%
Diabetes	2,069	29.73%
Renal disease	4,131	59.37%
Pre hospitalization cardiological evaluation	2,509	36.05%
Admission in cardiological ward	1377	19.79%
Comorbidities>0	4,625	66.47%
Charlson index>0	5,711	82.07%
Charlson index	m=2.44	sd=2.19
Values refered to the last event		
ICU>0	1,840	26.44%
Mean number of ICU	1.67	
Mean length of ICU	101.35	
IHC>0	3,072	44.15%
Mean number IHC	3.07	
Mean length of IHC	101.56	
ICU/IHC>0	3,621	52.04%

We analyse the frequency of ICU/IHC activations looking at variable ICU/ICH index at last hospitaliziation: in their history, 52% of patients has experienced at least one of this events. In particular, 26% of patients had ICU transition and 44% had an IHC activated. The ICU/IHC mean length is 101 days. The mean number of ICU events for a single patient (if we consider only who has experience it) is 1.67, while for the IHC events is 3.07.

In Table 3.7 a summary of the population features in terms of the number of hospitalizations is given considering only patients having up to 4 hospitalizations. We stop descripties at 4^{th} hospitalization since using this criterion we include about 70% of the entire population.

Looking at the rows referring to morbidity in Table 3.7, we can immediately note that the progressive number of readmission rates is associated with an increasing in comorbidity burden: we can see the Charlson index starting with a mean value of 2.36 and finishing with a mean value of 3.75. This increasing tendency is evident in every morbidity percentage: pulmonary disease, renal disease, diabetes and cancer.
The presence of a cardiological pre hospitalization visit is more common among patients who have a greater number of hospitalizations.
Another increasing index is the one that indicates the presence of ICU/IHC events. This seems reasonable because having more hospitalizations implies a longer permanence in the study and this, in turn, means having more probabilities of experience this kind of events. In particular the growth of the combined index rise from the 8% to the 70%.
A higher percentage of deaths is recorded among patients with multiple events. In particular, the number of deaths rise from the 60% to the 67%, the number of in-hospital deaths rise from the 50% of the deaths to the 65%. Conversely, the admission in cardiological ward decreases whit the progressive number of readmission rates.

In order to analyze the difference between the four populations described in Table 3.7, we made tests on proportions for the binary covariates and Kruscal-Wallis tests for continuous covariates. In both cases H_0 = same covariate distribution, H_1 = different covariate distribution.
The p-value referring to variable sex is 0.521, the one referring to variable

Table 3.7: Summary of the overall population features in terms of the number of hospitalizations considering only patients having up to 4 hospitalizations.

	1 hosp	2 hosp	3 hosp	4 hosp
General characteristics				
Num of patients	1,538 (22.10%)	1,392 (20%)	1,081 (15.53%)	849 (12.2%)
Male	682 (44.34%)	607 (43.60%)	464 (42.92%)	392 (46.17%)
DeNovo	1,206 (78.41%)	1,084 (77.87%)	806 (74.56%)	601 (70.78%)
Deaths	936 (60.85%)	896 (64.36%)	728 (67.34%)	570 (67.13%)
In hosp death	477 (50.96%)	562 (62.72)	451 (61.95%)	373 (65.43%)
Mean LOS (sd)	12.18 (13.45)	11.95 (12.41)	11.81 (12.50)	11.87 (11.60)
Values refered to the last event				
Mean age (sd)	82.20 (11.41)	83.05 (10.06)	83.50 (10.51)	83.26 (9.45)
Pulm. disease	333 (21.65%)	361 (25.93%)	328 (30.34%)	314 (36.98%)
Cancer	180 (11.70%)	240 (17.24%)	194 (17.94%)	166 (19.55%)
Diabetes	428 (27.82%)	365 (26.22%)	325 (30%)	290 (34.15%)
Renal disease	908 (59.03%)	871 (62.57%)	726 (67.16%)	593 (69.84%)
Pre hosp	443 (28.08%)	578 (41.52%)	460 (42.55%)	419 (49.35%)
CW	285 (18.53%)	175 (12.57%)	82 (7.58%)	70 (8.24%)
Mean Ch ind (sd)	2.36 (2.35)	2.9 (2.48)	3.33 (2.5)	3.75 (2.64)
ICU>0	35 (2.27%)	215 (15.44%)	267 (24.69%)	311 (36.63)
IHC>0	116 (7.54%)	409 (29.38%)	519 (48.01%)	496 (58.42%)
ICU/IHC>0	134 (8.71%)	529 (38%)	642 (59.38%)	601 (70.78%)

LOS is 0.621 and the one referring to variable age is 0.463. We can not reject the null hypothesis so we conclude that the distribution of these three variables in the four populations do not differ.

The p-values of the other variables are all very low: worsening index $= 8.34e-05$, polmonary desease $= 2.874e-15$, cancer $= 2.353e-07$, diabetes $= 4.794e-04$, renal desease $= 1.404e-07$, pre hospitalization cardiological evaluation $< 2.2e{-}16$, CW $< 2.2e{-}16$, Charlson index $< 2.2e{-}16$, ICU/IHC index $< 2.2e-16$. We can reject the null hypothesis and conclude that the distribution of these latter variable are different among the four populations.

3.2.2 De Novo data-DN

We proceed reporting the summary of the dataset regarding only *de novo* patients, namely that haven't experienced hospitalizations for HF in the five years preceding index admission.
In Table 3.8 we can find informations about all the *de novo* population, while in Table 3.9 a summary of the population features in terms of the number of hospitalizations is given considering only patients having up to 4 hospitalizations.

De novo patients are 5,057 (72% of the original data). The overall mean age is 80.7 with standard deviation of 10.54. Men are 46% and their mean age at first admission is 77.2, with standard deviation of 11.01. Women are 54% and their first admission mean age is 83.69 with relative standard deviation of 9.18.

Looking at Table 3.8, we note the presence of different morbidity at the first hospitalization: 22% of patients present pulmonary disease, 10% cancer, 27% diabetes and 56% renal disease. Overall 66% of patients has at least one morbidity and 78% has Charlson index grater than zero, with a mean value equal to 2.09 (sd=1.98).
36% of patients has a cardiological evaluation before the first admission and for 19% of them first hospitalization take place in CW.

Looking at the last admissions we can observe that the 23% of patients experiences at least one event in ICU, 38% an IHC activation, and overall the 46% at least one of the two, i.e. 46% has index ICU/IHC activated at the end of the observing period.
At the end of the study, 2,978 (59%) deaths are recorded; about the 60% of the deaths take place in hospital. The mean number of events recorded before death is 3.72.

We will now focus our attention on the Table 3.9, where patients are divided according to their total number of hospitalizations. There are the same trends we noted for the overall dataset in Section 3.2.1.

Table 3.8: Summary of the *de novo* population features.

General characteristics		
Number of patients	5,057	
Male	2,327	46.01%
Deaths	2,978	58.88%
In hospital death	1,784	59.9%
Mean number of ospedalizations before death	3.72	
Length of last ospedalization (in hospital death)	m=11.84	sd=14.42
LOS	m=11.48	sd=16.93
Values refered to the first event		
Age	m=80.7	sd=10.54
Male age	m=77.2	sd=11.01
Female age	m=83.69	sd=9.18
Pulmonary disease	1144	22.62%
Cancer	515	10.18%
Diabetes	1,403	27.74%
Renal disease	2,868	56.71%
Pre hospitalization cardiological evaluation	1,817	35.93%
Admission in cardiological ward	1,179	23.31%
Comorbidities>0	3,385	66.93%
Charlson index>0	3,948	78.07%
Charlson index	m=2.09	sd=1.98
Values refered to the last event		
ICU>0	1,180	23.33%
Mean number of ICU	1.61	
Mean length of ICU	93.13	
IHC>0	1,960	38.75%
Mean number IHC	2.83	
Mean length of IHC	93.22	
ICU/IHC>0	2,358	46.62%

The progressive number of readmission rates is associated with an increasing Charlson index and comorbidity burden at first admission: mean Charlson index ranges between 1.97 and 3.55. We can observe that these values are smaller than the corresponding ones in the complete dataset, due to the absence of worsening patients, whose clinical history starts before the events we are looking at in this table and consequently they could have a worse clinical condition.
The presence of a cardiological pre hospitalization visit also increases with the number of events recorded.
As we observed in Section 3.2.1, also the index indicating the presence of ICU/IHC events increases going from the 6% to the 67%.
Also an increase in mortality and in in-hospital mortality is observed.
Conversely, a decrease in admission in CW is pointed out.

In order to analyze the difference between the four populations described in Table 3.7, we carried out tests on proportions for the binary covariates and Kruscal-Wallis tests for continuous covariates (In both cases H_0 = same covariate distribution, H_1 = different covariate distribution).

The p-value referring to variable CW is 3.08e-12, the ones referring to variable Charlson index and ICU/IHC index are <2.2e-16. We can reject the null hypothesis and conclude that the distributions of these three variable are different among the four populations.
The p-values of the other variables are quite high: sex = 0.473, LOS = 0.462, age = 0.092, pulmonary disease = 0.346, cancer = 0.189, diabetes = 0.339, renal disease = 0.678, pre hospitalization cardiological evaluation = 0.510. In these cases we can not reject the null hypothesis and we conclude that the distributions of these variables is the same among the four populations. These populations are more similar to each other than the ones analyzed in Table 3.7. A possible explaination could be found in the fact that patients in DN dataset are more omogeneous than the ones in OD, where there are both worsening and *denovo* patients.

Table 3.9: Summary of the *de novo* population features in terms of the number of hospitalizations considering only patients having up to 4 hospitalizations.

	1 hosp	2 hosp	3 hosp	4 hosp
General characteristics				
Num of patients	884 (17.48%)	786 (15.54%)	590 (11.66%)	439 (8.68%)
Male	407 (46.04%)	350 (44.52%)	247 (41.86%)	208 (47.38%)
Deaths	463 (52.37%)	464 (59.03%)	359 (60.84%)	265 (60.36%)
In hosp death	202 (43.62%)	295 (63.57)	223 (62.11%)	174 (65.66%)
Mean LOS (sd)	11.36 (10.61)	11.49 (12.02)	11.73 (12.16)	11.55 (11.85)
Values refered to the last event				
Mean age (sd)	81.7 (11.82)	82.69 (10.19)	83.04 (11.50)	82.63 (9.96)
Pulm. disease	159 (17.98%)	177 (22.51%)	178 (30.16%)	152 (34.62%)
Cancer	91 (10.29%)	142 (18.06%)	101 (17.11%)	83 (18.9%)
Diabetes	234 (26.47%)	188 (23.91%)	167 (28.3%)	149 (33.94%)
Renal disease	479 (54.18%)	483 (61.45%)	396 (67.11%)	298 (67.88%)
Pre hosp	257 (29.07%)	343 (43.63%)	258 (43.72%)	224 (51.02%)
CW	189 (21.38%)	108 (13.74%)	82 (7.58%)	45 (10.25%)
Mean Ch ind (sd)	1.97 (2.02)	2.73 (2.42)	3.18 (2.44)	3.55 (2.56)
ICU>0	12 (1.35%)	99 (12.59%)	146 (24.74%)	149 (33.94%)
IHC>0	49 (5.54%)	204 (25.95%)	261 (44.23%)	234 (53.3%)
ICU/IHC>0	53 (5.99%)	257 (32.69%)	335 (56.77%)	294 (66.97%)

3.3 Inferential analysis

In this section we will present the inferential analysis made on data presented in Section 3.1.
In particular, in Section 3.3.1 we will explain the model we decided to implement on such data, whereas in Section 3.3.2 we will report the main results obtained from the application of the multi-state model.

3.3.1 Application of multi-state model to Trieste dataset

We implemented a multi-state model (as in Chapter 1.3) to jointly evaluate the impact of different risk factors on multiple hospital admissions, discharges and death.
In particular the model considered fits a *Cox-type* regression (2.1) for each transition. This kind of model provides a convenient description of the admission/discharge dynamics, pointing out which covariates act in which transitions and how they affect the istantaneous probability of going from one state to another.
As explained in Section 1.2, in the Cox model the hazard rate for a patient i at observation time t has the following form:

$$\alpha(t) = \alpha_0(t)\, exp(\beta_1 z_1 + ... + \beta_p z_p),$$

where β is a p-vector of unknown regression coefficients, $\alpha_0(t)$ is the baseline hazard and $exp(\beta_1 z_1 + ... + \beta_p z_p)$ is the hazard ratio (HR).

The possible states of the model are in this case hospitalizations (IN states in Figure 3.1), discharges (OUT states in Figure 3.1) and death (D state in Figure 3.1). We considered the hospitalizations as separated events until the sixth hospitalization. The adverse outcome of death is an absorbing state and a competing event with respect to all the other transitions. Overall there are 12 possible states and 21 possible transitions, represented in Figure 3.1.

The transitions can be grouped in 4 categories: admissions to hospital, hospital discharges, in-hospital deaths, out-of-hospital deaths. All the possible transitions with the corresponding categories, starting states and ending

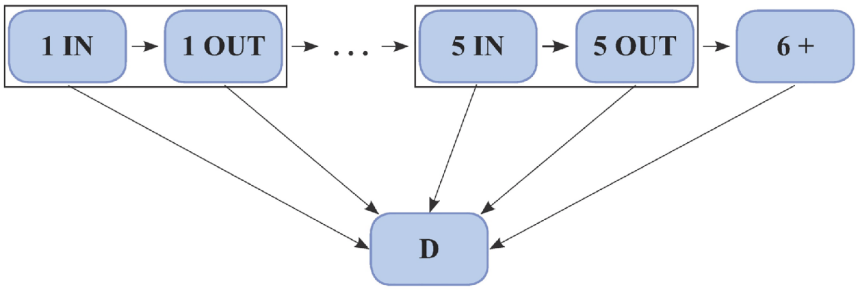


Figure 3.1: Sketch of multi-state model implemented: 12 possible states, 21 possible transitions.

states are shown in Table 3.10 .

We applied this model, sketched in Figure 3.1, to both the OD and DN dataset.

One appealing features of this model is the flexibility in introducing transition specific covariates. In order to decide in which transitions insert a particular variable, we tested the full model with all covariates in all transitions, then we retained only those which resulted to be significant from a statistical and/or clinical point of view.

The resulting choice is:

- Age is present in all transitions;
- Sex is present in hospital discharges and in-hospital deaths transitions;
- Charlson indexis present in all transitions;
- Pre hospitalization cardiological evaluation is present in in-hospital death and out-of-hospital deaths transitions;
- Admission in CW is present in hospital discharges, in-hospital death and out-of-hospital deaths transitions;
- Worsening index: is present in admission to hospital, in-hospital death and out-of-hospital deaths transitions;
- ICU/IHC indexis present in all transitions;

Table 3.10: Transition categories.

States connected	Transition ID	Transition type
1IN → 1OUT	1	hospital discharge
2IN → 2OUT	5	
3IN → 3OUT	9	
4IN → 4OUT	13	
5IN → 5OUT	17	
1IN → D	2	in-hospital death
2IN → D	6	
3IN → D	10	
4IN → D	14	
5IN → D	18	
6+ → D	21	
1OUT → 2IN	3	admission to hospital
2OUT → 3IN	7	
3OUT → 4IN	11	
4OUT → 5IN	15	
5OUT → 6+IN	19	
1OUT → D	4	out-of-hospital death
2OUT → D	8	
3OUT → D	12	
4OUT → D	16	
5OUT → D	20	

3.3.2 Results

In this section we report the results of fitting the Cox model, described in section 3.3.1, to both the OD and DN dataset.

Our quantities of interest are the hazard rates estimates, the istantaneous probabilities of going from a state to another. In particular we focus on the hazard ratios concerning each covariate.

In order to be able to correctly interpret the following tables and figures, we underline that each hazard ratio estimate is computed once fixed all the others pairs of variables/transitions.

OD

In Tables from 3.11 to 3.17, we report the hazard ratio estimates for each pair of variable/transition considered in the Cox model together with the p-value of the significance test of the single HR ($H_0 : exp(\beta_i) = 1$ VS $H_1 : exp(\beta_i) \neq 1$, with i from 1 to p), the number of the asterisks reflects the importance of the coefficient.

In every table we group together transitions belonging to the same category. For example, in Table 3.11 there are 4 blocks, the first from above contains hospital discharges transitions, the second in-hospital deaths transitions, the third admissions to hospital transitions and the last out-of-hospital deaths transitions.
Moreover, in every row, the number near the variable name indicates the transition considered, as reported in column *Transition ID* in Table 3.10.

Table 3.11: Hazard ratios estimates for the transitions of variable age, estimated fitting Cox model to OD.

		HR	Pr(>|z|)	
Hospital discharge	AGE.1	0.997	0.038	*
	AGE.5	0.997	0.071	.
	AGE.9	0.997	0.153	
	AGE.13	0.997	0.242	
	AGE.17	1.001	0.750	
In-hospital death	AGE.2	1.079	$< 2e-16$	***
	AGE.6	1.070	$< 2e-16$	***
	AGE.10	1.060	$< 2e-16$	***
	AGE.14	1.040	$< 2e-16$	***
	AGE.18	1.077	$< 2e-16$	***
	AGE.21	1.047	$< 2e-16$	***
Admissions to hospital	AGE.3	1.014	$< 2e-16$	***
	AGE.7	1.010	$< 2e-16$	***
	AGE.11	1.010	$< 2e-16$	***
	AGE.15	1.010	$< 2e-16$	***
	AGE.19	1.010	0.001	***
Out-of-hospital death	AGE.4	1.087	$< 2e-16$	***
	AGE.8	1.073	$< 2e-16$	***
	AGE.12	1.067	$< 2e-16$	***
	AGE.16	1.080	$< 2e-16$	***
	AGE.20	1.032	0.007	**

Table 3.12: Hazard ratios estimates for the transitions of variable sex, estimated fitting Cox model to OD.

		HR	Pr(>|z|)	
Hospital discharge	SEX.1	1.067	0.015	*
	SEX.5	1.084	0.008	**
	SEX.9	1.042	0.245	
	SEX.13	1.055	0.196	
	SEX.17	0.975	0.614	
In-hospital death	SEX.2	1.228	0.035	*
	SEX.6	1.107	0.263	
	SEX.10	1.197	0.073	.
	SEX.14	0.988	0.913	
	SEX.18	1.395	0.011	*
	SEX.21	1.244	0.001	**

Table 3.13: Hazard ratios estimates for the transitions of variable Charlson index, estimated fitting Cox model to OD.

		HR	Pr(>|z|)	
Hospital discharge	CHARLSON.1	0.949	$< 2e-16$	***
	CHARLSON.5	0.956	$< 2e-16$	***
	CHARLSON.9	0.976	0.001	***
	CHARLSON.13	0.960	$< 2e-16$	***
	CHARLSON.17	0.969	0.001	**
In-hospital death	CHARLSON.2	1.087	$< 2e-16$	***
	CHARLSON.6	1.089	$< 2e-16$	***
	CHARLSON.10	1.072	$< 2e-16$	***
	CHARLSON.14	1.045	0.037	*
	CHARLSON.18	1.058	0.024	*
	CHARLSON.21	1.089	$< 2e-16$	***
Admissions to hospital	CHARLSON.3	1.096	$< 2e-16$	***
	CHARLSON.7	1.081	$< 2e-16$	***
	CHARLSON.11	1.075	$< 2e-16$	***
	CHARLSON.15	1.073	$< 2e-16$	***
	CHARLSON.19	1.071	$< 2e-16$	***
Out-of-hospital death	CHARLSON.4	1.199	$< 2e-16$	***
	CHARLSON.8	1.190	$< 2e-16$	***
	CHARLSON.12	1.145	$< 2e-16$	***
	CHARLSON.16	1.108	$< 2e-16$	***
	CHARLSON.20	1.150	$< 2e-16$	***

Table 3.14: Hazard ratios estimates for the transitions of variable pre hospitalization cardiological evaluation, estimated fitting Cox model to OD.

		HR	Pr(>|z|)	
In-hospital death	PRE_HOSP.2	0.943	0.601	
	PRE_HOSP.6	1.002	0.983	
	PRE_HOSP.10	0.787	0.025	*
	PRE_HOSP.14	0.840	0.119	
	PRE_HOSP.18	0.915	0.501	
	PRE_HOSP.21	0.895	0.094	.
Out-of-hospital death	PRE_HOSP.4	0.687	0.002	**
	PRE_HOSP.8	0.582	$< 2e-16$	***
	PRE_HOSP.12	0.840	0.199	
	PRE_HOSP.16	0.951	0.746	
	PRE_HOSP.20	0.807	0.269	

Table 3.15: Hazard ratios estimates for the transitions of variable admission in CW, estimated fitting Cox model to OD.

		HR	Pr(>|z|)	
Hospital discharge	CW.1	1.522	$< 2e-16$	***
	CW.5	1.454	$< 2e-16$	***
	CW.9	1.495	$< 2e-16$	***
	CW.13	1.412	$< 2e-16$	***
	CW.17	1.420	$< 2e-16$	***
In-hospital death	CW.2	0.378	0.001	**
	CW.6	0.576	0.017	*
	CW.10	0.541	0.040	*
	CW.14	0.550	0.039	*
	CW.18	0.327	0.028	*
	CW.21	0.477	$< 2e-16$	***
Out-of-hospital death	CW.4	0.562	0.005	**
	CW.8	0.679	0.108	
	CW.12	0.484	0.029	*
	CW.16	0.624	0.181	
	CW.20	0.602	0.241	

Table 3.16: Hazard ratios estimates for the transitions of variable worsening index, estimated fitting Cox model to OD.

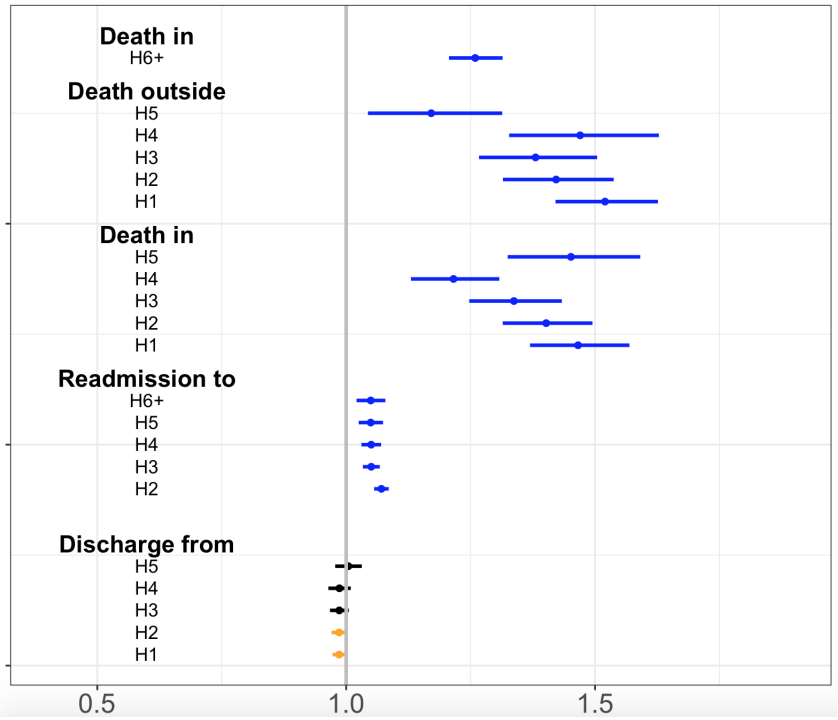
		HR	Pr(>|z|)	
In-hospital death	WORSENING.2	1.088	0.402	
	WORSENING.6	0.835	0.063	.
	WORSENING.10	0.837	0.084	.
	WORSENING.14	1.014	0.902	
	WORSENING.18	0.749	0.033	*
	WORSENING.21	0.916	0.177	
Admissions to hospital	WORSENING.3	1.148	$< 2e-16$	***
	WORSENING.7	1.180	$< 2e-16$	***
	WORSENING.11	1.138	0.001	**
	WORSENING.15	1.163	0.001	**
	WORSENING.19	1.102	0.071	.
Out-of-hospital death	WORSENING.4	0.698	0.002	**
	WORSENING.8	0.917	0.484	
	WORSENING.12	0.868	0.295	
	WORSENING.16	0.997	0.987	
	WORSENING.20	0.895	0.558	

Table 3.17: Hazard ratios estimates for the transitions of variable ICU/IHC index, estimated fitting Cox model to OD.

		HR	Pr(>|z|)	
Hospital discharge	ICU_IHC.1	0.827	$< 2e-16$	***
	ICU_IHC.5	0.871	$< 2e-16$	***
	ICU_IHC.9	0.826	$< 2e-16$	***
	ICU_IHC.13	0.884	0.003	**
	ICU_IHC.17	0.871	0.006	**
In-hospital death	ICU_IHC.2	1.774	$< 2e-16$	***
	ICU_IHC.6	1.298	0.003	**
	ICU_IHC.10	1.483	$< 2e-16$	***
	ICU_IHC.14	1.855	$< 2e-16$	***
	ICU_IHC.18	1.683	0.001	**
	ICU_IHC.21	1.011	0.093	.
Admissions to hospital	ICU_IHC.3	0.989	0.857	
	ICU_IHC.7	1.143	$< 2e-16$	***
	ICU_IHC.11	1.044	0.255	
	ICU_IHC.15	1.000	0.994	
	ICU_IHC.19	0.987	0.810	
Out-of-hospital death	ICU_IHC.4	1.447	0.039	*
	ICU_IHC.8	1.150	0.220	
	ICU_IHC.12	1.312	0.031	*
	ICU_IHC.16	1.530	0.009	**
	ICU_IHC.20	2.316	0.001	***

In Figures, from 3.2 to 3.8, we can find the plots of the hazard ratio estimates and their corresponding 95% confidence intervals. We consider one variable at a time, separating the different transitions for a better visualization.

Figure 3.2: 95% confidence intervals for hazard ratios of age (all other covariates fixed) estimated fitting Cox model to OD.



We point out that if the coefficient β_i of a variable x_i is bigger than zero then it acts increasing the hazard ratio, that consequently will be greater than one, this latter will increase the probability for the considered transition to happen. Conversely, if the same coefficient is smaller than zero then it acts decreasing the hazard ratio, that consequentely will be smaller than one and will decrease the probability of the transition.
In order to distinguish the positive/negative effects of the hazard ratios, we colored the confidence intervals accordingly: the orange ones are lower than one, the blue ones are greater than one and the black ones are straddling the value one, i.e., they come out not to be significant.

Figure 3.2 shows the confidence intervals for hazard ratios of age variable. This variable increases the probabilities of every transition of kind admission to hospital, in-hospital deaths and out-of-hospital deaths. In particular the effects on death transitions are greater than the others.
Conversely, this variable decreases the probabilities of discharge from hospital transitions or not influences them.

The same behavior is observed in the confidence intervals of hazard ratios of Charlson index variable, reported in Figure 3.3.

Figure 3.3: 95% confidence intervals for hazard ratios of Charlson index (all other covariates fixed) estimated fitting Cox model to OD.

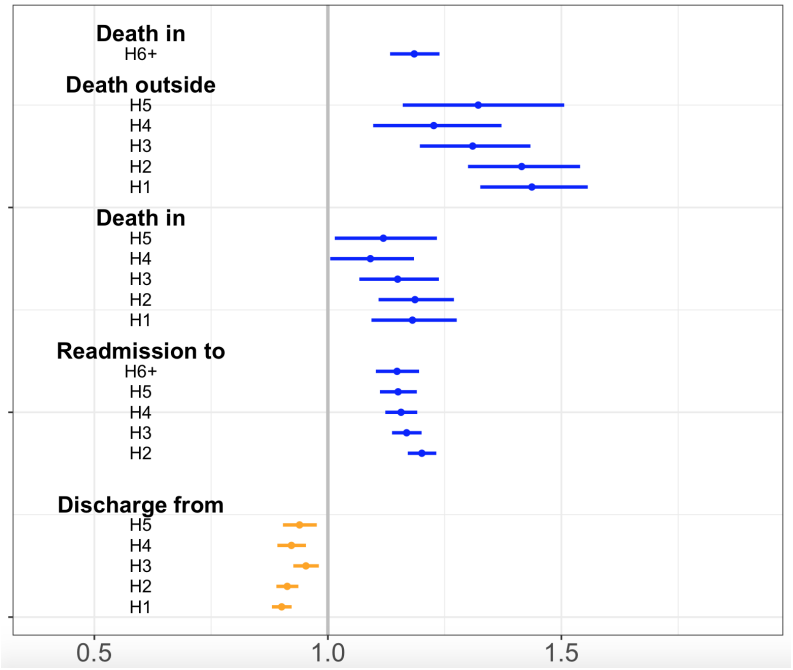


Figure 3.4 shows the confidence intervals for hazard ratios of sex variable. We note that the significant terms are all positive. This indicates that being a man increases the probabilities of being discharged in the first hospitalization but also of dying in hospital, in the first and in the latest hospitalizations.

Figure 3.5 shows the confidence intervals for hazard ratios of pre hospitalization cardiological evaluation variable.

Only few terms are significative: one in-hospital deaths transition and two out-of-hospital deaths transitions. Both terms are smaller than one meaning that having a pre hospitalization decreases the probabilities of having a transition to death.

Figure 3.6 shows the confidence intervals for hazard ratios of CW admission variable.
The hazard ratios related to the discharge from hospital are all bigger than one, indicating that being admitted in CW increases the LOS of a patients, probably due to the more severe conditions of the patient. On the other hand, the hazard ratios related to death inside or outside hospital are all smaller than one. Being admitted in this ward is a protected factor for the deaths since it decreases the probability to have a transition to death.

Figure 3.4: 95% confidence intervals for hazard ratios of sex (all other covariates fixed) estimated fitting Cox model to OD.

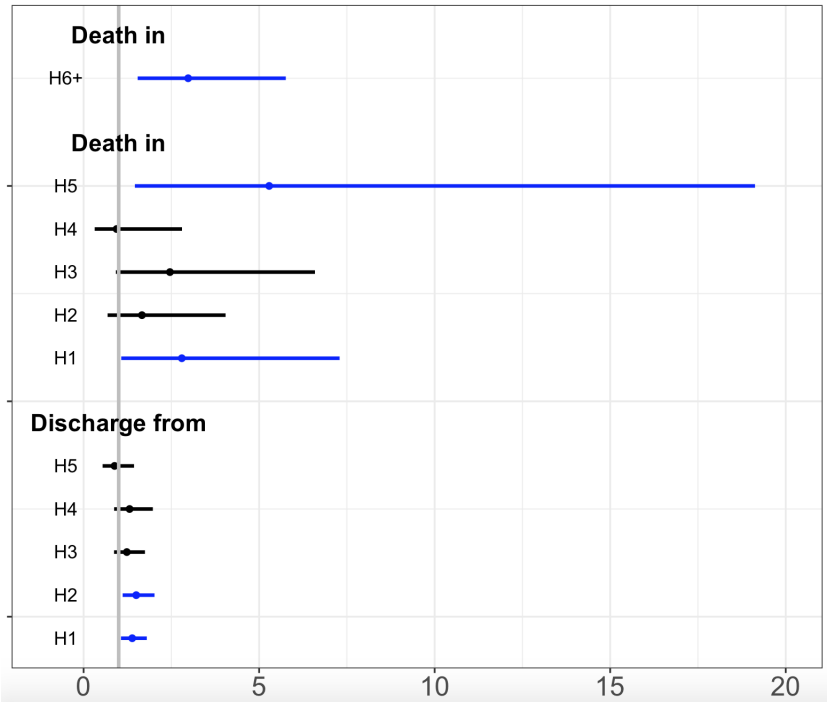


Figure 3.7 shows the confidence intervals for hazard ratios of worsening index variable.
Having a hospitalization in the five years before the index admission in-

Figure 3.5: 95% confidence intervals for hazard ratios of pre hospitalization cardiological evaluation (all other covariates fixed) estimated fitting Cox model to OD.

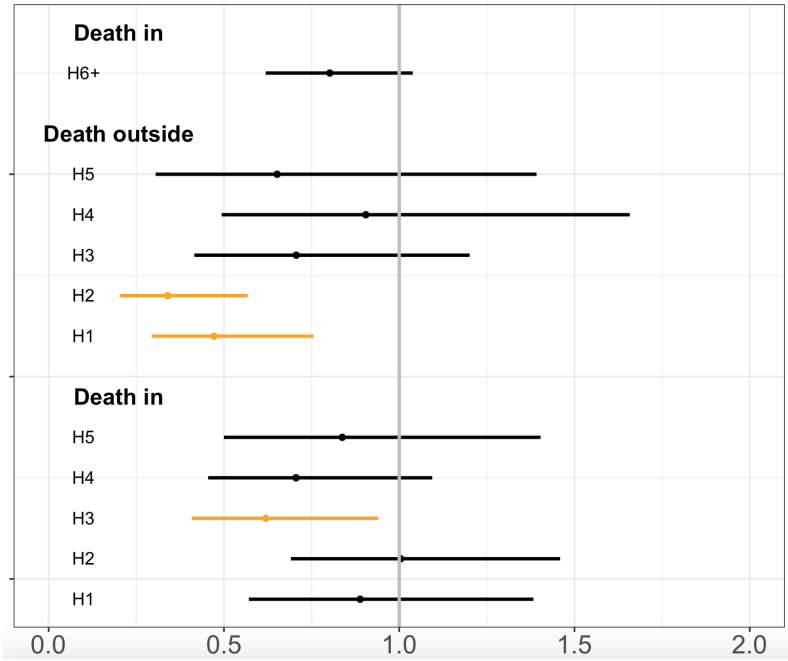
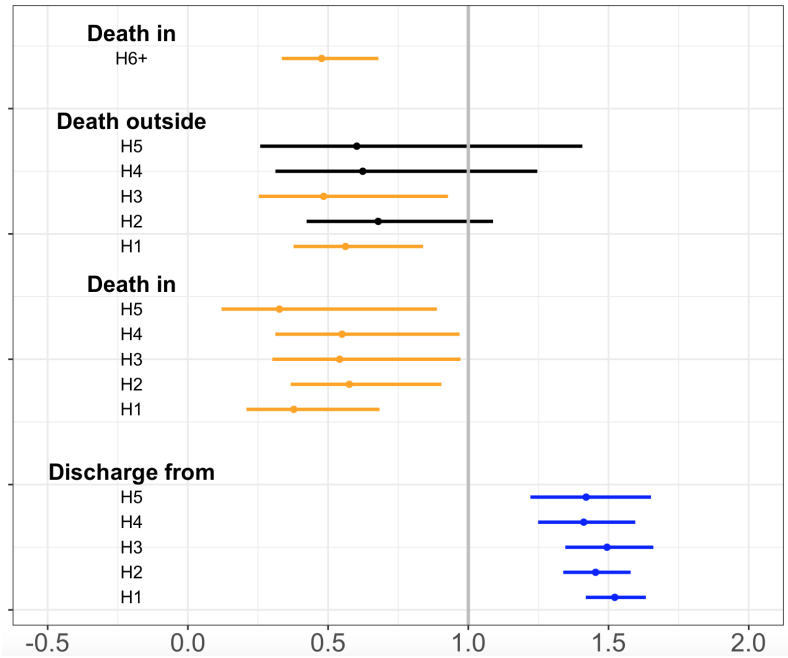


Figure 3.6: 95% confidence intervals for hazard ratios of admission in CW (all other covariates fixed) estimated fitting Cox model to OD.



creases the istantaneous risk of being readmitted and decreases the risk of death (in-hospital for the last hospitalization and out-of-hospital for the first one).

Figure 3.8 shows the confidence intervals for hazard ratios of ICU/IHC index variable.
Looking at them we can suppose that ICU/IHC index allows to identify the most fragile population, because experiencing one of these events decreases the probabilities of being discharged from hospital and increases the ones of being readmitted in hospital and of dying. However we will focus on this assumption later in Section 3.5.

Figure 3.7: 95% confidence intervals for hazard ratios of worsening index (all other covariates fixed) estimated fitting Cox model to OD.

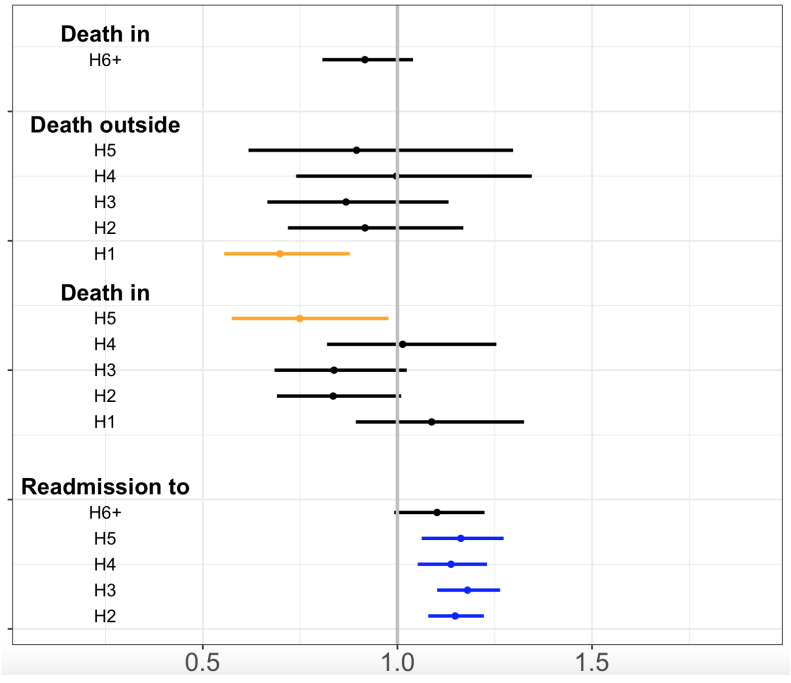
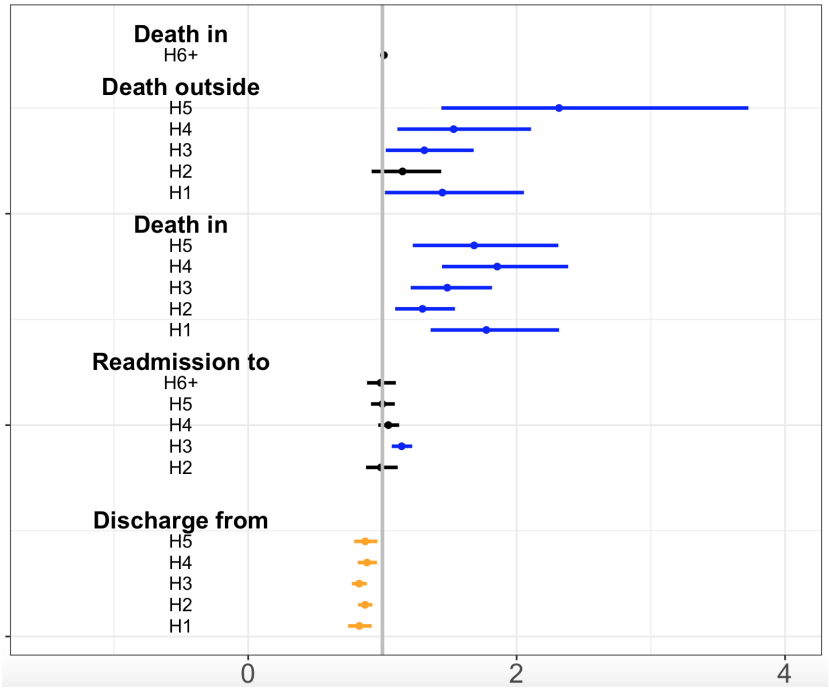


Figure 3.8: 95% confidence intervals for hazard ratios of ICU/IHC index (all other covariates fixed) estimated fitting Cox model to OD.



DN

In this paragraph we report the results of the application of the model described in Section 3.3.1 to the DN data.
We remind that this dataset consist only of *de novo* patients, namely the incident cases, not having a HF hospitalization in the five years preceding their index admission. For this reason, the *worsening* variable won't be necessary.
As it can be evinced by Tables from 3.18 to 3.23 and from Figures from 3.9 to 3.14, all the conclusion pointed out in Section 3.3.2 hold also in this case.

Table 3.18: Hazard ratios estimates for the transitions of variable age, estimated fitting Cox model to DN data.

		HR	Pr(>|z|)	
Hospital discharge	AGE.1	0.997	0.038	*
	AGE.5	0.997	0.071	.
	AGE.9	0.997	0.153	
	AGE.13	0.997	0.242	
	AGE.17	1.001	0.750	
In-hospital death	AGE.2	1.080	$< 2e-16$	***
	AGE.6	1.071	$< 2e-16$	***
	AGE.10	1.060	$< 2e-16$	***
	AGE.14	1.040	$< 2e-16$	***
	AGE.18	1.078	$< 2e-16$	***
	AGE.21	1.048	$< 2e-16$	***
Admissions to hospital	AGE.3	1.014	$< 2e-16$	***
	AGE.7	1.010	$< 2e-16$	***
	AGE.11	1.010	$< 2e-16$	***
	AGE.15	1.010	$< 2e-16$	***
	AGE.19	1.010	0.001	***
Out-of-hospital death	AGE.4	1.087	$< 2e-16$	***
	AGE.8	1.073	$< 2e-16$	***
	AGE.12	1.067	$< 2e-16$	***
	AGE.16	1.080	$< 2e-16$	***
	AGE.20	1.032	0.007	**

Table 3.19: Hazard ratios estimates for the transitions of variable sex, estimated fitting Cox model to DN data.

		HR	Pr(>|z|)	
Hospital discharge	SEX.1	1.067	0.015	*
	SEX.5	1.084	0.008	**
	SEX.9	1.042	0.245	
	SEX.13	1.055	0.196	
	SEX.17	0.975	0.614	
In-hospital death	SEX.2	1.228	0.036	*
	SEX.6	1.106	0.265	
	SEX.10	1.187	0.089	.
	SEX.14	0.988	0.915	
	SEX.18	1.373	0.015	*
	SEX.21	1.242	0.001	**

Table 3.20: Hazard ratios estimates for the transitions of variable Charlson index, estimated fitting Cox model to DN data.

		HR	Pr(>|z|)	
Hospital discharge	CHARLSON.1	0.949	$< 2e-16$	***
	CHARLSON.5	0.956	$< 2e-16$	***
	CHARLSON.9	0.976	0.001	***
	CHARLSON.13	0.960	$< 2e-16$	***
	CHARLSON.17	0.969	0.001	***
In-hospital death	CHARLSON.2	1.091	$< 2e-16$	***
	CHARLSON.6	1.084	$< 2e-16$	***
	CHARLSON.10	1.068	$< 2e-16$	***
	CHARLSON.14	1.045	0.035	*
	CHARLSON.18	1.052	0.040	*
	CHARLSON.21	1.088	$< 2e-16$	***
Admissions to hospital	CHARLSON.3	1.104	$< 2e-16$	***
	CHARLSON.7	1.087	$< 2e-16$	***
	CHARLSON.11	1.080	$< 2e-16$	***
	CHARLSON.15	1.076	$< 2e-16$	***
	CHARLSON.19	1.073	$< 2e-16$	***
Out-of-hospital death	CHARLSON.4	1.180	$< 2e-16$	***
	CHARLSON.8	1.186	$< 2e-16$	***
	CHARLSON.12	1.140	$< 2e-16$	***
	CHARLSON.16	1.107	$< 2e-16$	***
	CHARLSON.20	1.147	$< 2e-16$	***

Table 3.21: Hazard ratios estimates for the transitions of variable pre hospitalization cardiological evaluation, estimated fitting Cox model to DN data.

		HR	Pr(>|z|)	
In-hospital death	PRE_HOSP.2	0.940	0.582	
	PRE_HOSP.6	1.007	0.937	
	PRE_HOSP.10	0.796	0.032	*
	PRE_HOSP.14	0.839	0.116	
	PRE_HOSP.18	0.944	0.660	
	PRE_HOSP.21	0.901	0.113	
Out-of-hospital death	PRE_HOSP.4	0.688	0.002	**
	PRE_HOSP.8	0.585	$< 2e-16$	***
	PRE_HOSP.12	0.849	0.226	
	PRE_HOSP.16	0.951	0.746	
	PRE_HOSP.20	0.816	0.292	

Table 3.22: Hazard ratios estimates for the transitions of variable admission in CW, estimated fitting Cox model to DN data.

		HR	Pr(>|z|)	
Hospital discharge	CW.1	1.522	$< 2e-16$	***
	CW.5	1.454	$< 2e-16$	***
	CW.9	1.495	$< 2e-16$	***
	CW.13	1.412	$< 2e-16$	***
	CW.17	1.420	$< 2e-16$	***
In-hospital death	CW.2	0.373	0.001	**
	CW.6	0.578	0.017	*
	CW.10	0.542	0.041	*
	CW.14	0.550	0.038	*
	CW.18	0.323	0.027	*
	CW.21	0.474	$< 2e-16$	***
Out-of-hospital death	CW.4	0.584	0.008	**
	CW.8	0.681	0.111	
	CW.12	0.486	0.029	*
	CW.16	0.624	0.181	
	CW.20	0.600	0.238	

Table 3.23: Hazard ratios estimates for the transitions of variable ICU/IHC index, estimated fitting Cox model to DN data.

		HR	Pr(>|z|)	
Hospital discharge	ICU_IHC.1	0.827	$< 2e-16$	***
	ICU_IHC.5	0.871	$< 2e-16$	***
	ICU_IHC.9	0.826	$< 2e-16$	***
	ICU_IHC.13	0.884	0.003	**
	ICU_IHC.17	0.871	0.006	**
In-hospital death	ICU_IHC.2	1.826	$< 2e-16$	***
	ICU_IHC.6	1.255	0.008	**
	ICU_IHC.10	1.453	$< 2e-16$	***
	ICU_IHC.14	1.859	$< 2e-16$	***
	ICU_IHC.18	1.633	0.002	**
	ICU_IHC.21	1.009	0.137	
Admissions to hospital	ICU_IHC.3	1.038	0.534	
	ICU_IHC.7	1.170	$< 2e-16$	***
	ICU_IHC.11	1.060	0.124	
	ICU_IHC.15	1.008	0.853	
	ICU_IHC.19	0.998	0.973	
Out-of-hospital death	ICU_IHC.4	1.270	0.170	
	ICU_IHC.8	1.136	0.258	
	ICU_IHC.12	1.292	0.041	*
	ICU_IHC.16	1.530	0.009	**
	ICU_IHC.20	2.284	0.001	***

Figure 3.9: 95% confidence intervals for hazard ratios of age (all other covariates fixed) estimated fitting Cox model to DN data.

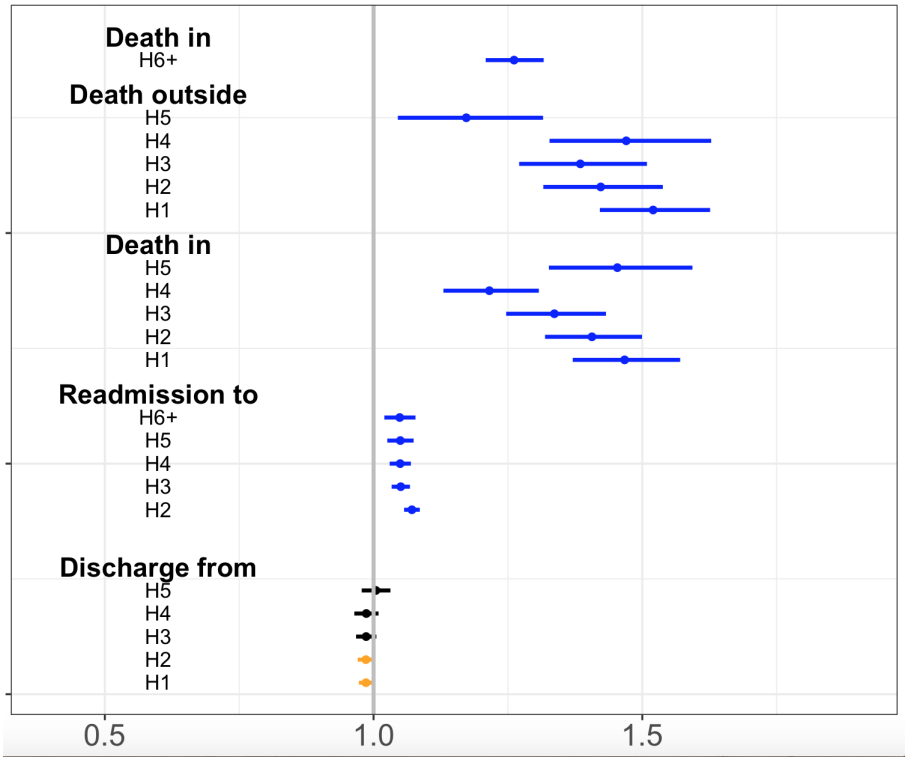


Figure 3.10: 95% confidence intervals for hazard ratios of sex (all other covariates fixed) estimated fitting Cox model to DN data.

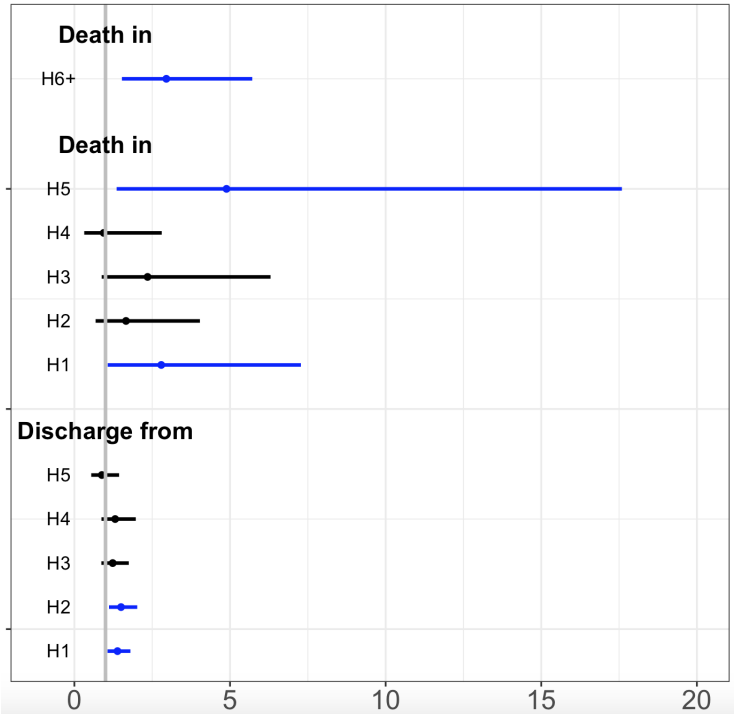


Figure 3.11: 95% confidence intervals for hazard ratios of Charlson index (all other covariates fixed) estimated fitting Cox model to DN data.

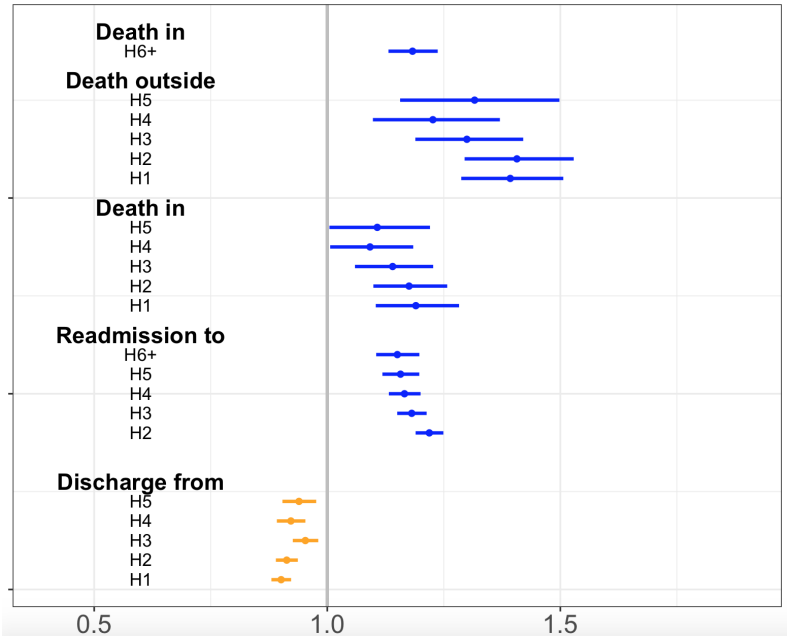


Figure 3.12: 95% confidence intervals for hazard ratios of pre hospitalization cardiological evaluation (all other covariates fixed) estimated fitting Cox model to DN data.

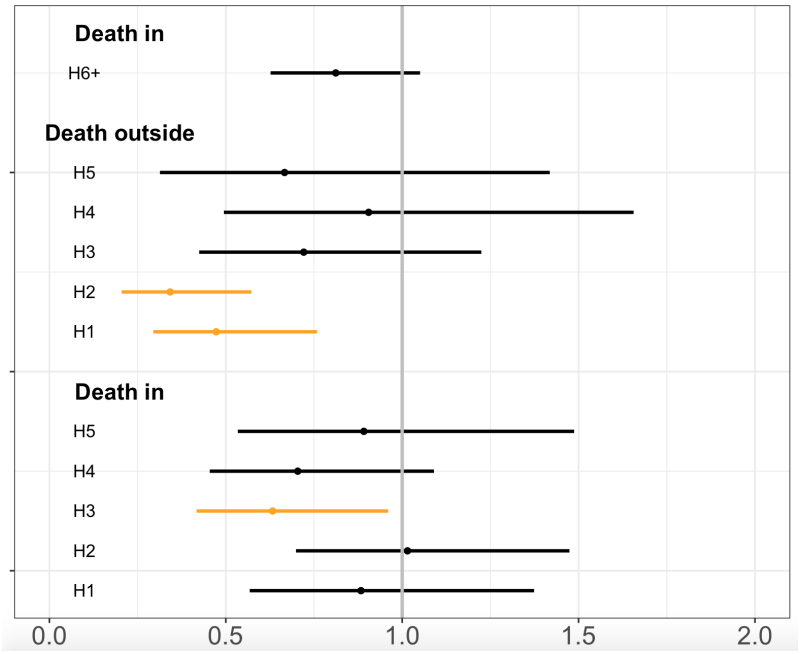


Figure 3.13: 95% confidence intervals for hazard ratios of admission in CW (all other covariates fixed) estimated fitting Cox model to DN data.

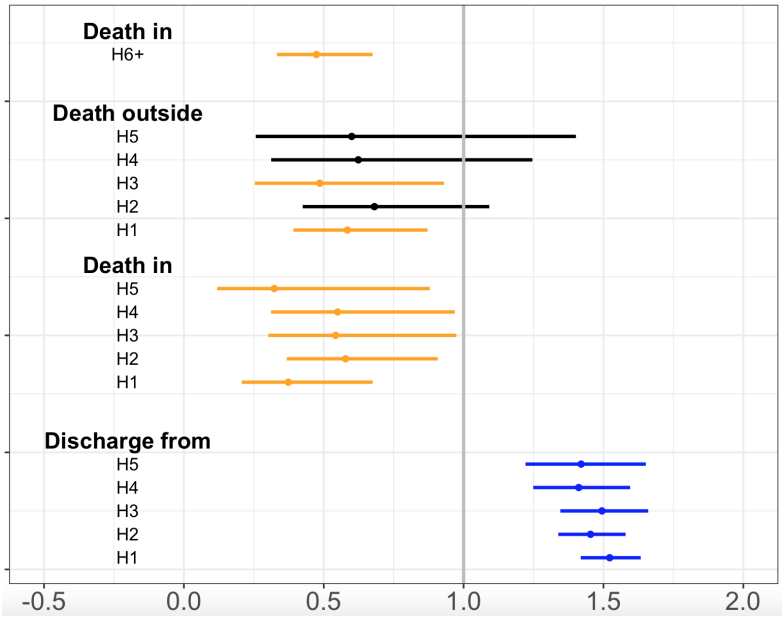
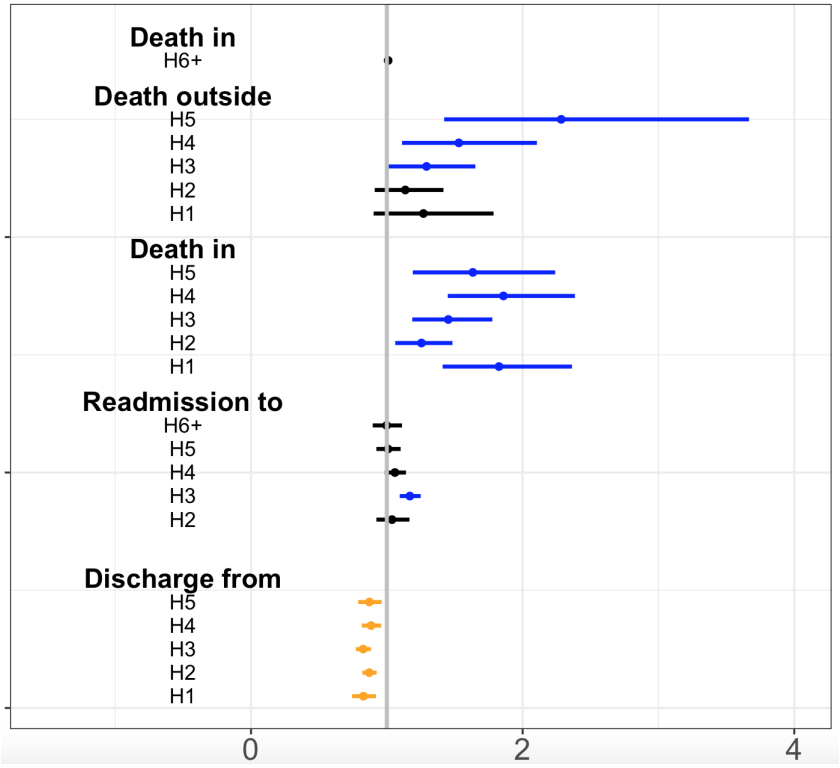


Figure 3.14: 95% confidence intervals for hazard ratios of ICU/IHC index (all other covariates fixed) estimated fitting Cox model to DN data.



3.4 Kaplan-Meier curves

In this section we report the analysis carried out on OD dataset using Kaplan-Meier (KM) curves, introduced in Section 1.2.3.

In this section we will focus on two different times to event. In Section 3.4.1 we will study the survival time, hence the time between the first admission in the study and the death of the patients, if it occurs. In Section 3.4.2 we will study the time between the first and the second hospitalization, if it occurs. In this latter case, we are interested in the time needed for the second hospitalization to happen, so we don't consider patients who die between the two hospitalizations. We decide to focus on second admission because it could be seen as a marker of success of the initial treatment.

In any survival plot, from 3.15 to 3.23, on the x-axis there is the calendar time, from 0 to the last time recorded, on the y-axis there is the probability for a people to not experience the event of interest, the death for figures in Section 3.4.1, the second hospitalization for figures in Section 3.4.2.

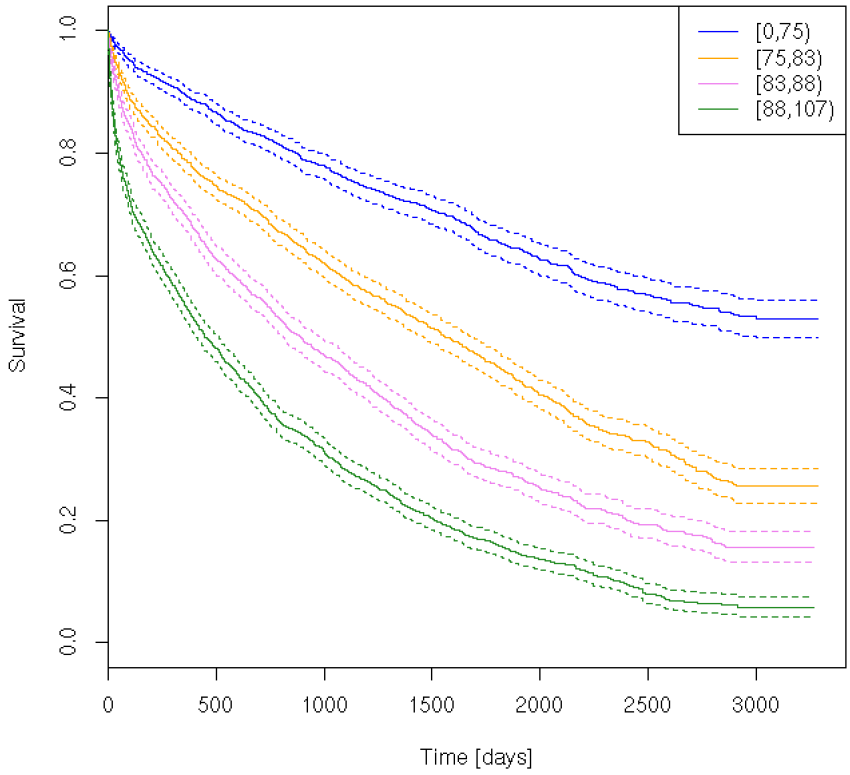
It is possible to stratify a survival curves according to some characteristics of interest, in order to investigate the effect of the stratifying variable on the outcome.
In our case, we will consider stratifications induced by age at first admission, sex, Charlson index at first admission, presence of pre hospital cardiological evaluation before first admission and presence of ICU admission or IHC activation among all the registered events.
All the plots reported are created using the OD dataset.

3.4.1 KM curves on survival time

Figures from 3.15 to 3.19 are obtained considering the difference between the first admission in the study and the death of the patients, if it occurs.
We stratify the survival curve for each variable described in Section 3.1.2, in order to analyze if being part of a given population's subgroup makes differences in terms of survival probability.

In Figure 3.15, survival time is stratified according to the age of the patients at first admission. We categorized the age variable using the empirical quartiles, i.e. we grouped togheter the patients whose age at first admission was below the first quatile (75), between the first and the second quartile (83), between the second and the third quartile (88) and above the third quartile.
In order to analyze the difference between the curves we use the *Log-Rank* test, where the null hypothesis is H_0: no difference between survival curves. The corresponding p-value, reported in Table 3.24, is <2.2e-16, so we can conclude that the curves are different and, consequently, that age influences the survival probability, in particular the latter decreases with aging.

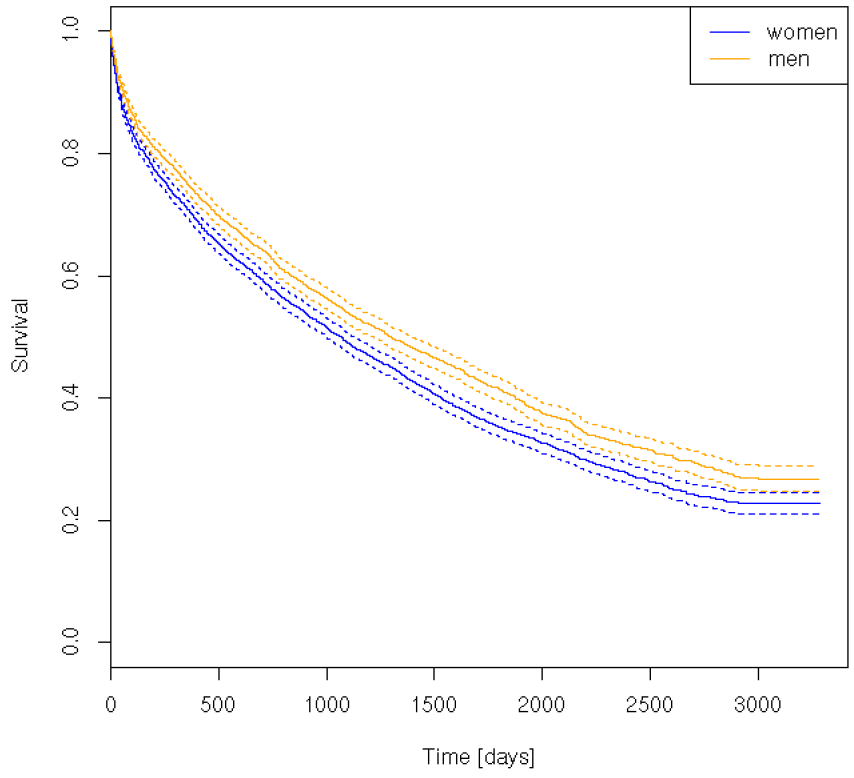
Figure 3.15: KM curves of survival time stratified by age of patients at their first admission.



In Figure 3.16 survival is stratified according to the sex of the patients.

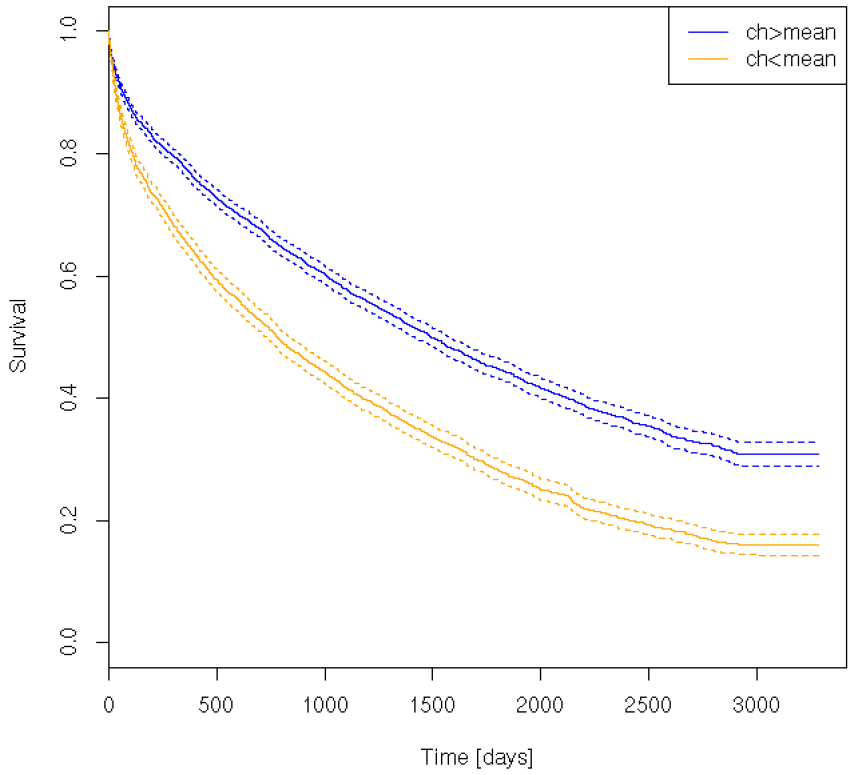
In Table 3.24 the p-value of the *Log-Rank* test is shown, its low value (6.62e-07) allows us to consider the two curves as different. We can notice that the curve of the women is always below the one of the men, indicating that women have a smaller survival probability. The fact that women mean age is higher than men mean age can explain this results.

Figure 3.16: KM curves of survival time stratified by sex of patients.



In Figure 3.17 survival is stratified according to the Charlson index of the patients computed at first hospitalization.
Since this index ranges between 0 and 17, we dichotomized the variable according to the mean value (2.45).
In Table 3.24 the p-value of the *Log-Rank* test is shown, from his low value (<2.2e-16) we can conclude that the curves are different and that Charlson index influence the survival: the higher the commorbidity level the lower the survival.

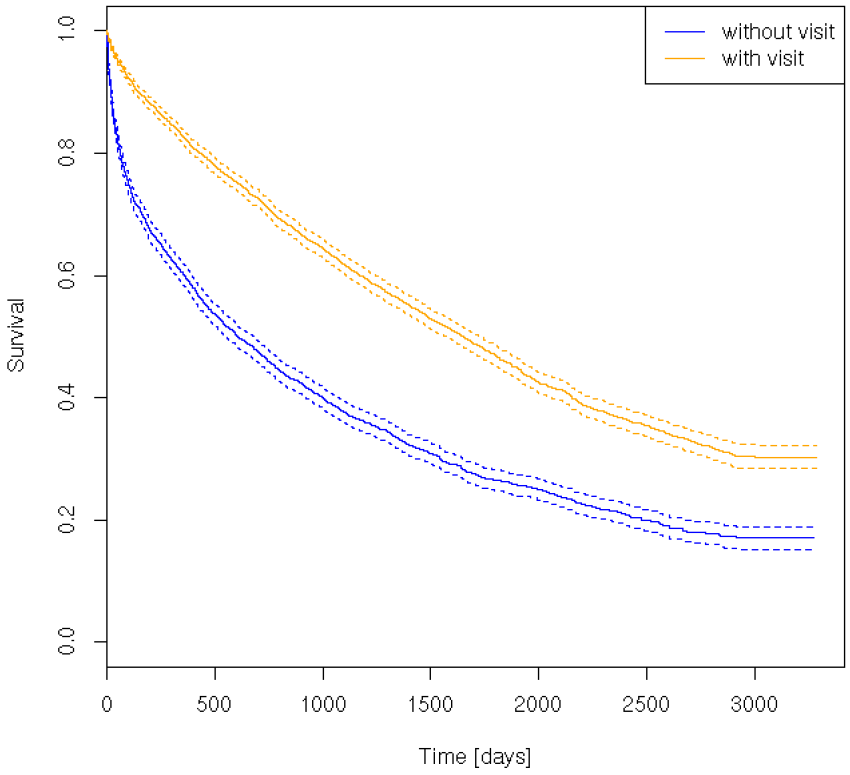
Figure 3.17: KM curves of survival time stratified by Charlson index of patients at their first admission: above the mean, below the mean.



In Figure 3.18 survival is stratified according to the presence of at least one pre hospital cardiological evaluation in patient's history. The patients who had at least one of this visits have the survival curve above the patients who hadn't it, the difference between the curves is confirmed by the p-value of the *Log-Rank* test, reported in Table 3.24.
We can observe the same trend also if we consider the presence of this visit only before the first hospitalization.

In Figure 3.19 survival is stratified according to the variable ICU/IHC index, distinguishing between patients who have at least one of these events in all their clinical history or who haven't it.
The behaviour of the curves is twofold: for the first 3 years having an admis-

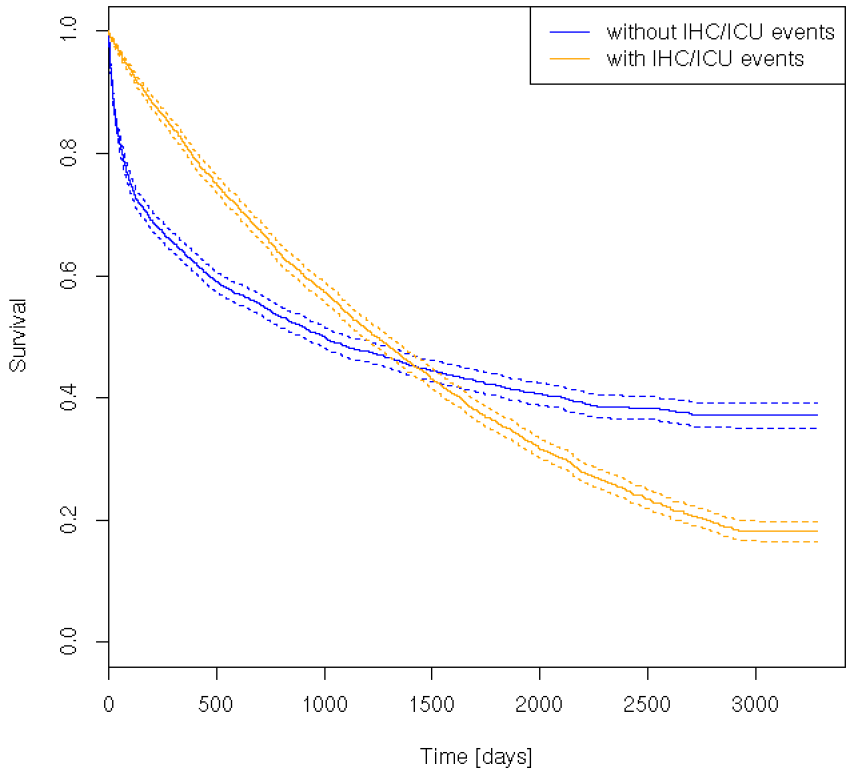
Figure 3.18: KM curves of survival time stratified by the presence of al least one pre hospital cardiological evaluation in patients's clinical history.



sion/activation of ICU/IHC seems to be protective. We could suppose that this variable is usefull in identifying the most frailty patients, whose survival is more at risk.
The p-value of the *Log-Rank* test, reported in Table 3.24 does not allow us to conclude that the two curves are different, due to the intersection.

We can conclude saying that all the considered variables turn out to be a discriminating factor for the estimation of the survival probability, as expected from observing their significance in the Cox model. The only exception is the ICU/IHC index, however we believe it is significant in identifying the most fragile patients's subgroup.

Figure 3.19: KM curves of survival time stratified by the presence of al least one ICU admission or IHC activation in patients's clinical history.



Variable	p-value
Age	<2.2e-16
Sex	6.62e-07
Charlson index	<2.2e-16
Pre hosp card eval	<2.2e-16
ICU/IHC index	0.329

Table 3.24: P-value of the *Log-Rank* test for the different survival curves esitimates created with the Kaplan-Meier estimator.

3.4.2 KM curves on time to second hospitalization

We now focus only on the second hospitalization. We consider only the time needed for the possible second admission to hospital: the *time* variable is now the time spent between the discharge from the first hospitalization and the second admission to hospital, if it happens.
Patients who die after the first admission and before the first discharge, as well as patients who die after the first discharge and before the second admission, are not considered for the creation of these curves, as explained and sketched in Figure 3.20.
The variables considered for the stratifications of the curves are the same ones used for the previous analysis. For this reason we report now only the significant plots.
An important thing to note is the difference in the asymptotic value between the plots about survival times and the plots about times to second rehospitalization: the former have a bigger asymptotic mean value. This is due to the fact that peolple who die at the end of the study are 65% of the population, while people who experience second hospitalization are 78% and the asymptotic value expresses the percentage of patients who not experience the event of interest, death in the previous plots and second hospitalization in these plots.

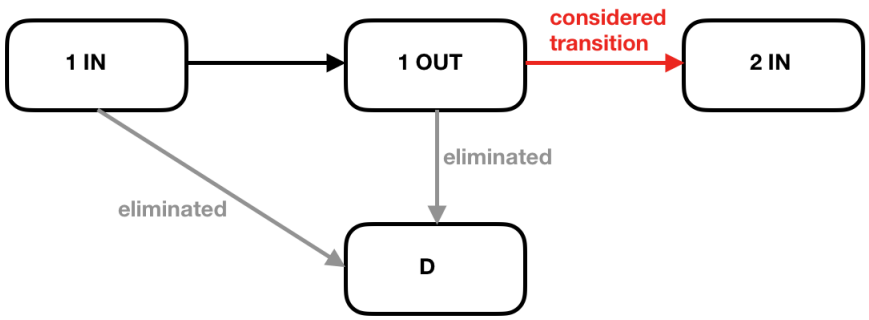
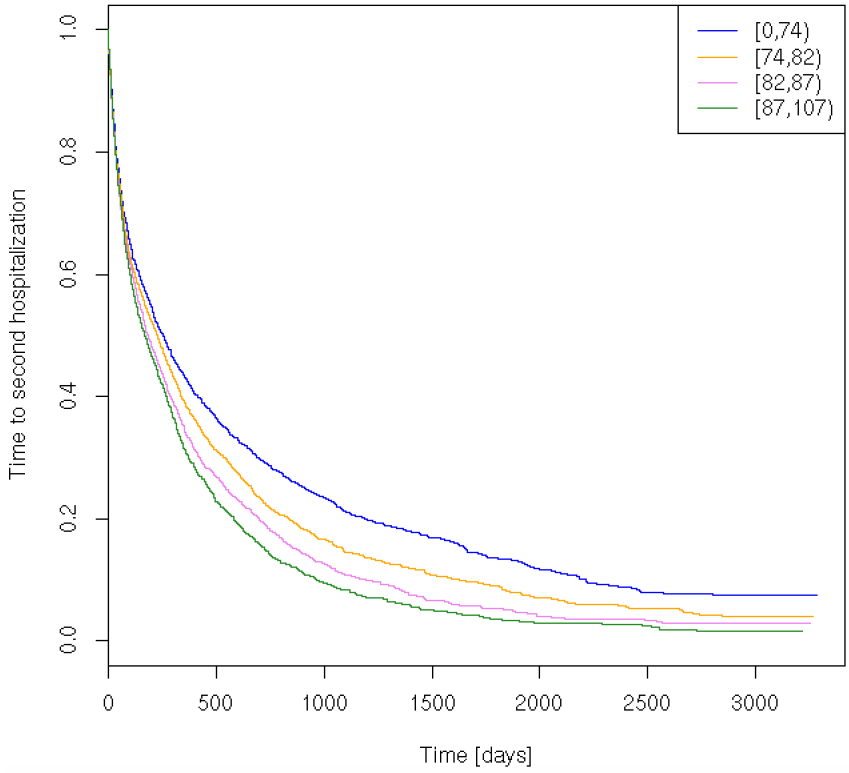


Figure 3.20: Sketch of considered patients for the creation of Kaplan-Meier curves about the survival to second hospitalization.

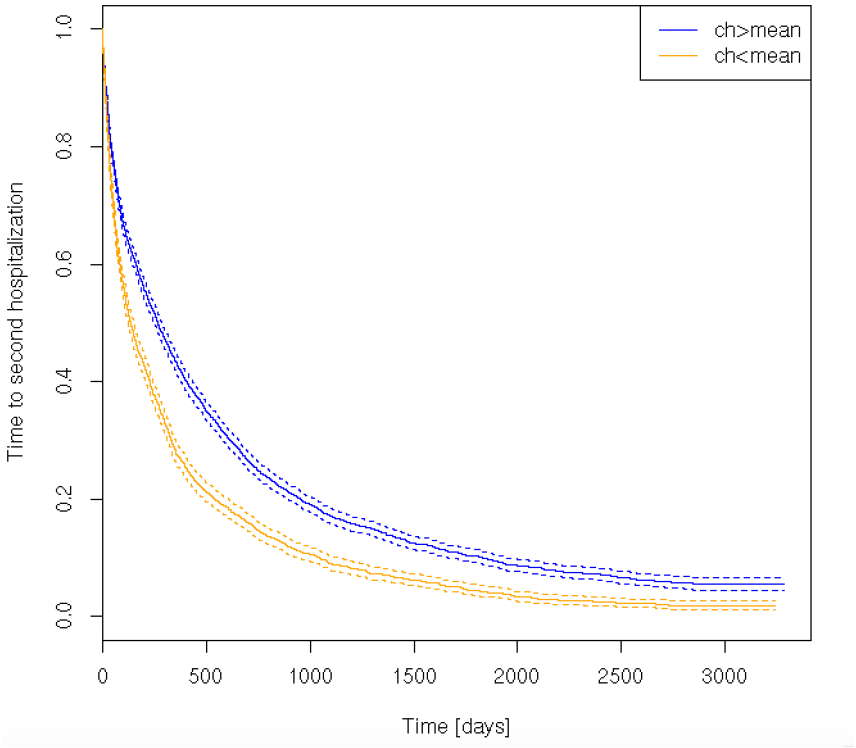
In Figure 3.21 survival is stratified according to the age of the patients at first admission. As in the previous analysis, we categorized the age variable using the empirical quartiles, i.e. 74, 82 and 87.
The difference between the curves related to different levels of the variables are less evident than in the corresponding previuos plot; nevertheless the p-value of the *Log-Rank* test, reported in Table 3.25, is very low and it allows us to conclude that aging is a factor of risk. We omit the confidence intervals of the curves in order to have a clear visibility of the differences.

Figure 3.21: KM curves stratified by age.



In Figure 3.22 survival is stratified according to the Charlson index of the patients computedat first admission. We create again two levels: above the mean (2.37) and under the mean. The two curves behave as we described before, having a bigger Charlson index is a factor of risk, because the probabilities of being readmitted in hospital raise. This behaviour is confirmed by the p-value of the *Log-Rank* test, reported in Table 3.25.

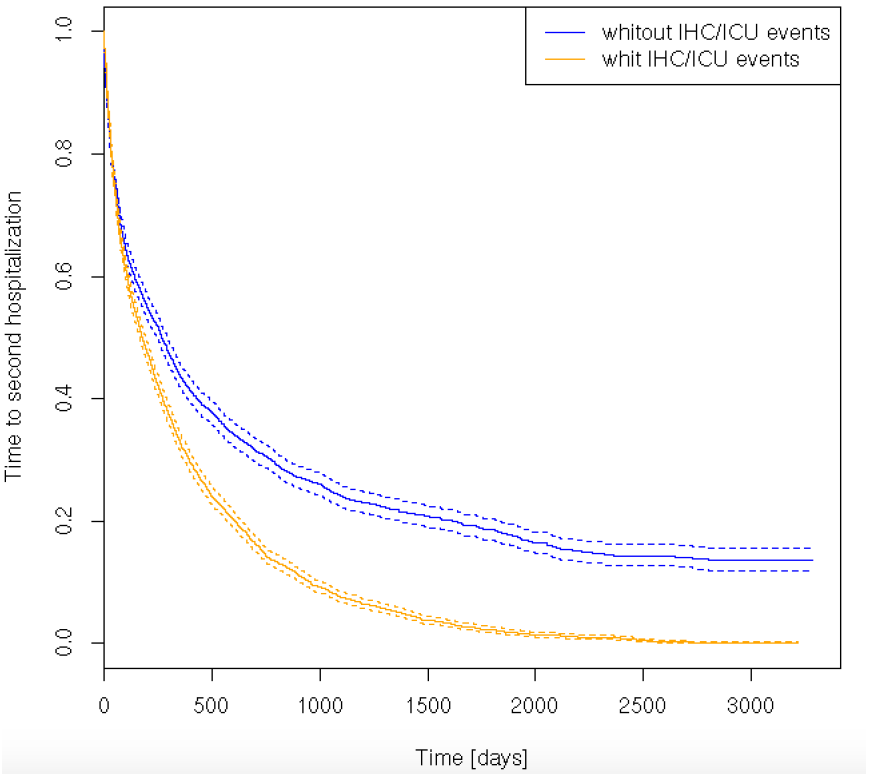
Figure 3.22: KM curves stratified by Charlson index at first hospitalization: above the mean, below the mean.



In figure 3.23 survival is stratified according to the ICU/IHC index, differentiating between patients that have one of these events before first admission and patients that haven't it. We remind the having this index active at first hospitalization is possible only for patients whose history starts before 2009 and consequently has to be cleaned from events preceding that year. The patients in this situation are a small percentage of the total population. The low p-value of the *Log-Rank* test, reported in Table 3.25, express their bigger probability to experience a rehospitalization. This fact seems reasonable because this kind of patients has a longer clinical history and consequently they should have a worse health.

We can conclude saying that all the considered variables are discriminating factors in the calculation of the probability of experiencing a second hospitalization.

Figure 3.23: KM curves stratified by the presence of at least one events of type ICU or IHC at first hospitalization.



Variable	p-value
Age	<2.2e-16
Charlson index	<2.2e-16
ICU/IHC index	<2.2e-16

Table 3.25: P-value of the *Log-Rank* test for the difference between survival curves about second hospitalization, created with the Kaplan-Meier estimator.

3.5 Comparison among Overall Data-OD and De Novo data-DN

In this section we compare patients who experienced an ICU/IHC event with patients who did not; we make this comparison both for the OD and the DN dataset.

Summaries of these population subgroups are shown in Table 3.26.
Looking at the complete dataset we can observe that patients with ICU/IHC index active present a higher commorbidity load than the other cohort: in fact the 95% (vs 86%) has Charlson index greater than zero, with mean value of 2.57 (vs 2.31). The 93% (vs 78%) has more than one morbidity. The mean number of comorbidities at last hospitalization is 1.1 (vs 1.01). The number of total hospitalizations recorded is 5.33 (vs 2.41).
This indicates that having experienced at least once an ICU/IHC event allows to classify a patient as a fragile subject.
To confirm this hypothesis of fragility, we can note that among patients with ICU/IHC events 2,623 deaths are recorded (72%), against the 1,910 (57%) recorded between patients with hospitalization events only.
The same trend can be observed in the dataset containing only *de novo* patients.

	OD		DN	
	ICU/ICH=0	ICU/ICH>0	ICU/ICH=0	ICU/ICH>0
Num of patients	3,377	3,621	2,699	2,358
Death	1,910 (57.23%)	2,623 (72.43%)	1,407 (52.13%)	1,571 (66.62%)
In hosp death	1,058 (55.39%)	1,730 (65.95%)	752 (53.44%)	1,032 (65.69%)
Ch ind>0	2,882 (86.36%)	3,459 (95.52%)	2,276 (84.32%)	2,224 (94.31%)
Mean Ch ind (sd)	2.31 (2.19)	2.57 (2.18)	2.07 (2.02)	2.12 (1.93)
Com>0	2,622 (78.57%)	3,370 (93.06%)	2,110 (78.17%)	2,192 (92.96%)
Com mean	1.03	1.12	1.03	1.07
Com at first hosp	1.01	1.02	1.03	1.03
Com at last hosp	1.01	1.10	0.99	1.04
Hospitalizations	2.41	5.33	2.35	5.06
Hosp/ICU/IHC		8.79		8.23

Table 3.26: Summaries of population subgroups: OD vs DN, ICU/IHC index active vs ICU/IHC index inactive.

Chapter 4

Analysis of Friuli Venezia Giulia dataset

In this chapter we will report the analysis of the Friuli Venezia Giulia dataset, in particular we will present an application of the Non Parametric Discrete Frailty Cox model (npdf Cox), described in Chapter 2, and an application of the multi-state model, described in Chapter 1.
In Section 4.1 we will introduce the dataset and we will report the descriptive analysis. In Section 4.2 we will analyze the homogeneity of the residence districts while in Section 4.3 we will analyze the homogeneity of the cohort. In Section 4.4 we will perform a multi-state modelling of the Friuli Venezia Giulia data.

4.1 Presentation of dataset from Friuli Venezia Giulia

In this section we will introduce the dataset, we will explain the variables we decided to exctract from the dataset for the analysis and we will describe the preprocessing of data carried out on the dataset.

The dataset has the same structure of the one used in Chapter 3 and described in Section 3.1, since it is the extension of the Trieste dataset to all the Friuli Venezia Giulia region.

It is composed by informations about 26,303 patients, identified by an univocal anonymous personal code, hospitalized with HF in the Friuli Venezia Giulia region.
The cohort considered is composed of patients hospitalized between 2009 and 2017. The five-year period from 2004 to 2008 was used for the calculation of significant clinical quantities.
Each row of the dataset refers to a specific event. Possible events are:

- hospitalization for HF;
- hospitalization for any cause;
- Intermediate Care Unit admission (ICU);
- Integrated Home Care (IHC) activations.

Several patient specific informations are recorded for each event: gender, age, length of stay, department of admission, presence of cardiological evaluation before hospitalization, laboratory tests, comorbidities, residence district, hospital where the hospitalization takes place etc..

4.1.1 Variables description

The variables used for the analysis are:

- **Sex**: 1 for male, 0 for female;
- **Age** [years]: time difference between the considered event and the date of birth;
- **Worsening index**: 1 if patient has an hospitalization for HF in the five years preceding the index admission (worsening patient), 0 otherwise (*de novo* patient);
- **Charlson index**: index of comorbidity;
- **Pre hospitalization cardiological evaluation**: 1 if the patient has an hospitalization in cardiology before the considered event, 0 otherwise;
- **Admission in Cardiological Ward (CW)**: 1 if patient is admitted in a cardiological ward, 0 otherwise;

- **ICU/IHC index**: 1 if patient have at least one events of type ICU or IHC before the considered event, 0 otherwise;
- **Residence district**: a number between 1 and 20 indicating the residence district where the patient lives;
- **Death index**: 1 if patient die before the end of the study, 0 otherwise.

We decide to consider the residence district instead of the hospital where the patient is admitted since it is a more robust indicator. Indeed, it is possible that a patient visits different hospitals during his clinical history while it is less probable that he changes residence district, especially considering that the cohort is composed of very old patients.

4.1.2 Dataset transformation

We remind that our dataset is in the format of one row for every patient's event. In order to implement both the npdf Cox model and the multi-state model we have to reshape the dataset, in two different ways.

In order to implement the npdf Cox model, see Section 4.2, we reshape the dataset obtaining the *one row for patient* dataset. This dataset, used in general to fit survival models, contains one row for each patient and as many columns as many informations are collected for each patient. Since we start with one row for each event and we need one row for each patient, we decide to synthesize the informations gathered in the different hospitalizations reporting only informations computed at first admission or at last admission. In particular age, Charlson index, pre hospital cardiological evaluation, worsening index and admission in CW are computed at first admission while ICU/IHC index is computed at last admission. Sex, residence district and death index are time independent variables.

In order to implement the the multi-state model, see Section 4.4, we reshape the dataset obtaining the *long format* dataset. This transformation is explained in Section 3.1.3.

4.1.3 Descriptive analysis

In this section we will report the descriptive analysis of the patients in the Friuli Venezia Giulia dataset.
In Table 4.1 we can find a summary of the population features.

Variable name	# patients	% patients
N	26,303	
Male	11,937	45.38%
Age	m=81.82	sd=9.6
DeNovo	23,646	89.89%
Deaths	16,625	63.2%
Survival time	m=897.9	sd= 856.2
Pre hosp card visit	14,540	55.14%
CW	2,252	8.56%
Charlson index	m=2.93	sd=1.96
Ch >0	24,013	91.29%
ICU/IHC >0	10,127	38.5%
ICU >0	4,424	16.81%
IHC >0	8,726	33.17%
Cancer	2,537	9.64%
Pulmonary desease	4,863	18.48%
Diabete	13,035	49.55%
Renal desase	15,042	57.18%

Table 4.1: Summary of the main population features.

The patients considered are 26,303. Among these 11,937 (45%) are males. 23,646 patients (90% of the total) did not have a HF hospitalization in the five years preceding the index admission (*de novo* patients). The mean age at first hospitalization is 81.82 years, the corresponding standard deviation is 9.6. The mean survival time is 897.9 days, the corresponding standard deviation is 856.2. Over the observing period 16,625 deaths (63%) are recorded.

ID	District name	# patients	% patients
1	Valmaura	1,302	4.95%
2	Via Stock	1,104	4.19%
3	Alto Isontino	1,463	5.56%
4	Cervignano del Friuli	1,213	4.61%
5	Codroipo	1,053	4%
6	San Daniele	999	3.79%
7	Cividale del Friuli	1,029	3.91%
8	Udine	3,283	12.48%
9	Maniago (Nord)	1,057	4.01%
10	Sacile (Ovest)	1,261	4.79%
11	San Giovanni	1,149	4.36%
12	Via della Pieta	1,134	4.31%
13	Basso Isontino	1,565	5.94%
14	Latisana	1,133	4.3%
15	Gemona del Friuli	928	3.52%
16	Tolmezo	904	3.43%
17	Tarcento	996	3.78%
18	Azzano decimo (Sud)	1,100	4.18%
19	Pordenone	1,913	7.27%
20	San Vito al Tagliamento	827	3.14%
	N.D.	890	3.36%

Table 4.2: List of all the Friuli Venezia Giulia residence districts together with the numer and the percentage of patients living there.

Focusing on indices at first hospitalization, there is a prevalence of non-cardiac comorbidities in patient's background (18% has pulmonary disease, 9% cancer, 49% diabetes and 57% renal disease). The mean Charlson index of 2.93 denotes an high level of comorbidity burden. The 91% of patients has this index grater that zero. The rate of admission in CW is 9%. Finally, 55% of patients has a cardiological pre hospitalization visit.

We analyse the frequency of ICU/IHC activations looking at variable ICU/ICH index at last hospitaliziation: in their history, 38% of patients ex-

perienced at least one of this events. In particular, 16% of patients had an ICU transition and 33% had an IHC activation.

In Table 4.2 we can find all the residence districts together with the number and the percentage of patients living there. We can note that the patients are equally distributed in the different districts. In every district we can find about 3%-5% of the patients, with the exception of the Udine district (12%) and of the Pordenone district (7%).

4.2 Analysis of the homogeneity of the residence districts

In this section we will report the application of the npdf Cox model with residence distric specific frailty on the Friuli Venezia Giulia dataset.
Our goal is to evaluate if, after specific patient and specific procedure adjustments, the residence distric, see Table 4.2 for all the possible residence districts, has some further measurable influence that our covariates can not explain.

We want to clarify that, for this specific analysis, we only consider patients whose information about the residence district is present. Hence, as reported in Table 4.2, we do not include 890 patients (3.36% of the total population) in the following analysis.

First of all we compare the Kaplan-Meier estimates of patients's survival, stratified according to all the possible residence districts. These estimates are reported in Figure 4.1. We can notice that the curves have a similar trend and do not differ from each other, being almost overlapped. We can suppose that the difference in hospital treatment among all the residence districts is minimal so we hypothesize the homogeneity of the residence districts.

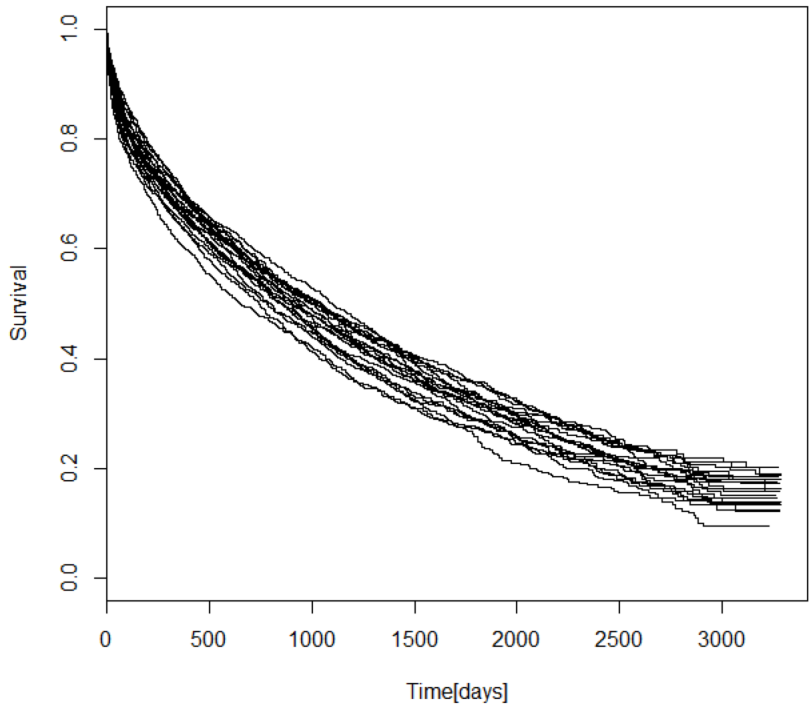


Figure 4.1: Kaplan-Meier estimates of the survival stratified according to the residence district.

We proceed by analyzing the variables distribution among the twenty different residence districts, reported in Table 4.2. In this way we can see if any particular behaviour of the variables is present. We report the comparison in Table 4.3 and 4.4. We can notice that there are not big differences among the variables distributions between the residence districts. In every residence district, the percentage of people presenting the characteristics reported in Table 4.3 and 4.4 is about the same as if we consider the entire population, as reported in Table 4.1.
This absence of huge differences is in line with our first hypothesis, deduced from the Kaplan-Meier estimates in Figure 4.1, of homogeneity among the residence districts.

In order to confirm that the residence district has no influence on the survival we decide to implement the npdf Cox model with residence district specific frailty, described in Chapter 2. In this way we can verify, in a

ID	N	Male	Age	Denovo	Deaths	Survival time
1	1,302	44%	81.8(8.6)	87%	63%	965.6(687)
2	1,104	45%	82.8(8.2)	86%	66%	912.6(826)
3	1,463	46%	81.3(10.3)	92%	65%	782.4(758)
4	1,213	46%	81.7(9.6)	87%	66%	902.6(877)
5	1,053	46%	82.0(9.3)	89%	66%	840.6(856)
6	999	43%	82.7(9.3)	88%	68%	770.9(774)
7	1,029	48%	81.2(9.1)	89%	61%	927(854)
8	3,283	45%	82.1(9.4)	89%	63%	884.9(851)
9	1,077	45%	82.5(9.6)	90%	66%	727(791)
10	1,261	43%	82.4(9.3)	90%	64%	869(859)
11	1,149	40%	83.7(8.5)	87%	70%	869(823)
12	1,134	36%	83.6(8.7)	88%	69%	887.5(835)
13	1,565	49%	80.1(10.5)	90%	60%	927.3(874)
14	1,133	48%	81.4(9.7)	89%	63%	887.6(846)
15	928	43%	81.1(9.3)	89%	65%	919.4(858)
16	904	49%	80.1(9.6)	91%	61%	915.2(838)
17	996	45%	81.9(9.2)	90%	67%	900.9(845)
18	1,100	45%	81.4(10.0)	90%	61%	922.3(874)
19	1,913	44%	82.4(9.4)	91%	60%	915.4(871)
20	827	41%	82.1(9.7)	92%	61%	876.1(842)

Table 4.3: Main features comparison among patients living in different residence districts.

modellistic and quantitative way, if the available covariates are able to catch almost all the present variability. This kind of model takes advantage of the nonparametric and discrete distribution of the frailty term in order to build a probabilistique clustering tecnique with whom we can investigate the influence of the residence districts on the survival.

We built five different models, one for each potential number of clusters K, from 1 to 5, and then we compute model selection criteria such as AIC, BIC, or search for the optimal K using the approach proposed by Laird [22]. We always consider seven individual-level predictors: age, sex, worsening index, Charlson index, pre hospitalization cardiological evaluation, admission in CW, ICU/IHC index.

ID	Pre hosp card eval	CW	Ch index	Ch index >0	ICU/IHC>0
1	56%	4%	2.9(2.1)	89%	47%
2	55%	5%	2.9(1.9)	89%	51%
3	61%	29%	3.3(1.8)	99%	30%
4	62%	5%	2.7(1.7)	90%	39%
5	59%	3%	2.9(1.8)	91%	43%
6	54%	2%	2.7(1.9)	89%	34%
7	64%	3%	2.8(2.0)	89%	43%
8	58%	3%	2.9(1.9)	90%	36%
9	51%	4%	3.3(2.1)	93%	32%
10	54%	11%	2.3(2.0)	83%	37%
11	51%	4%	2.9(2.0)	90%	48%
12	44%	4%	2.8(2.1)	87%	46%
13	60%	33%	3.4(1.8)	99%	32%
14	62%	2%	2.9(1.8)	92%	41%
15	57%	1%	3.3(1.8)	98%	45%
16	59%	2%	3.3(2.8)	99%	45%
17	62%	3%	3.1(1.8)	93%	40%
18	46%	9%	2.6(2.0)	86%	39%
19	42%	11%	2.3(1.9)	82%	37%
20	61%	3%	3.4(1.9)	97%	31%

Table 4.4: Main features comparison among patients living in different residence districts.

Criterion	K optimum
AIC	1
BIC	1
Laird	1

Table 4.5: Results of the model selection criteria.

The model selection criteria, as reported in Table 4.5, are all consistent: the optimum number of clusters is 1. We can conclude that the covariates have caught almost all the heterogeneity present in the residence districts that, hence, have no further influence on the survival.

Parameters	$\hat{\beta}$	HR	se Louis
Age	0.062	1.064	0.001
Sex	0.213	1.237	0.016
Charlson index	0.111	1.117	0.004
Pre hosp card visit	-0.063	0.939	0.016
Worsening index	0.342	1.408	0.023
CW	-0.342	0.710	0.037
ICU/IHC	-0.307	0.735	0.016

Table 4.6: Estimate of the parameters ($\hat{\beta}$ and HR) of the npdf Cox model with no latent populations, together with Louis standard errors.

We report the results of the npdf Cox model for K=1 in Table 4.6.
The HR of variable age is 1.064, this means that being older increases the probability of dying.
The HR of variable sex is 1.237, this means that being a woman increases the probability of dying.
The HR of variable Charlson index is 1.117, this means that having a bigger index of comorbidity increases the probability of dying.
The HR of variable pre hospitalization cardiological evealuation is 0.939, this means that having this kind of pre hospitalization decreases the probability of dying.
The HR of variable worsening index is 1.408, this means that being a worsening patient increases the probability of dying.
The HR of variable CW is 0.710, this means that being admitted in a CW decreases the probability of dying.
The HR of variable ICU/IHC index is 0.735, this means that experience at least one of these events decreases the probability of dying.

We decide to compare the results of the npdf Cox together with the results of the traditional Cox model without the frailty term and with a traditional Cox model where Gamma and Normal frailties are specified for the district term. The results of the different Cox models are reported in Table

4.7. We can notice that the coefficients estimates of the three Cox models reported in Table 4.7 are very similar to the ones of the npdf Cox model reported in Table 4.6. The coefficients referring to the same variable have all the same sign in the different models, meaning that in every model the variables have the same effect on the outcome.
For the Cox frailty models, both with Gamma distribution and Normal distribution, we can note that the variance of random effect is very low: 0.005 for the first model and 0.003 for the second one. This indicates that the residual variability, not explained by the covariates, is very low, in agreement with the fact that only one cluster is significant. The residence districs can hence be considered homogeneous, after specific patient and specific procedure adjustments. This means that we have no evidence to conclude that the clinical treatment is different among the districts of the Friuli Venezia Giulia region.

	Cox		Cox frailty Gamma		Cox frailty Normal	
Parameters	HR	se exact	HR	se exact	HR	se exact
Age	1.064	0.001	1.063	0.001	1.063	0.001
Sex	1.237	0.016	1.238	0.016	1.238	0.016
Charlson ind.	1.117	1.116	0.110	1.116	0.110	0.004
Pre hosp c.v.	0.937	0.934	-0.068	0.935	-0.067	0.016
Worsening ind.	1.322	1.329	0.285	1.328	0.284	0.023
CW	0.709	0.037	0.696	0.038	0.697	0.037
ICU/IHC	0.735	0.016	0.737	0.016	0.737	0.016
Variance of random effect			0.005		0.003	

Table 4.7: Parameters estimates (HR) together with exact standard errors of the simple Cox model, hence without frailty term, of the Cox model with residence district specific Gamma distributed frailty and of the Cox model with residence district specific Normal distributed frailty. For frailty models we reported also the variance of random effects.

4.3 Analysis of the homogeneity of the cohort

In this section we will analyze the homogeneity of the cohort and the possible presence of significant subgroups. In this context, homogeneity means that the statistical properties of any subset of the overall dataset are the same as those of the latter, hence can not be found groups of patients characterized by strongly different features.
In Section 4.3.1 we will implement Cox models with different parametric patient specific frailty, in order to evaluate the importance of this patient specific adjustment, and we will discuss the impossibility to apply the npdf Cox model with patient specific frailty. In Section 4.3.2, we will apply the k-means algorithm in order to analyze if we can cluster our cohort at least according to the covariates.

4.3.1 Parametric frailty Cox models

In order to evaluate the impact of a patient specific adjustment, hence if the patient specific frailty is a significant term that can catch and explain a part of variability or if the covariates are sufficiently to explain it, we decide to compare the results of the simple Cox model, of the Cox model with patient specific frailty with Gamma distribution and of the Cox model with patient specific frailty with Normal distribution. The results of the different Cox models are reported in Table 4.8.
First of all, we can notice that the coefficients estimates referring to the same variable of the three Cox models are quite similar and, in particular, they have all the same sign, meaning that, in every model, the variables have the same effect on the outcome.
We can also notice that the estimated variance of random effect is quite low: 0.617 for the Cox model with Gamma distributed frailty and 0.431 for the Cox model with Normal distributed frailty. This means that the residual variability, hence not explained by the covariates, is not so high and that our cohort could be considered homogeneous with respect to it.

In order to confirm that the variability that the covariates can not ex-

plain is low, we would like to apply the npdf Cox model with patient specific frailty. Indeed, this model, thanks to the discrete distribution of the frailty, see Section 2.2, could allow us to evaluate if our cohort could be partitioned in some subgroups, according to the residual variability.
Nevertheless, as we show in Appendix A, we can not trust the results of the npdf Cox model with patient specific frailty. We would have 26,303 groups (that correspond to the number of the patients in our cohort) made by only one unit and our simulation study highlights how, in such stressed situation, the npdf Cox algorithm could not estimate the correct number of latent populations.

	Cox		Cox frailty Gamma		Cox frailty Normal	
Parameters	HR	se exact	HR	se exact	HR	se exact
Age	1.064	0.001	1.080	0.001	1.073	0.001
Sex	1.237	0.016	1.295	0.017	1.272	0.016
Charlson ind.	1.117	0.004	1.164	0.004	1.144	0.004
Pre hosp c.v.	0.937	0.016	0.906	0.016	0.920	0.016
Worsening ind.	1.322	0.023	1.421	0.031	1.382	0.023
CW	0.709	0.037	0.640	0.044	0.673	0.037
ICU/IHC ind.	0.735	0.016	0.567	0.021	0.638	0.016
Variance of random effect			0.617		0.431	

Table 4.8: Parameters estimates (HR) together with exact standard errors of the simple Cox model, hence without frailty term, of the Cox model with patient specific Gamma distributed frailty and of the Cox model with patient specific Normal distributed frailty. For frailty models we reported also the variance of random effects.

Since we can not apply the npdf Cox model in order to detect a possible clustering structure of our cohort, we decide to investigate it using the collected informations. Hence, we choose to apply the k-means algorithm on the available measured covariates.

4.3.2 K-means algorithm

As we explained in Section 4.3.1, in order to analyze the possible clustering of patients according to the covariates, we decide to proceed with the application of the k-means algorithm.

Firstly we use the "elbow" method [21] in order to identify an optimal number of clusters.
This method runs a k-means clustering on the dataset for a range of values of k (in our case, as we can see in Figure 4.2, from 1 to 10). For each value of k it calculate the between (BSS) and within (WSS) sum of square and compute the ratio BSS/(BSS+WSS). This ratio is a measure of the percentage of varibility explained. We look at it as a function of the number of clusters. We should choose a number of clusters so that the variability explained is sufficiently high and has not a significant increase after the addition of another cluster. A threshold value for the percentage of variability explained could be 80%. As we can see in Figure 4.2, at some point the marginal gain will drop, giving an angle in the graph of the function. The number of clusters is chosen at this point. In Figure 4.2 we can find the "elbow" function.

We choose a number of clusters k=4, since adding another cluster does not highly increase the percentage of variance explained, that is over the 80% yet.
Once the number of clusters is decided, we analyze the distribution of the available variables in the four different groups identified.
We report these comparisons in Table 4.9, where the p-value refers to the tests on the proportions, for binary covariates, and to the Kruskal-Wallis tests, for continuous covariates.

Looking at the p-values in the last column, we can immediately notice that all the variables are significantly differently distributed among the four groups. In particular the different percentage of deaths and the different mean survival times seem characterize the four groups.
For example, Group 1 is characterized by the highest mean survival times (1372 days) and the lowest percentage of deaths recorded (34%), and, coherently with the results obtained in Section 3.3.2, it also presents the lowest

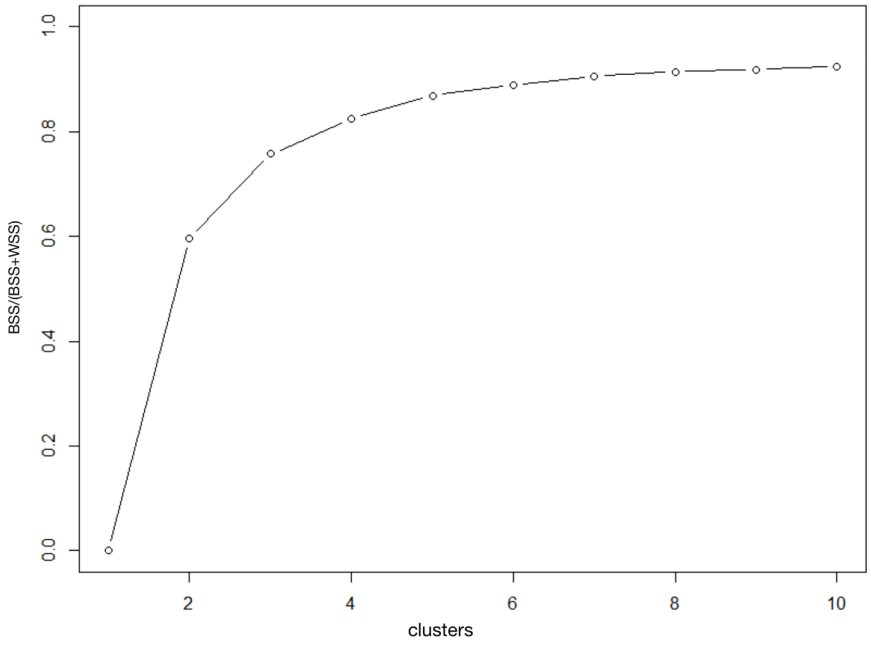


Figure 4.2: "Elbow" function: on the x-axis there is the number of cluster considered while on the y-axis there is the ratio BSS/(BSS+WSS), a measure of the percentage of varibility explained by the corresponding clusters.

mean age (62.04 years) and the highest percentage of men (70%).
On the contrary, Group 4 is characterized by the lowest mean survival times (500 days) and the highest percentage of deaths recorded (80%), and, coherently, it also presents the highest mean age (93.02 years) and the lowest percentage of men (25%).
In Group 2 and 3 are intermediate results.

In Figure 4.3 we can find the Kaplan-Meier estimate of the survival stratified according to the four groups identified through the k-means.
We can see that the four groups have survival curves with different behaviors, that reflect the different mean survival times reported in Table 4.9. The different survival of the groups is confirmed by the p-value of the Log-Rank test, that is <2.2e-15.

	Group 1	Group 2	Group 3	Group 4	P-value
Deaths	34%	52%	70%	80%	<2.2e-16
Mean survival time(sd)	1372(973)	1112(907)	803(782)	500(592)	<2.2e-16
N	2,723	8,232	10,044	5,304	
Male	70%	57%	39%	25%	<2.2e-16
Mean age(sd)	62.04(7.08)	76.5(3.3)	85.6(2.2)	93.02(2.9)	<2.2e-16
DeNovo	92%	89%	88%	91%	3.03e-08
Visit pre hosp.	51%	63%	56%	42%	<2.2e-16
CW	28%	11%	4%	1%	<2.2e-16
Mean Charlson index(sd)	2.51(2)	3(2.13)	3(1.9)	2.9(1.7)	<2.2e-16
ICU/IHC >0	29%	40%	42%	32%	<2.2e-16

Table 4.9: Main features comparison among the four groups identified thorugh k-means algortithm together with the p-values of tests on the proportion, for binary covariates, and of Kruskal-Wallis tests, for continuous covariates.

In order to better analyze the difference between the four groups identified by the k-means algorithm, we fit a simple Cox model on each group. The estimates of the coefficients variables togheter with the relative standard errors and the p-values of the significance tests are reported in Table 4.10 and in Table 4.11.

We can note that Group 1 (i.e. the youngest group, since it contains the patients with the lowest mean age), is the group with the highest, nevertheless significant, p-values, hence, there is less evidence in the significance of the coefficients. We can explain this fact supposing that being hospitalized at a "young" age is a relevant factor that, once taken into account, puts the other factors in the background.

Moreover, in all the four groups, the coefficients of variables age, Charlson index and worsening index are significant and all positive. This means that being older as well as having a bigger Charlson index as well as being a worsening patient increases the probability of death. These results are consistent with the ones obtained in Section 3.3.2.

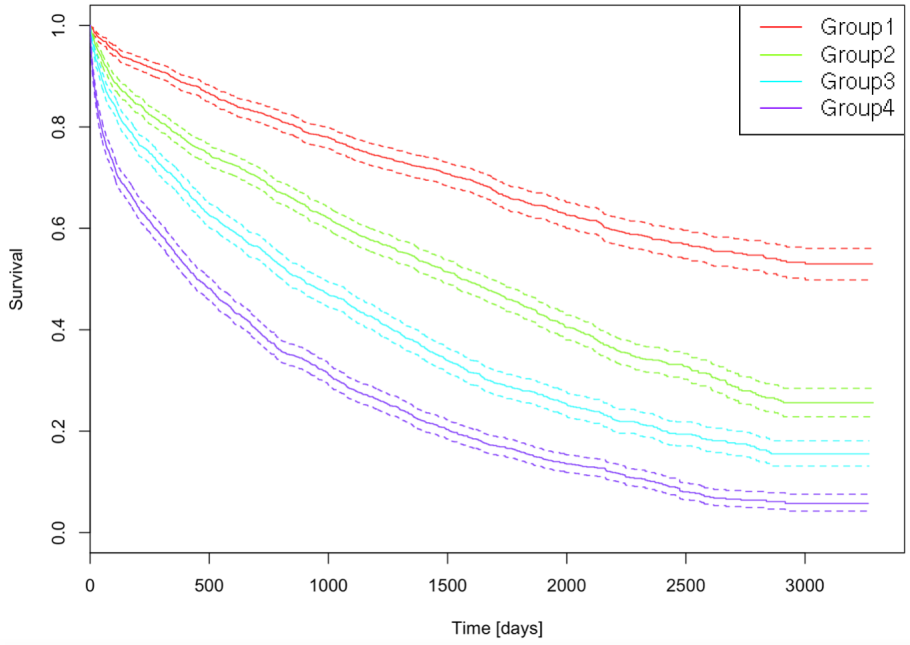


Figure 4.3: Kaplan-Meier estimate of the survival stratified according to the four groups identified thorugh the k-means algorithm.

Looking at the estimates of the ICU/IHC variable coefficients we can note that the significant ones have different sign in different groups. In particular the coefficient is positive in Group 1, the youngest group, and is negative in Group 3 and 4, that are the older groups. This means that experience an ICU or an IHC event is a risk factor for the younger patients while it is a protective factor for the older patients. Indeed, generally, if a young patient has an ICU admission or a IHC activation means that he has a complicated clinical situation. Instead, if an old patient has an ICU admission or a IHC activation implies that he will receive more specific and careful care.
This different behavior is reflected in the Kaplan-Meier estimate of the survival stratified by the presence of at least one ICU/IHC events in patient's clinical history, reported in Figure 3.19, where we can note that the two curves intersect. The intersection can be justified by these reseults: before the intersection, where in general we can find the mean survival time of the older patients, experiencing an ICU/IHC event is a protective factor, after the intersection, where in general we can find the mean survival time of the younger patients, experiencing an ICU/IHC event becomes a risk factor.

	Group 1 34% deaths			Group 2 52% deaths		
Parameters	HR	se	p-value	HR	se	p-value
Age	1.032	0.005	3.11e-08	1.058	0.004	<2e-16
Sex	1.201	0.074	0.014	1.278	0.031	6.55e-15
Charlson index	1.217	0.015	<2e-16	1.144	0.006	<2e-16
Pre hosp card visit	1.204	0.067	0.005	1.031	0.031	0.323
Worsening index	1.312	0.094	0.004	1.423	0.042	<2e-16
CW	0.711	0.088	0.0001	0.733	0.056	2.95e-08
ICU/IHC index	1.275	0.069	0.0004	1.021	0.031	0.498

Table 4.10: Estimates of the coefficients variables (HR) togheter with the relative standard errors and the p-values of the significance tests obtained fitting the simple Cox model on the first two groups identified thorugh the k-means algorithm.

	Group 3 70% deaths			Group 4 80% deaths		
Parameters	HR	se	p-value	HR	se	p-value
Age	1.072	0.005	<2e-16	1.052	0.005	<2e-16
Sex	1.236	0.024	<2e-16	1.155	0.035	4.95e-05
Charlson index	1.105	0.006	<2e-16	1.073	0.008	<2e-16
Pre hosp card visit	0.938	0.024	0.007	0.967	0.032	0.292
Worsening index	1.291	0.034	9.66e-15	1.104	0.050	0.050
CW	0.757	0.064	1.54e-05	0.746	0.130	0.024
ICU/IHC index	0.727	0.024	<2e-16	0.620	0.032	<2e-16

Table 4.11: Estimates of the coefficients variables (HR) togheter with the relative standard errors and the p-values of the significance tests obtained fitting the simple Cox model on the last two groups identified thorugh the k-means algorithm.

4.4 Analysis of Friuli Venezia Giulia dataset thorugh multi-state model

In this section we will present the main results obtained from the application of the same multi-state model described in Section 3.3.1 to the Friuli Venezia Giulia dataset.
The only difference between this model and the one described in Section 3.3.1 is that here we do not consider the ICU/IHC index variable in the transitions with ID equal to 1, 2, 3 and 4. Indeed, in the Friuli Venezia Giulia dataset,

any patient has this index active at the beginning of his clinical history. The patients who need it are admitted in ICU only after the first hospitalization and the same happens for the IHC activation.

We want to see if the results obtained using the Friuli Venezia Giulia dataset are similar with the results obtained in Section 3.3.2, where we used a subset of this regional dataset, the Trieste dataset.

In Tables from 4.12 to 4.18, we report the hazard ratio estimates for each pair of variable/transition considered in the Cox model together with the p-value of the significance test of the single HR ($H_0 : exp(\beta_i) = 1$ VS $H_1 : exp(\beta_i) \neq 1$, with i from 1 to p), the number of the asterisks reflects the importance of the coefficient.
In Figures, from 4.5 to 4.10, we can find the plots of the hazard ratio estimates and their corresponding 95% confidence intervals. We consider one variable at a time, separating the different transitions for a better visualization.

Table 4.12: Hazard ratios estimates for the transitions of variable sex, estimated fitting Cox model to Friuli Venezia Giulia dataset.

		HR	Pr(>|z|)	
Hospital discharge	SEX.1	1.082	$< 2e-16$	***
	SEX.5	1.042	0.017	*
	SEX.9	1.019	0.358	
	SEX.13	1.027	0.280	
	SEX.17	1.014	0.637	
In-hospital death	SEX.2	1.222	$< 2e-16$	***
	SEX.6	1.054	0.275	
	SEX.10	1.219	$< 2e-16$	***
	SEX.14	1.143	0.028	*
	SEX.18	1.175	0.031	*
	SEX.21	1.185	$< 2e-16$	***

Table 4.13: Hazard ratios estimates for the transitions of variable age, estimated fitting Cox model to Friuli Venezia Giulia dataset.

		HR	Pr(>|z|)	
Hospital discharge	AGE.1	0.996	$< 2e-16$	***
	AGE.5	0.998	0.066	*
	AGE.9	1.000	0.723	
	AGE.13	1.000	0.901	
	AGE.17	1.003	0.086	*
In-hospital death	AGE.2	1.081	$< 2e-16$	***
	AGE.6	1.058	$< 2e-16$	***
	AGE.10	1.064	$< 2e-16$	***
	AGE.14	1.051	$< 2e-16$	***
	AGE.18	1.069	$< 2e-16$	***
	AGE.21	1.044	$< 2e-16$	***
Admissions to hospital	AGE.3	1.005	$< 2e-16$	***
	AGE.7	1.006	$< 2e-16$	***
	AGE.11	1.005	$< 2e-16$	***
	AGE.15	1.005	0.001	**
	AGE.19	1.003	0.085	*
Out-of-hospital death	AGE.4	1.079	$< 2e-16$	***
	AGE.8	1.068	$< 2e-16$	***
	AGE.12	1.068	$< 2e-16$	***
	AGE.16	1.061	$< 2e-16$	***
	AGE.20	1.055	$< 2e-16$	***

Table 4.14: Hazard ratios estimates for the transitions of variable Charlson index, estimated fitting Cox model to Friuli Venezia Giulia dataset.

		HR	Pr(>|z|)	
Hospital discharge	CHARLSON.1	0.949	$< 2e-16$	***
	CHARLSON.5	0.948	$< 2e-16$	***
	CHARLSON.9	0.961	$< 2e-16$	***
	CHARLSON.13	0.959	$< 2e-16$	***
	CHARLSON.17	0.963	$< 2e-16$	***
In-hospital death	CHARLSON.2	1.096	$< 2e-16$	***
	CHARLSON.6	1.079	$< 2e-16$	***
	CHARLSON.10	1.083	$< 2e-16$	***
	CHARLSON.14	1.075	$< 2e-16$	***
	CHARLSON.18	1.080	$< 2e-16$	***
	CHARLSON.21	1.076	$< 2e-16$	***
Admissions to hospital	CHARLSON.3	1.084	$< 2e-16$	***
	CHARLSON.7	1.072	$< 2e-16$	***
	CHARLSON.11	1.071	$< 2e-16$	***
	CHARLSON.15	1.069	$< 2e-16$	***
	CHARLSON.19	1.073	$< 2e-16$	***
Out-of-hospital death	CHARLSON.4	1.110	$< 2e-16$	***
	CHARLSON.8	1.143	$< 2e-16$	***
	CHARLSON.12	1.130	$< 2e-16$	***
	CHARLSON.16	1.140	$< 2e-16$	***
	CHARLSON.20	1.129	$< 2e-16$	***

Table 4.15: Hazard ratios estimates for the transitions of variable pre hospitalization cardiological evaluation, estimated fitting Cox model to Friuli Venezia Giulia dataset.

		HR	Pr(>|z|)	
In-hospital death	PRE_HOSP.2	0.782	$< 2e-16$	***
	PRE_HOSP.6	0.887	0.010	**
	PRE_HOSP.10	0.955	0.399	
	PRE_HOSP.14	0.836	0.004	**
	PRE_HOSP.18	0.898	0.179	
	PRE_HOSP.21	0.837	$< 2e-16$	***
Out-of-hospital death	PRE_HOSP.4	0.908	0.050	*
	PRE_HOSP.8	0.766	$< 2e-16$	***
	PRE_HOSP.12	0.844	0.013	*
	PRE_HOSP.16	0.787	0.005	**
	PRE_HOSP.20	0.578	$< 2e-16$	***

Table 4.16: Hazard ratios estimates for the transitions of variable CW admission, estimated fitting Cox model to Friuli Venezia Giulia dataset.

		HR	Pr(>|z|)	
Hospital discharge	CW.1	1.174	$< 2e-16$	***
	CW.5	1.203	$< 2e-16$	***
	CW.9	1.217	$< 2e-16$	***
	CW.13	1.187	$< 2e-16$	***
	CW.17	1.318	$< 2e-16$	***
In-hospital death	CW.2	0.840	0.216	
	CW.6	0.411	$< 2e-16$	***
	CW.10	0.495	$< 2e-16$	***
	CW.14	0.646	0.003	**
	CW.18	0.842	0.340	
	CW.21	0.485	$< 2e-16$	***
Out-of-hospital death	CW.4	0.763	0.031	*
	CW.8	0.425	$< 2e-16$	***
	CW.12	0.486	$< 2e-16$	***
	CW.16	0.442	$< 2e-16$	***
	CW.20	0.440	0.006	**

Table 4.17: Hazard ratios estimates for the transitions of variable worsening index, estimated fitting Cox model to Friuli Venezia Giulia dataset.

		HR	Pr(>|z|)	
In-hospital death	WORSENING.2	1.054	0.455	
	WORSENING.6	0.940	0.352	
	WORSENING.10	1.032	0.653	
	WORSENING.14	0.944	0.463	
	WORSENING.18	0.833	0.052	*
	WORSENING.21	0.835	$< 2e-16$	***
Admissions to hospital	WORSENING.3	1.356	$< 2e-16$	***
	WORSENING.7	1.227	$< 2e-16$	***
	WORSENING.11	1.209	$< 2e-16$	***
	WORSENING.15	1.209	$< 2e-16$	***
	WORSENING.19	1.195	$< 2e-16$	***
Out-of-hospital death	WORSENING.4	0.968	0.686	
	WORSENING.8	0.932	0.396	
	WORSENING.12	0.767	0.010	**
	WORSENING.16	0.892	0.319	
	WORSENING.20	1.025	0.854	

Table 4.18: Hazard ratios estimates for the transitions of variable ICU/IHC index, estimated fitting Cox model to Friuli Venezia Giulia dataset.

		HR	Pr(>|z|)	
Hospital discharge	ICU/IHC.5	0.852	$< 2e-16$	***
	ICU/IHC.9	0.876	$< 2e-16$	***
	ICU/IHC.13	0.891	$< 2e-16$	***
	ICU/IHC.17	0.887	$< 2e-16$	***
In-hospital death	ICU/IHC.6	1.315	$< 2e-16$	***
	ICU/IHC.10	1.218	$< 2e-16$	***
	ICU/IHC.14	1.453	$< 2e-16$	***
	ICU/IHC.18	1.227	0.009	**
	ICU/IHC.21	1.538	$< 2e-16$	***
Admissions to hospital	ICU/IHC.7	1.014	0.531	
	ICU/IHC.11	1.030	0.196	
	ICU/IHC.15	1.009	0.730	
	ICU/IHC.19	0.978	0.498	
Out-of-hospital death	ICU/IHC.8	1.382	$< 2e-16$	***
	ICU/IHC.12	1.459	$< 2e-16$	***
	ICU/IHC.16	1.691	$< 2e-16$	***
	ICU/IHC.20	1.516	$< 2e-16$	***

Figure 4.4: 95% confidence intervals for hazard ratios of sex (all other covariates fixed) estimated fitting Cox model to Friuli Venezia Giulia dataset.

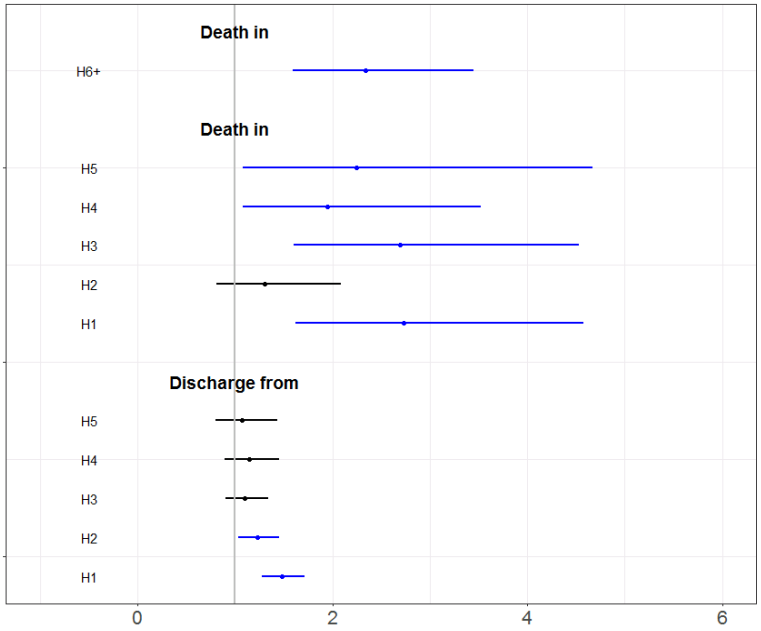


Figure 4.4 shows the confidence intervals for hazard ratios of sex variable. We note that the significant terms are all positive. This indicates that being a man increases the probabilities of being discharged and of dying in hospital.

Figure 4.5 shows the confidence intervals for hazard ratios of age variable. This variable increases the probabilities of every transition of kind admission to hospital, in-hospital deaths and out-of-hospital deaths. In particular the effects on death transitions are greater than the others.
Conversely, this variable decreases the probabilities of discharge from hospital transitions or not influences them.

Figure 4.6 shows the confidence intervals for hazard ratios of Charlson index variable.
This variable increases the probabilities of every transition of kind admission to hospital, in-hospital deaths and out-of-hospital deaths and decreases the probabilities of discharge from hospital transitions.

Figure 4.7 shows the confidence intervals for hazard ratios of pre hospitalization cardiological evaluation variable.

Figure 4.5: 95% confidence intervals for hazard ratios of age (all other covariates fixed) estimated fitting Cox model to Friuli Venezia Giulia dataset.

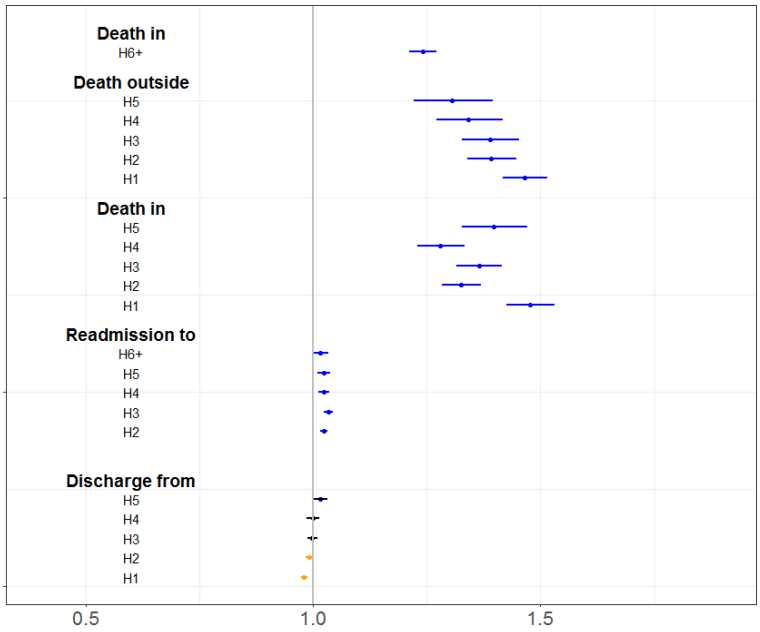


Figure 4.6: 95% confidence intervals for hazard ratios of Charlson index (all other covariates fixed) estimated fitting Cox model to Friuli Venezia Giulia dataset.

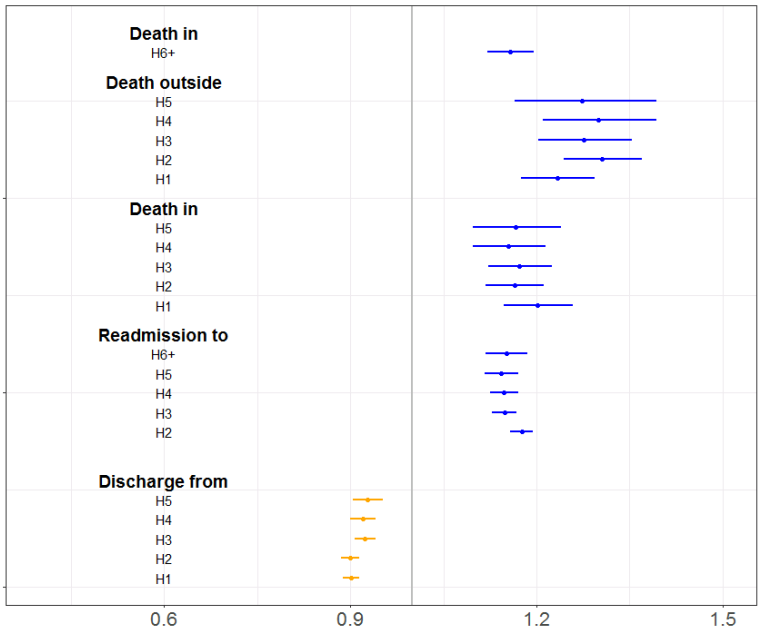
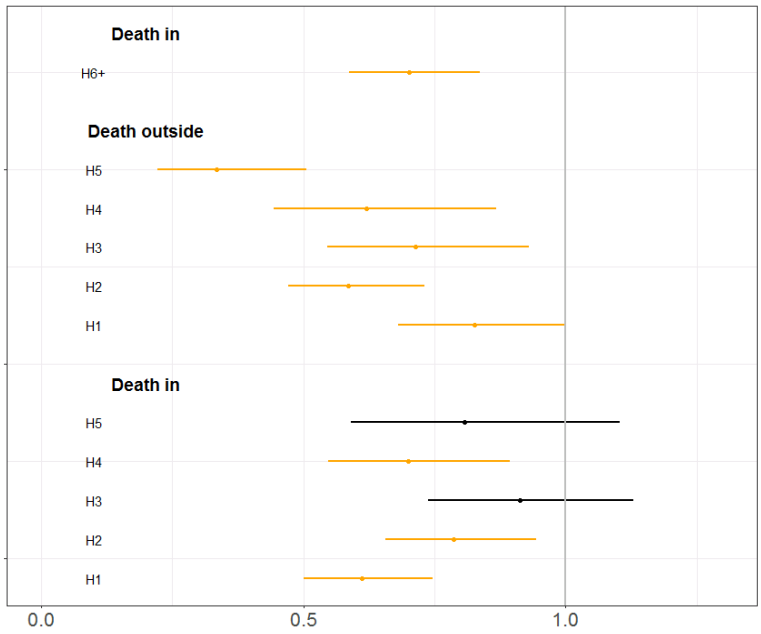


Figure 4.7: 95% confidence intervals for hazard ratios of pre hospitalization cardiological evaluation (all other covariates fixed) estimated fitting Cox model to Friuli Venezia Giulia dataset.



We note that the significant terms are all negative. This means that having a pre hospitalization decreases the probabilities of having a transition to death.

Figure 4.8 shows the confidence intervals for hazard ratios of CW admission variable.
The hazard ratios related to the discharge from hospital are all bigger than one, indicating that being admitted in CW increases the LOS of a patients. On the other hand, the hazard ratios related to death inside or outside hospital are all smaller than one. Being admitted in this ward is a protective factor for the deaths since it decreases the probability to have a transition to death.

Figure 4.9 shows the confidence intervals for hazard ratios of worsening index variable.
Having a hospitalization in the five years before the index admission increases the istantaneous risk of being readmitted and decreases or does not influence the risk of death.

Figure 4.8: 95% confidence intervals for hazard ratios of CW admission (all other covariates fixed) estimated fitting Cox model to Friuli Venezia Giulia dataset.

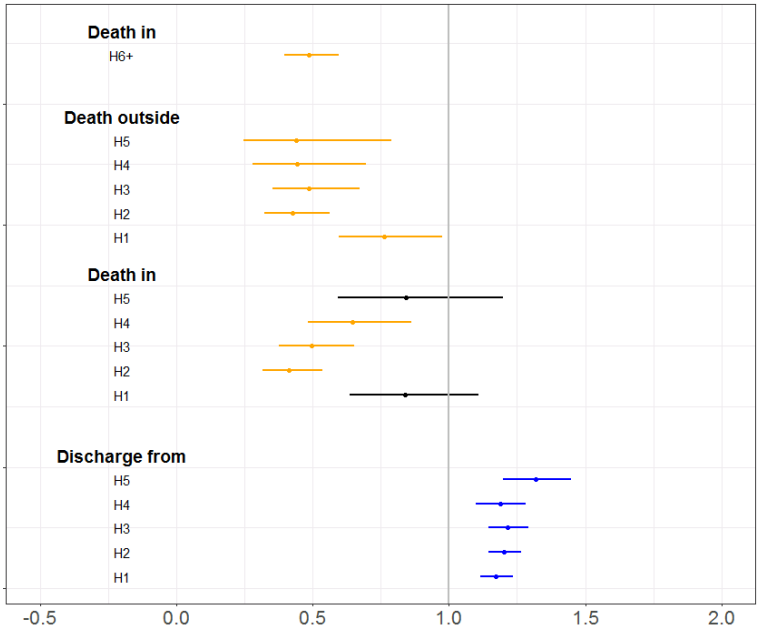


Figure 4.9: 95% confidence intervals for hazard ratios of worsening index (all other covariates fixed) estimated fitting Cox model to Friuli Venezia Giulia dataset.

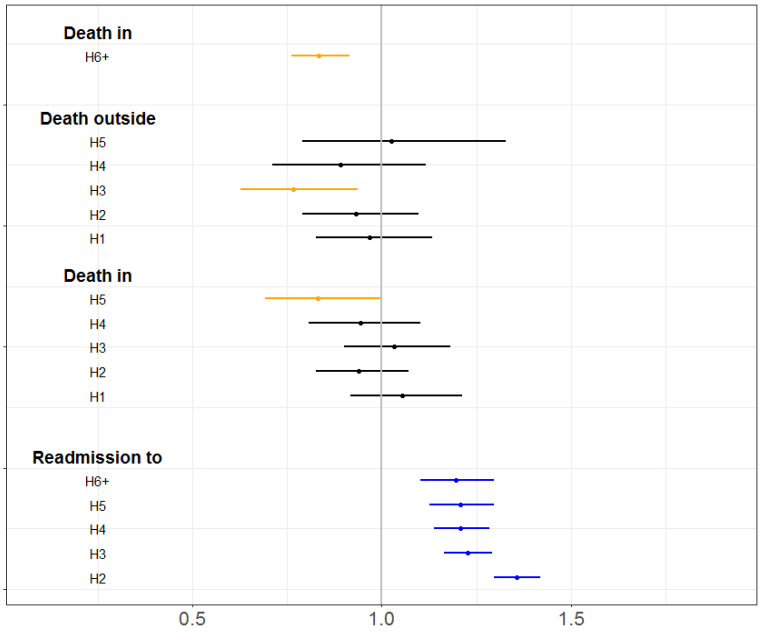


Figure 4.10: 95% confidence intervals for hazard ratios of ICU/IHC (all other covariates fixed) estimated fitting Cox model to Friuli Venezia Giulia dataset.

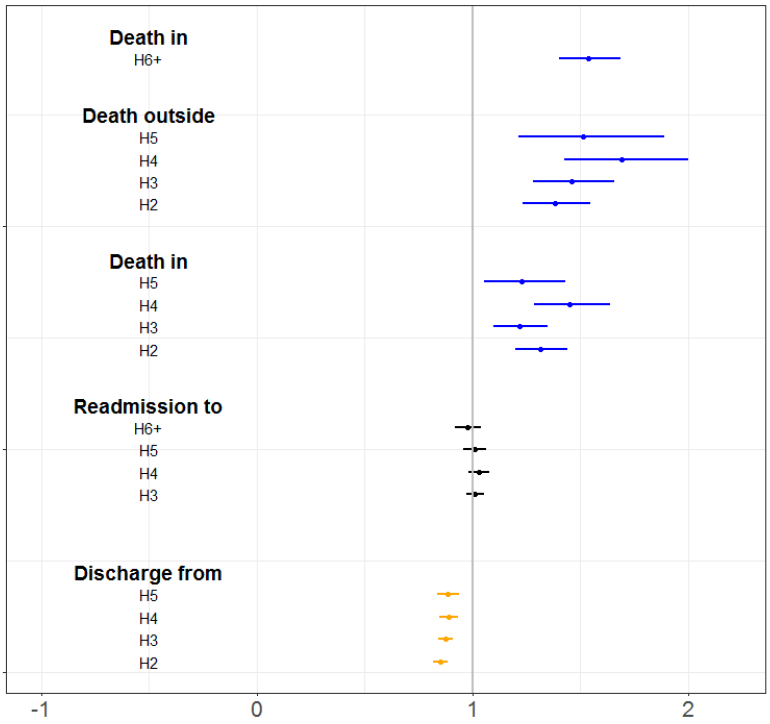


Figure 4.10 shows the confidence intervals for hazard ratios of ICU/IHC index variable.
Experiencing one of these events decreases the probabilities of being discharged from hospital and increases the probabilities of dying. The probabilities of being readmitted in hospital are not influenced by this variable.

We can conclude saying that all the results obtained fitting the multistate model on the Friuli Venezia Giulia dataset are coherent with the results obtained fitting the same model on the Trieste dataset.
The influence of the different variables on the different transitions is the same in both cases.
Anyway, we have to note that the significant terms of the model fitted on the Friuli Venezia Giulia dataset are more than the significant terms of the model fitted on the Trieste dataset. This could be due to the fact that the first dataset is bigger and more comprehensive, and this can lead to more significative estimates.

Chapter 5

Conclusions and further developments

In this thesis work we analized two different administrative dataset, in which several information about HF patients from Trieste and from Friuli Venezia Giulia region are collected.

Firstly, we focused on Trieste dataset. We selected the most relevant administrative information and we computed some clinical indices. Afterwards, we applied a musti-state model in order to analyze the effect of different covariates (i.e. sex, age, Charlson index, worsening index, presence of a pre hospital cardiological evaluation, ICU/IHC index) on the admission to hospital, discharge from hospital and death dinamic. We showed that some variables act as a protective factor in some transitions and as a risk factor in other transitions.
Then, we analyzed how the previous variables affect the survival outcome, through the Kaplan-Meier estimate of the survival, stratified according to the different variables (age at first admission, sex, Charlson index at fist admission, presence of a pre hospitalization cardiological evaluation, presence of an ICU/IHC event). We noticed that age is an important discriminant factor and that the presence of an ICU/IHC event has a double effect that depends on the time from the first admission.
We also analyzed how the previous variables affect the time to second admission outcome, throught the Kaplan-Meier estimate of the time to second admission, stratified according to the different variables (age, Charlson index

and ICU/IHC index). In this case the effect of the variables is less evident, even though, they are always statistically significant.

Afterwards, we focused on Friuli Venezia Giulia dataset, a bigger dataset that contains also Trieste area.
First of all, we analyzed the distribution of the main features among the 20 existing residence districts and we plotted the Kaplan-Meier estimate of the survival, stratified according to these districts. Through this first analysis we noted that the residence district does not significantly affect the survival. In order to confirm this hypothesis we applied a npdf Cox model with residence district specific frailty. Indeed, through this model, we can evaluate if a possible clustering structure, based on the residual variability, could be found among the residence districts. If no clustering structure is detected, it means that the residual variability should not be modeled through a discrete random variable. The results of this model suggested the absence of a possible clustering structure, hence we could state that there is homogeneity among the hospital treatment in the different residence districts.
Subsequently, we analyzed the cohort distribution and homogeneity. Firstly, we applied different Cox models with a parametric shared frailty term, distributed as a Gamma or a Normal, in order to evaluate if the residual variability can be explained with this term. Since the estimate of the frailty variance is low (0.005 and 0.003), we concluded that the residual variability should not be explained through an individual-specific parametric frailty term. Then, we applied the npdf Cox algorithm with patient specific frailty. However, the simulation study reported in Appendix A showed us that the estimates of this algorithm are not reliable in such specific case. We could not detect a possible clustering structure based on the residual variability but we could detect a possible clustering structure based on the recorded covariates. So, we decided to proceed fitting a k-means algorithm. We detected four clusters characterized by different mean survival times and different percentages of deaths, primarily justified by the different mean ages of the patients. In order to better understand the differences among the four clusters, we fitted a simple Cox model on each cluster. An interesting aspect that emerged, and that confirmed the results obtained with the analysis of the Trieste dataset, is the double effect of the ICU/IHC index: it acts as a protective factor for older patients and as a risk factor for younger patients.

Finally, we concluded the analysis of Friuli Venezia Giulia dataset fitting a multi-state model whose results confirmed the ones obtained through the analysis of Trieste dataset.

As concerns the limitations and the possible future developments, we are aware of the fact that Charlson index is not the best index to measure the comorbidity load in such an old cohort. In this case, considering CIRS index (Cumulative Illness Rating Scale) would have been more appropriate, but we were not able to compute this index because some of the required variables (i.e. psychiatric conditions) are not routinely measured. Moreover, in this work we considered the combined binary index ICU/IHC, but two separate indices or an index that takes into account the length of these events could be considered and could be a target to be addressed in future works.
After highlighting the fragility of the npdf Cox algorithm with patient specific frailty, we concluded the study of the cohort homogeneity through the application of the k-means algorithm, that highlighted the presence of possible clusters based on the recorded covariates. This clustering structure should be studied in deep and other clustering techniques could be considered in future works.

Appendix A

Simulation study

A.1 Npdf Cox model

A simulation study is conducted to evaluate the performance of the estimators obtained with the npdf Cox algorithm, described in Section 2.2.
We simulate 100 datasets for each value of N=number of groups (e.g. residence districts) $\in \{10, 50, 100\}$ and S=statistical units per group (e.g. patients) $\in \{1, 5, 20, 50, 100\}$.
The number of latent populations is varied in two scenarios with $K = 2$ and 4, while the mixing proportions and frailty ratios are fixed at balanced values listed in Table A.1.

K	π_1	π_2	π_3	π_4	w_1	w_2	w_3	w_4
2	0.5	0.5	-	-	1	3	-	-
4	0.3	0.2	0.3	0.2	1	1.5	3	5

Table A.1: Values of mixing proportions and frailty ratios used in the simulation staudy for each number of latent populations K.

For all the simulations, we set the covariate-related log hazard ratio = 0.4, and define the baseline cumulative hazard so that $\Lambda_0^{-1}(t) = 0.01 \cdot t^{1.9}$ in order to mimic the dataset that motivated this study.
The aim of the simulation is to estimate how well the algorithm estimates

the number of latent populations K for various values of N and S.

In Table A.2 and A.3 we report the results of the simulations, respectively for K=2 and 4 latent populations. For each combination of N and S we report the number of latent populations estimated by AIC, BIC and Laird, together with the estimated mixing proportions $\boldsymbol{\pi}$. When the number of estimated latent populations is smaller that the real number of latent populations, we report an asterisk, since it is not possible to compute the estimate.

N	S	AIC	BIC	Laird	Estimated $\boldsymbol{\pi}$
10	1	1	1	1	*
10	5	1	1	1	*
10	20	2	2	3	0.4 0.6
10	50	2	2	2	0.5 0.5
10	100	2	2	3	0.5 0.5
50	1	1	1	2	0.16 0.84
50	5	1	1	3	0.24 0.76
50	20	2	2	2	0.5 0.5
50	50	2	2	2	0.5 0.5
50	100	2	2	2	0.5 0.5
100	1	1	1	2	0.3 0.7
100	5	1	1	2	0.5 0.5
100	20	2	2	3	0.52 0.48
100	50	2	2	4	0.5 0.5
100	100	2	2	2	0.5 0.5

Table A.2: Simulation study results for each combination of N and S when 2 latent populations are present.

N	S	AIC	BIC	Laird	Estimated $\boldsymbol{\pi}$
10	1	1	1	1	*
10	5	1	1	2	*
10	20	2	2	3	*
10	50	2	2	3	*
10	100	3	3	3	*
50	1	1	1	2	*
50	5	1	1	3	*
50	20	2	2	4	0.36 0.12 0.42 0.1
50	50	4	4	4	0.28 0.22 0.3 0.2
50	100	4	4	4	0.34 0.16 0.32 0.18
100	1	1	1	1	*
100	5	1	1	3	*
100	20	2	2	4	0.22 0.31 0.3 0.17
100	50	4	4	5	0.3 0.21 0.3 0.19
100	100	4	4	5	0.28 0.22 0.3 0.2

Table A.3: Simulation study results for each combination of N and S when 4 latent populations are present.

Looking at the results in Table A.2 and A.3, we can note that AIC and BIC estimate, in the majority of the scenarios, the same number of latent populations, while Laird tends to estimate higher values, as expected.
We can observe that the algorithm performs well only when the number of statistical units per group is sufficiently high. For example in Table A.2, where two latent populations are present, the correct number of latent population is estimated only when the statistical units per group are more than 20, both when we have 10, 20 or 100 groups. Otherwise, in Table A.3, where four latent populations are present, the correct number of latent population is estimated only when we have more than 50 statistical units per group in case of 20 and 100 groups. When we have one unit per group the algorithm always estimates only one latent population. This means that in such cases others clustering techniques have to be considered.
Another thing to note is that, when the real number of latent population

increases, the estimates are more precise if we have many groups; we can see for example the results of the first block from the top in Table A.3. This is quite reasonable, in fact, when there are 10 groups it is unlikely that there are 4 latent populations. We can conclude saying taht this model is a good method of screening when we have complex and large databases.

A.2 Cox model with a shared gamma frailty term

In this section we want to analyze if, even with an algorithm different from the npdf Cox one, the estimation of the model parameters is difficult when the groups are made by only one unit. For this reason a simulation study is conducted to evaluate the performance of the estimators obtained with the Cox algorithm with gamma distributed frailty. In particular we evaluate the performance of the *coxph* function of the *survival* pakage of R. We simulate 100 datasets for each value of N=number of groups (e.g. residence districts) $\in \{10, 50, 100\}$ and S=statistical units per group (e.g. patients) $\in \{1, 5, 20, 50, 100\}$. We repeat the simulations for four different values of θ=variance of the gamma distributed frailty $\in \{0.3, 0.7, 1.5, 2\}$. We set the covariate-related log hazard ratio = 0.4, and we define the baseline cumulative hazard so that $\Lambda_0^{-1}(t) = 0.01 \cdot t^{1.9}$, in order to mimic the dataset that motivated this study.
The aim of the simulation is to evaluate how well the algorithm estimates the variance of random effect and the covariate-related log hazard ratio $\hat{\beta}$ for variuous values of N and S.

In Table A.4, A.5, A.6 and A.7 we report the results of the simulations, respectively for $\theta = 0.3, 0.7, 1.5$ and 2. For each combination of N and S we report the mean value and the standard deviation of the variance of random effect and of $\hat{\beta}$, hence we compute the mean and the standard deviation of the estimates obtained fitting the Cox model on the 100 simulated dataset.

Looking at the results in Table A.4, A.5, A.6,A.7 we can note that, in the majority of the scenarios, the covariate-related log hazard ratio is correctly estimated, since the values referring to $\hat{\beta}$ are around 0.4, that is the real value. However, we note that the $\hat{\beta}$ estimates, in the case of groups made by

N	S	Variance of random effect mean (sd)	$\hat{\beta}$ mean (sd)
10	1	0.004 (0.048)	0.376 (0.601)
10	20	0.309 (0.158)	0.389 (0.081)
10	50	0.477 (0.225)	0.408 (0.058)
10	100	0.734 (0.190)	0.403 (0.032)
50	1	0.023 (0.117)	0.369 (0.195)
50	20	0.370 (0.104)	0.403 (0.031)
50	50	0.606 (0.105)	0.401 (0.020)
50	100	0.810 (0.064)	0.398 (0.014)
100	1	0.052 (0.170)	0.363 (0.141)
100	20	0.388 (0.077)	0.401 (0.027)
100	50	0.633 (0.063)	0.405 (0.015)
100	100	0.829 (0.043)	0.402 (0.009)

Table A.4: Cox algorithm simulation study results for each combination of N and S when the real variance of the gamma distributed frailty is 0.3.

only one unit, are much less accurate. We can see for example how, when S=1 and $\theta = 2$ in Table A.7, the estimates are 0.306, 0.159 and 0.182.
Looking at the estimates of the variance of random effect, we can note that when $S = 1$ the estimates are always inaccurate. Instead, for bigger values of S, the estimate depends on θ. When $\theta = 0.3$, in Table A.4, the estimate is very accurate for $S = 20$ and less accurate for bigger values of S. When $\theta = 0.7$, in Table A.5, the estimate is not very accurate for any S values bigger than 1. When $\theta = 1.5$ or 2, in Table A.6 and A.7, the estimate is fairly accurate for each value of S bigger than 1.

These results are in line with our hypothesis: the parameters estimates of a shared frailty survival model when all the groups are made by only one unit are not accurate and other methods are needed.

N	S	Variance of random effect mean (sd)	$\hat{\beta}$ mean (sd)
10	1	0.007 (0.073)	0.393 (1.187)
10	20	0.690 (0.299)	0.412 (0.092)
10	50	0.882 (0.214)	0.395 (0.044)
10	100	0.954 (0.148)	0.403 (0.040)
50	1	0.025 (0.126)	0.259 (0.189)
50	20	0.806 (0.121)	0.402 (0.035)
50	50	0.932 (0.062)	0.404 (0.020)
50	100	1.002 (0.029)	0.398 (0.014)
100	1	0.095 (0.229)	0.290 (0.157)
100	20	0.826 (0.080)	0.404 (0.025)
100	50	0.954 (0.042)	0.399 (0.014)
100	100	1.000 (0.018)	0.401 (0.011)

Table A.5: Cox algorithm simulation study results for each combination of N and S when the real variance of the gamma distributed frailty is 0.7.

N	S	Variance of random effect mean (sd)	$\hat{\beta}$ mean (sd)
10	1	4.8e-07 (8.4e-08)	0.263 (0.574)
10	20	1.433 (0.488)	0.407 (0.074)
10	50	1.364 (0.519)	0.398 (0.051)
10	100	1.449 (0.481)	0.406 (0.034)
50	1	0.030 (0.123)	0.189 (0.163)
50	20	1.552 (0.280)	0.398 (0.035)
50	50	1.526 (0.278)	0.401 (0.020)
50	100	1.521 (0.279)	0.398 (0.015)
100	1	0.118 (0.292)	0.202 (0.167)
100	20	1.505 (0.214)	0.402 (0.026)
100	50	1.538 (0.202)	0.400 (0.013)
100	100	1.617 (0.222)	0.399 (0.010)

Table A.6: Cox algorithm simulation study results for each combination of N and S when the real variance of the gamma distributed frailty is 1.5.

N	S	Variance of random effect mean (sd)	$\hat{\beta}$ mean (sd)
10	1	0.007 (0.075)	0.306 (0.599)
10	20	1.832 (0.756)	0.401 (0.099)
10	50	1.925 (0.739)	0.393 (0.055)
10	100	1.860 (0.696)	0.396 (0.034)
50	1	0.057 (0.227)	0.159 (0.202)
50	20	2.018 (0.375)	0.398 (0.032)
50	50	2.000 (0.322)	0.396 (0.020)
50	100	1.958 (0.289)	0.400 (0.013)
100	1	0.133 (0.309)	0.182 (0.144)
100	20	2.025 (0.307)	0.400 (0.025)
100	50	2.014 (0.220)	0.402 (0.016)
100	100	2.013 (0.235)	0.400 (0.010)

Table A.7: Cox algorithm simulation study results for each combination of N and S when the real variance of the gamma distributed frailty is 2.

Appendix B

Code

In this appendix we will report the main functions that we used to obtain the results reported in Chapter 3.

```
library(survival)

# COX MULTI-STATE MODEL

# dat_expanded, as required by the "coxph" function, is a
# dataset in long format whose columns are:
# AGE.1, AGE.2,...,R_A.20.

modello_dummy <- coxph(Surv(time,status) ~

  AGE.1 + AGE.5 + AGE.9 + AGE.13 + AGE.17
+ AGE.2 + AGE.6 + AGE.10+ + AGE.14 + AGE.18 + AGE.21 +
+ AGE.3 + AGE.7 + AGE.11 + AGE.15 + AGE.19 +
+ AGE.4 + AGE.8 + AGE.12 + AGE.16 + AGE.20 +

+ SEX.1 + SEX.5 + SEX.9 + SEX.13 + SEX.17
+ SEX.2 + SEX.6 + SEX.10 + + SEX.14 + SEX.18 + SEX.21 +

+ CH.1 + CH.5 + CH.9 + CH.13 + CH.17
+ CH.2 + CH.6 + CH.10+ + CH.14 + CH.18 + CH.21 +
+ CH.3 + CH.7 + CH.11 + CH.15 + CH.19 +
+ CH.4 + CH.8 + CH.12 + CH.16 + CH.20 +

+ PRE.2 + PRE.6 + PRE.10 + PRE.14 + PRE.18 + PRE.21 +
```

```
+ PRE.4 + PRE.8 + PRE.12 + PRE.16 + PRE.20 +

+ CW.1 + CW.5 + CW.9 + CW.13 + CW.17 +
+ CW.2 + CW.6 + CW.10 + CW.14 + CW.18 + CW.21 +
+ CW.4 + CW.8 + CW.12 + CW.16 + CW.20 +

+ WOR.3 + WOR.7 + WOR.11 + WOR.15 + WOR.19 +
+ WOR.2 + WOR.6 + WOR.10 + WOR.14 + WOR.18 + WOR.21 +
+ WOR.4 + WOR.8 + WOR.12 + WOR.16 + WOR.20 +

+ R_A.1 + R_A.5 + R_A.9 + R_A.13 + R_A.17
+ R_A.2 + R_A.6 + R_A.10+ + R_A.14 + R_A.18 + R_A.21 +
+ R_A.3 + R_A.7 + R_A.11 + R_A.15 + R_A.19 +
+ R_A.4 + R_A.8 + R_A.12 + R_A.16 + R_A.20 +

+ strata( trans ), data=dat_expanded, method="breslow")

summary(  modello_dummy )


# KAPLAN-MEIER ESTIMATES OF THE SURVIVAL CURVES
# in dati_km we have one row for each patient
dati_km<-read.table("dati_vita_km.txt",header=T)

# as example we report the computation to obtain the
# estimate of the survival stratified by ICU/IHC index,
# the stratifications according to all the others
# variables are similar

# we create a survival object
km <- survfit(Surv(vita,stato_vita) ~
r_a, data = dati_km, conf.type = "log-log")

# we plot the KM curve
plot(km, conf.int=T,col=c('blue','orange'),
xlab='Time[days]', ylab='Survival')

# we perform the log rank test
logrank<-survdiff(Surv(vita,stato_vita) ~
r_a,data = dati_km)
```

Bibliography

[1] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[2] Per Kragh Andersen, Steen Z Abildstrom, and Susanne Rosthøj. Competing risks as a multi-state model. *Statistical methods in medical research*, 11(2):203–215, 2002.

[3] Per Kragh Andersen and Richard David Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

[4] Per Kragh Andersen and Niels Keiding. Multi-state models for event history analysis. *Statistical methods in medical research*, 11(2):91–115, 2002.

[5] Jeffrey A Bakal, Finlay A McAlister, Wei Liu, and Justin A Ezekowitz. Heart failure re-admission: measuring the ever shortening gap between repeat heart failure hospitalizations. *PloS one*, 9(9):e106494, 2014.

[6] Jan Beyersmann, Arthur Allignol, and Martin Schumacher. *Competing risks and multistate models with R*. Springer Science & Business Media, 2011.

[7] Jose Cortinas Abrahantes and Tomasz Burzykowski. A version of the em algorithm for proportional hazard model with random effects. *Biometrical Journal*, 47(6):847–862, 2005.

[8] Liesbeth C De Wreede, Marta Fiocco, and Hein Putter. The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Computer methods and programs in biomedicine*, 99(3):261–274, 2010.

[9] Liesbeth C de Wreede, Marta Fiocco, Hein Putter, et al. mstate: an r package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, 38(7):1–30, 2011.

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[11] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

[12] Joshua J Gagne, Robert J Glynn, Jerry Avorn, Raisa Levin, and Sebastian Schneeweiss. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *Journal of clinical epidemiology*, 64(7):749–759, 2011.

[13] Francesca Gasperoni, Francesca Ieva, Giulia Barbati, Arjuna Scagnetto, Annamaria Iorio, Gianfranco Sinagra, and Andrea Di Lenarda. Multi-state modelling of heart failure care path: A population-based investigation from italy. *PloS one*, 12(6):e0179176, 2017.

[14] Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Christopher H. Jackson, and Linda D. Sharples. Nonparametric frailty cox models for hierarchical time-to-event data. Technical report.

[15] Philip Hougaard. Multi-state models: a review. *Lifetime data analysis*, 5(3):239–264, 1999.

[16] Philip Hougaard. *Analysis of multivariate survival data*. Springer Science & Business Media, 2012.

[17] Francesca Ieva, Christopher H Jackson, and Linda D Sharples. Multi-state modelling of repeated hospitalisation and death in patients with

heart failure: the use of large administrative databases in clinical epidemiology. *Statistical methods in medical research*, 26(3):1350–1372, 2017.

[18] Statistiche Istat. Censimento popolazione istat; 2016, 2017.

[19] Christopher Jackson. Multi-state modelling with r: the msm package. *Cambridge, UK*, 2007.

[20] Søren Johansen. An extension of cox's regression model. *International Statistical Review/Revue Internationale de Statistique*, pages 165–174, 1983.

[21] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[22] Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

[23] Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

[24] Douwe Postmus, Dirk J Veldhuisen, Tiny Jaarsma, Marie Louise Luttik, Johan Lassus, Alexandre Mebazaa, Markku S Nieminen, Veli-Pekka Harjola, James Lewsey, Erik Buskens, et al. The coach risk engine: a multistate model for predicting survival and hospitalization in patients with heart failure. *European journal of heart failure*, 14(2):168–175, 2012.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[26] Terry M Therneau and PM Grambsch. Extending the cox model. *Edited by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg*, page 51, 2000.

[27] Florin Vaida, Ronghui Xu, et al. Proportional hazards model with random effects. *Statistics in medicine*, 19(24):3309–3324, 2000.

Ringraziamenti

Vorrei innanzitutto ringraziare la professoressa Francesca Ieva che mi ha proposto questa tesi e che mi ha seguito durante tutto il lavoro con disponibilità e precisione.

Vorrei ringraziare la dottoressa Francesca Gasperoni che con la sua gentilezza e infinita disponibilità mi ha aiutato a non perdermi tra tutte le righe di codice e mi ha accolto ad ogni ora alla sua scrivania per rispondere a tutte le mie domande.

Vorrei ringraziare soprattutto la mia famiglia, mia mamma e mio papà, che mi hanno sempre supportato e sopportato in questi lunghi anni di studi e di esami, perchè hanno sempre creduto in me più di quanto ci credessi io. A mia mamma va anche un grazie speciale per tutta la consulenza psicologica fornita prima e dopo ogni esame.

Vorrei ringraziare gli amici che, come me, hanno deciso di intraprendere questa folle avventura al Poli, senza di voi questi cinque anni non sarebbero stati gli stessi. Grazie per le intere giornate di studio disperato insieme, per i ricchi pranzi, per le serate in compagnia, per i viaggi e per le mille risate. Un grazie speciale alle foxes, fedeli compagne di progetti, e agli amici che mi accompagnano dai tempi del Lussana, Simo e Ari, i caffè insieme per sopravviviere allo studio non si contano ormai più.

Infine vorrei ringraziare Fabio, che ormai conosce ogni parola di questa tesi meglio di me, per la sua presenza e il suo continuo appoggio.