POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete nonparametric frailty approach

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Riccardo Scaramuzza**

Student ID: 944904
Advisor: Prof. Francesca Ieva
Co-advisors: Dr. Chiara Masci, Dr. Marta Spreafico
Academic Year: 2020-21

Abstract

Heart Failure is a largely diffused chronic condition, consisting in the deterioration of the function of the heart, which may ends in the patient's death. In particular, in Italy almost 600.000 persons suffer from this disease, which is considered the main cause of hospitalization for 65+ patients. The disease prevalence and the impact over the sanitary system have made this pathology object of study, both from the clinical and administrative point of view. In particular, several researches have focused on the investigation of the effect that different drugs associated to the treatment and the corresponding adherence to treatment have on the survival as well as on the clinical path a patient comes across.
In this work we analyze data coming from the Lombardy region administrative database, focusing on the joint modelling of the patient's clinical path evolution in terms of hospitalizations and survival, evaluating the effect that the adherence to a Heart Failure specific drug (i.e. ACE inhibitors) treatment has upon them.
The goal of this thesis is a methodological discussion, aimed at spotting the most promising tool to jointly model a recurrent events process and a terminal event one, with a focus on frailty models. We present as the main result our innovative model, in which the two processes of interest are jointly modelled through a multivariate discrete non parametric frailty. This frailty formulation reveals a big potential from an interpretative point of view, especially for the application at hand. In fact, it enables a more direct analysis of the induced partition of patients in subpopulations characterized by different levels of frailty, which can be easily translated in a providers' assessment. We also provide a comparative study to assess the effectiveness of our model in a controlled setting.

Keywords: Heart Failure, joint models, recurrent events, non parametric frailty models

Sommario

Lo Scompenso Cardiaco è una condizione cronica largamente diffusa, che consiste nel deterioramento della funzione del cuore, e che può, nei casi più gravi, portare alla morte. In Italia, in particolare, circa 600,000 persone soffrono di questa patologia, che è la prima causa di ospedalizzazione nella fascia degli ultrasessantacinquenni. L'incidenza sulla popolazione, la possibilità di gravi conseguenze e l'impatto sul sistema sanitario hanno reso questa malattia oggetto di studio, sia dal punto di vista clinico che amministrativo. In particolare, numerose ricerche si sono concentrate sull'effetto che i diversi farmaci usati nel trattamento della malattia, e la corrispondente aderenza a tale trattamento, hanno sia sulla probabilità di sopravvivenza sia sul decorso clinico dei pazienti.
In questo lavoro analizziamo i dati provenienti dalla banca dati amministrativa di Regione Lombardia, concentrandoci sulla modellazione congiunta dell'evoluzione del percorso clinico dei pazienti in termini di ospedalizzazioni e di sopravvivenza, valutando l'effetto che l'aderenza a un trattamento con farmaci specifici (ACE inibitori) ha su queste ultime.
Lo scopo di questa tesi è una discussione metodologica, il cui obiettivo è trovare lo strumento più promettente per modellare un processo di eventi ricorrenti congiuntamente a un evento terminale, con focus sui modelli frailty. Presentiamo come risultato principale un nostro modello innovativo, in cui i due processi sono modellati congiuntamente tramite una frailty multivariata che segue una distribuzione discreta non parametrica. Tale formulazione della frailty mostra un particolare potenziale dal punto di vista interpretativo, specialmente per quanto riguarda l'applicazione considerata, visto che permette un'analisi diretta del raggruppamento dei pazienti in sottopopolazioni caratterizzate da differenti livelli di fragilità, che può facilmente tradursi in uno strumento per l'assesment ospedaliero. Forniamo anche uno studio comparativo con l'obiettivo di verificare l'efficacia del nostro modello in un ambiente controllato.

Parole chiave: Scompenso cardiaco, modelli congiunti, eventi ricorrenti, modelli con frailty non parametrica

Contents

Introduction

Heart failure (HF) is a condition characterized by a deterioration in the function of the heart that makes it unable to contract (systole) or release (diastole) adequately to pump enough blood to meet the body's needs. Due to heart failure, on the one hand, the organs and tissues receive insufficient quantities of oxygen and nutrients for their metabolic needs, on the other hand there tends to be an accumulation of excess fluid in the lungs and tissues [34]. The HF condition may be provoked by several different causes: the main ones are myocardial ischemia, high blood pressure, cardiomyopathies, valvular heart disease, pulmonary hypertension and congenital heart disease [41]. The diagnosis of HF is usually difficult due to the variety of its symptoms, which comprehends breathlessness, reduced exercise tolerance, fatigue and edema, which may worsen evolving, in the worst cases, to acute pulmonary edema and finally death. Besides its mortality, which reaches relatively high rates in the first period after a HF event but has a much more gradual slope afterwards, this disease is considered a public health problem of primarily importance due to its incidence over the population. According to [34], it represents the first cause of hospitalization in people over 65 years of age and, only in Italy, over 600,000 cases are recorded. Moreover, since it is a condition linked to the lengthening of the average life span, as its prevalence increases from year to year due to the general aging of the population (i.e. due to the general increase of survival in industrialized societies), its impact is likely to grow even bigger. The standard treatment is represented by pharmacological therapies, usually involving Angiotensin-Converting Enzyme inhibitors, Angiotensin Receptors Blockers, Anti Aldosterone agents, Beta Blocking agents and Diuretics [33].

This thesis arises in an attempt to respond to the renewed need of methods capable of properly model hospitalizations and survival of patients affected by major accident diseases, such as Heart Failure. In particular, we want to investigate tools able both to cope with two correlated processes, the first regarding recurrent events (i.e. hospitalizations) and the latter terminal ones (i.e. deaths), and to assess the effect that available therapies have on them. To do so, we deeply delve into the field of Survival analysis, offering an overview of the techniques developed in this context, which belong to the class of frailty

models. The latter are Cox models in which a random effect (i.e. the frailty) is added to the linear predictor, in order to account for unexplained heterogeneity at patients' level for the process of interest. In recent studies, their application allow the simultaneous modelling of the two mentioned processes through the linking of the hospitalizations and death frailties, which follow a joint (usually multivariate Normal) distribution. This methodological research allow us to investigate the issues and the corresponding solutions adopted in literature for this class of models, finally proposing an innovative approach, in which the two processes' frailties follow a non parametric, discrete distribution.

The dissertation is developed in five Chapters.
In Chapter 1 we presented some pharmacoepidemiology concepts, like duration of a prescription and adherence, which we consider necessary to understand the analysis and procedures we adopt in the following of our work.
In Chapter 2 we describe the features of the Heart Failure Lombardy Region administrative database and the information it furnishes. Moreover, we detail the adopted procedures to obtain from it the ACE inhibitors dataset, on which we perform our analysis. In particular, we describe how patients are selected, how the initial dataset is splitted in monotherapy subdatasets and how the variables involved in these latter are defined.
Chapter 3 represents the core of our work, as it presents the theoretical foundations of all the methodologies considered in our work. It comprehends an introduction to classical survival analysis tools, in particular to Cox models, whose aim is to present the framework and explain how the pharmacoepidemiological concepts of Chapter 1 can be integrated in the analysis. It follows a discussion of the recurrent events framework, its issues and the methods adopted to tackle them, with a focus about frailty models. Then the joint modelling of the recurrent events process and the terminal events one, achieved through the linking of the two processes' frailties, is presented. Finally, within this joint modelling context, we propose an original extension in which the frailties follows a discrete, non parametric distribution. This represents our main contribution, as we assume that this frailties characterization may result more exploitable from a hospital assessment point of view, in addition to opening the way to the analysis of the partition of patients in subpopulations characterized by different levels of fragility, which is induced by the identified discrete distribution.
In Chapter 4 we present the results obtained applying the techniques described in Chapter 3 to the ACE inhibitors dataset. Our goal, at this stage, is both to gain insights about the hospitalizations and death processes, paying particular attention to the effect that adherence to the ACE inhibitors treatment has upon them, and to conduct a compara-

tive study among the proposed jointly models, in order to assess the effectiveness of our proposed extension.
The thesis is concluded in Chapter 5 by a discussion which summarizes the results obtained and suggests some further developments, in particular for what it concerns our original model.

1 | Pharmacoepidemiology

As mentioned in the Introduction, our thesis has a strong methodological component, but the questions addressed arise from a specific application context, related to the field of pharmacoepidemiology. Actually, we aim at investigating suitable modelling tools to jointly assess the effect of an ACE inhibitors treatment on readmissions and mortality of heart failure patients. Because of that, in this Chapter we propose a very brief introduction to this field, stressing some concepts and definitions which are necessary to properly understand some of the analysis presented in Chapter 4. Firstly, in Section 1.1 we give a definition of Pharmacoepidemiology and Drug Utilization Research, specifying their goals. In Section 1.2 we describe useful information about drugs which are usually reported in administrative databases (like the Heart Failure dataset, described in Chapter 2), as they facilitate a proper managing of this kind of aggregated data. In Section 1.3 we explain the concept of adherence and how to compute corresponding quantitative indicators from the available data. Finally, in Section 1.4 we present briefly the drug of our concern, the ACE inhibitors.

1.1. Drug Utilization Research

One of the most popular definition of Pharmacoepidemiology was given by Strom in [58]

Definition 1.1. *(Pharmacoepidemiology)*
Pharmacoepidemiology is the study of the use, the effectiveness and safety of post-marketing drugs on a large sample with the purpose of supporting the rational and cost-effective use of drugs in the population in order to improve the health outcomes.

Actually, as it can be argued from the definition, pharmacoepidemiology is a broad field, encompassing several branches which specialize on different areas. In particular, the investigation about the use of drugs and their effect on patients' clinical courses refers to a branch known as Drug Utilization Research. Reporting the words used by the World Health Organization [38], it is defined as

Definition 1.2. *(DUR)*
Drug Utilization Research consists in the marketing, distribution, prescription, and use of a drug in the society, with special emphasis on the resulting medical, social and economic consequences.

The ultimate goal of DUR is to identify and communicate the proper use of drugs to patients, combining researches which belong to the medical, economical and social fields. It is important to stress the fact that the scope is not limited only to the clinical effect of drugs, even if that represents a very important part. As an example, in our case the focus is both the survival (clinical) outcome and the effect of the drug treatment on recurrent hospitalizations of a patient affected by a chronic disease (i.e. an important hospital management issue). Usually, when dealing with Drug Utilization Research, data are retrieved from two different types of databases, clinical and administrative. Clinical databases are focused on habits and lifestyles of patients, as well as diagnosis and intermediate clinical outcomes composing a patient's clinical course. Administrative databases, instead, contains information about personal data, prescriptions and hospitalization recordings. The integration of these two types of databases allows for measurement of a patient's drug use. Finally, we must recall that the main limitation of this approach consists in the strong assumption that prescribing a therapy to a patient actually coincides with the patient following it properly (i.e., we cannot verify that the patient is actually taking the medications).

1.2. Available information about drugs

Administrative databases usually records data in an aggregated manner. Thus, for our research purposes, we describe some of the information usually available, which is useful to organize data and compute quantitative indicators. In particular, we describe Anatomical Therapeutic Chemical codes and the information they provide as they allow to retrieve data related to a specific drug in the HF dataset, and present the concept of the duration of a prescription.

1.2.1. ATC codes

According to the European Medicines Agency website [16]

Definition 1.3. *(ATC code)*
The Anatomical Therapeutic Chemical code is a unique code assigned to a medicine according to the organ or system it works on and how it works. The classification system is

maintained by the World Health Organization (WHO).

ATC codes are composed of five different levels, each one of them provides different information as follows

- *first level*: consists of one letter, specifies the main anatomical group on which the drug acts;
- *second level*: consists of two digits, specifies the therapeutic subgroup;
- *third level*: consists of one letter, specifies the pharmacological/therapeutic subgroup;
- *fourth level*: consists of one letter, specifies the pharmacological/therapeutic/chemical subgroup;
- *fifth level*: consists of two digits, specifies the chemical substance;

As an example, Table 1.1 reports the deductible information from the different levels of the ATC code C09AA05 related to *ramipril*, an ACE inhibitor present in our dataset.

Level	**Code**	**Information**
first	C	Cardiovascular system
second	C09	Agents acting on the renin-angiotensin system
third	C09A	ACE Inhibitors, plain
fourth	C09AA	ACE Inhibitors, plain
fifth	C09AA05	Ramipril

Table 1.1: Example of the information retrievable from the different levels of the ATC code C09AA05 related to *ramipril*, an ACE inhibitor present is our dataset.

As mentioned, ATC codes play a fundamental role in administrative databases as they allow to classify prescriptions according to the different drug which they refer to. Besides of that, and in addition to the information embedded in the ATC encoding levels, the WHOCC (World Health Organization Collaborating Center for Drugs Statistics Methodology) website [63] provides an interactive tool to extract different useful information about a specific drug through its ATC code. In particular, Figure 1.1 shows the results of a search about the *ramipril*, characterized by code C09AA05. The tool provides information about the ATC levels, as well as the *defined daily dose* [38], the *administration route* and some additional notes.

C CARDIOVASCULAR SYSTEM
C09 AGENTS ACTING ON THE RENIN-ANGIOTENSIN SYSTEM

The DDDs are based on the treatment of mild-moderate hypertension.
See comments to C02L concerning the principles for assignment of DDDs for combined preparations.

C09A ACE INHIBITORS, PLAIN
All plain ACE inhibitors are classified in this group. No separate ATC codes are assigned for the active metabolites of the ACE inhibitors (e.g. enalaprilat, quinaprilat).
Combinations with diuretics, see C09BA - ACE inhibitors and diuretics.
Combinations with calcium channel blockers, see C09BB - ACE inhibitors and calcium channel blockers.
Combinations with beta blocking agents, see C09BX - ACE inhibitors, other combinations.

C09AA ACE inhibitors, plain

ATC code	Name	DDD	U	Adm.R	Note
C09AA05	ramipril	2.5	mg	O	

Figure 1.1: Search results in the WHOCC databases for ATC code C09AA05 related to *ramipril*.

1.2.2. Duration of a prescription

One of the most important information about each recorded prescription in administrative databases is represented by its duration. It expresses the number of *coverage days*, that correspond to the days in which the patient is supposed to consume the corresponding drug. As specified in the following, its importance is given by the fact that it is used to define quantitative indicators for patient's adherence to the pharmacological treatment. If the duration of a prescription is missing, there exists methods in order to exploit drug information (as the one retrieved in the *ramipril* example of Figure 1.1) to deduct it. We do not treat them as this is not our case, but an example can be found in [56].

1.3. Adherence

In order to assess the effect of a drug, the most common approach is to compare the clinical outcome of patients which effectively follow the treatment and patients which do not. In order to distinguish them, we rely on the concept of *adherence*, whose general definition is given in [5]

Definition 1.4. *Adherence (or compliance) generally refers to whether a patient takes a prescribed medication according to schedule.*

To quantify adherence through a numerical indicator, we adopt one of the most widely used methods in the pharmacoepidemiological literature [26], i.e., the *proportion of days covered* defined as

Definition 1.5. *(PDC) The Proportion of Days Covered is defined as*

$$PDC = \frac{number\ of\ distinct\ coverage\ days}{number\ of\ days\ in\ the\ observation\ period} \in [0, 1]. \tag{1.1}$$

Notice that the coverage days can be retrieved from a patient's prescriptions history, but overlapping days in different prescriptions of the same drug are considered only once. The *observation period* refers to a span of time, before the start of the actual survival study, in which the patient is alive and his pharmacological habits noted, and actually serves to define this adherence indicator. In our work we decided to adopt an observation period of 365 days, as suggested in [26], where a detailed overview about the concept of adherence and other possible ways to quantify it can be found. Finally, we use the PDC to define two alternative indicators of adherence status

- a *binary indicator* which specifies if a patient has to be considered adherent (1) or non adherent (0) to the treatment, based on the fact that its PDC reach a certain threshold, fixed at 0.8 according to literature:
  - non adherent (0) $0.00 \leq PDC < 0.80$
  - adherent (1) $0.80 \leq PDC \leq 1.00$;
- a *multi-levels indicator* which classifies patients in different levels, accordingly to their PDC with respect to the considered treatment:
  - level 1 (very low) $0.00 \leq PDC < 0.25$
  - level 2 (low) $0.25 \leq PDC < 0.50$
  - level 3 (medium) $0.50 \leq PDC < 0.75$
  - level 4 (high) $0.75 \leq PDC \leq 1.00$.

The detailed procedure is suitable to define an adherence quantifier variable to include in classical Survival analysis; however, the definition have to be extended when moving to a recurrent events framework. In this context, disposing of all the prescriptions in the study period, we would prefer an adherence quantifier capable of expressing the status of a

patient (i.e. whether he/she is adherent or not) all throughout its clinical course. For this reason, we develop a time dependent adherence quantifier, simply evaluating the PDC at the different patient's time t of interest, coinciding with events in its clinical course, and considering the observation period as an expanding time frame (see Section 2.3)

$$PDC(t) = \frac{\text{number of distinct coverage days up to time } t}{\text{number of days from index date to time } t} \in [0, 1]. \tag{1.2}$$

Based on this definition, the relative binary and multi-levels time-dependent indicators of adherence status at time t can then be calculated as mentioned above using the time-dependent $PDC(t)$.

1.4. ACE inhibitors

Pharmacological prescriptions in the Heart Failure dataset considered in our work refer to the disease-modifying drugs used in heart failure treatment [33], that are

- ACE Inhibitors (ACE)
- Angiotensin Receptors Blockers (ARB)
- Anti-Aldosterone Agents (AA)
- Beta-Blocking Agents (BB)
- Diuretics (DIU)

We decided to work in a monotherapy framework, focusing on the sole effect of ACE inhibitors on patient's hospitalizations and survival. The methodologies used in the analysis are adaptable to the other drugs mentioned above. However, it is important to stress the fact that, even if often adopted in practice, the monotherapy approach may result restrictive, since heart failure therapies usually involve different drugs at the same time.

Angiotensin-Converting Enzyme (ACE) inhibitors are medications that help relax the veins and arteries to lower blood pressure. ACE inhibitors prevent an enzyme in the body from producing angiotensin II, a substance that narrows blood vessels. Actually, in our analysis, we consider three different classes of ACE inhibitors, characterized by different third levels of their ATC codes, i.e., C09A, C09B or C09X as shown in Table 1.2.

ATC (3° level)	**Description**
C09A	Plain ACE Inhibitors
C09B	Combination of ACE Inhibitors
C09X	Other agents acting on the renin-angiotensin system

Table 1.2: Different classes of ACE inhibitors considered in our dataset and their characterizing ATC codes.

The first class refers to plain ACE inhibitors, while the second one to combined drugs in which ACE inhibitors represent the main component. Secondary components are, for example, diuretics, calcium channel blockers and beta blockers. The third class contains other agents which act on the renin-angiotensin system. They are treated as ACE inhibitors due to their similar effect, although they act on a different enzyme (e.g. renin inhibitors, which affects the production of angiotensin I instead of angiotensin II).

We now listed all the main pharmacoepidemiological concepts which are needed in the following analyses. In the next Chapter we will describe the datasets these analyses will be applied to.

2 | Heart Failure Dataset

In this Chapter we describe the dataset analyzed in our work, which was provided by *Regione Lombardia - Healthcare Division*, within the research project HFData (HFData-RF-2009-1483329) [47]. In Section 2.1 we present the Heart Failure (HF) dataset, specifying all the variables and information collected in it. Section 2.2 presents the cohort selection procedure which is adopted in our work, while Section 2.3 details the definition of auxiliary variables quantifying adherence to the considered treatment. Finally, Section 2.4 describes how we build the monotherapy dataset about ACE Inhibitors from the complete HF dataset.

2.1. Dataset Structure

The dataset to which we refer comes from an administrative database of Lombardy Region and reports information about patients which were hospitalized, due to heart failure, during the period from 2000 to 2012. In particular, it comprises hospitalization discharges, ambulatory care events, ER services and drugs prescriptions. A detailed description of the complete dataset and how it was built is available in [32].
Since drugs purchases are available only from 2006, we work on a reduced version composed by events from January 1st, 2006 ("2006-01-01", the study start date) until December 31st, 2012 ("2012-31-12", the study end date). Moreover, we consider only hospitalizations, removing records of other clinical events. This yields an initial dataset of 648,169 entries, related to 4,872 patients, which is composed of 625,388 pharmacological prescriptions and 22,781 hospitalization recordings.

At this initial stage, each entry of the considered dataset is characterized by different variables, which can be distinguished in three different levels: variables specifying information about the patient, variables related to an event (i.e. hospitalization or prescription) and variables reporting patient's clinical information at the event date (in case of hospitalizations).

Variable	Description
COD_REG	Patient's anonymous ID code
SESSO	Patient's gender
Eta_Min	Patient's age at first discharge due to HF
data_rif_ev	Date of discharge of first HF hospitalization
data_studio_out	Date of exit from the study
desc_studio_out	State of the patients at the end of the study. Labels are dead, truncated or lost.

Table 2.1: Variables at patient's level in the HF dataset.

Variable	Description	Hospitalization	Prescription
tipo_prest	Event classification code adopted in the original dataset	41	30
class_prest	String expressing additional notes	Cause of hospitalization	ATC code of the prescribed drug
data_prest	Date of event	Date of discharge	Date of prescription
qta_prest_Sum		Length of stay in hospital	Duration of the prescription

Table 2.2: Variables at event level in the HF dataset.

As reported in Table 2.1 the individual level characteristics are the patient's unique anonymous identification code, its gender, its age at the study entry, the date of enrolment in the study (i.e. and the date of discharge of the first hospitalization due to HF) and the date of exit from the study. Contextually, a label specifying the patient's survival status at the end of the study is provided. A patient is flagged as *dead* if he dies before the administrative end, December 31st, 2012; *truncated* if he is alive at such date; *lost* if he is lost to follow up (i.e. its traces are lost before the study ends).

Each record in the dataset is related to an event for which the patient-level variables share the same values. The specific information available for each event is coded through the variables shown in Table 2.2. The most important is the factor specifying if the entry refers whether to an hospitalization or to a drug prescription; accordingly, additional information are provided: the date of discharge and the length of stay in hospital in the first case, the date of prescription and its duration (see Section 1.2) in the second.

Finally, in case of a hospitalization event, further information about the patients' clinical history is given, consisting of binary factors specifying the onset of twenty common comorbidities at the time of each hospital discharge. Comorbidities are assessed as present

or absent according to the method proposed in [19], considering the conditions included either in Romano's adaptation of the Charlson index or in the Elixhauser system, as specified in Table 2.3. Notice that this set of variables will be involved in our analysis through the definition of a simple, summarizing comorbidity score, consisting in the sum of recorded comorbidities at each event.

Variable	Description	Comorbidity Index
metastatic	Binary flag specifying the presence of metastasis as a comorbidity	Romano
chf	Binary flag specifying the presence of CHF as a comorbidity	Romano
dementia	Binary flag specifying the presence of dementia as a comorbidity	Elixhauser
renal	Binary flag specifying the presence of renal issues as a comorbidity	Elixhauser
wtloss	Binary flag specifying the presence of weigth loss as a comorbidity	Elixhauser
hemiplegia	Binary flag specifying the presence of hemiplegia as a comorbidity	Romano
alcohol	Binary flag specifying the presence of alcohol abuse as a comorbidity	Elixhauser
tumor	Binary flag specifying the presence of tumours as a comorbidity	Romano
arrhythmia	Binary flag specifying the presence of arrhythmia as a comorbidity	Elixhauser
pulmonarydz	Binary flag specifying the presence of one or more pulmonary diseases as a comorbidity	Romano
coagulopathy	Binary flag specifying the presence of coagulopathy as a comorbidity	Elixhauser
compdiabetes	Binary flag specifying the presence of diabetes as a comorbidity	Elixhauser
anemia	Binary flag specifying the presence of anemia as a comorbidity	Elixhauser
electrolytes	Binary flag specifying the presence of electrolytes related issues as a comorbidity	Elixhauser
liver	Binary flag specifying the presence of peripheral vascular disease as a comorbidity	Elixhauser
pvd	Binary flag specifying the presence of liver issues as a comorbidity	Elixhauser
psychosis	Binary flag specifying the presence of psychosis as a comorbidity	Elixhauser
pulmcirc	Binary flag specifying the presence of pulmonary circulation issues as a comorbidity	Elixhauser
hivaids	Binary flag specifying the presence of HIV/AIDS as a comorbidity	Romano
hypertension	Binary flag specifying the presence of hypertension as a comorbidity	Elixhauser

Table 2.3: Additional information about comorbidities registered at each event. Column *Method* specifies according to which method, Romano's or Elixhauser's, the presence of the comorbidity was assessed.

2.2. Cohort selection

The next step is to select the appropriate cohort of patients for our analysis, starting from the HF dataset. In particular, in order to have an adequate observation period and properly assess the effect of the considered pharmacological treatment, we decide to select only those patients who survive at least one year after entering the study. This criterion is not strictly necessary when moving to the recurrent events framework, consistently with the corresponding redefinition of adherence (see Section 1.3), but we decide to stick with it for coherence, in order to consider the same cohort of patients all throughout our work. As index date we take for each patient the date of discharge due to his/her first HF hospitalization (provided by variable **data_rif_ev**). Moreover, only subjects who experienced at least one hospitalization and one pharmacological prescription are considered.

This selection yields a sample of 335,493 observations about 4,471 patients, comprehending 21,276 hospitalizations and 313,767 pharmacological prescriptions.

As we follow a monotherapy approach, the next step is represented by the splitting of the obtained dataset in different subdatasets, each one characterizing one of the disease-modifying drugs used in heart failure treatment [33]. We exploit ATC codes (see Section 1.2), encoded for pharmacological prescriptions events in variable **class_prest**, to build five different subdatasets: the considered drugs are ACE inhibitors (ACE), Angiotensin Receptors Blockers (ARB), Anti-Aldosterone agents (AA), Beta-Blockers (BB) and Diuretics (DIU). Each subdataset thus consists of records related to patients undergoing the considered treatment, where the events are drug prescriptions or hospitalizations. Table 2.4 reports the composition of the five obtained splitted datasets.

Drug	ATC codes	Hospitalizations	Prescriptions	Patients
ACE	3° level $\in \{C09A, C09B, C09X\}$	12,746	77,391	3,232
ARB	3° level $\in \{C09C, C09D\}$	10,349	34,474	1,958
AA	2° level = C07	16,760	76,512	2,601
BB	3° level $\in \{C03D, C03E\}$	13,612	34,783	3,333
DIU	3° level $\in \{C03C \mid ATC = C03BA08\}$	19,614	90,607	3,837

Table 2.4: Summary table of the HF subdatasets splitted by drugs.

2.3. Adherence Variables

As mentioned, our aim is to assess the effect of ACE inhibitors treatment on survival and hospitalizations in patients affected by Heart Failure. In our analysis we apply different tools (presented in the next Chapter) to model the process of patients' deaths and hospitalizations, including indicators of adherence to drug treatment as modelling variables. The adherence indicators presented in Section 1.3 are finally computed for the prescription data in each drug subdatasets.

Initially, for each patients we compute the *Proportion of days covered* (PCD) during the one year observation period, dividing the distinct number of coverage days by 365, and then define two variables as follows

$$\textbf{Adherent1Y} = \begin{cases} 0, & \text{if } PDC \in [0, 0.8) \\ 1, & \text{if } PDC \in [0.8, 0.1]; \end{cases} \tag{2.1}$$

$$\textbf{AdhLev1Y} = \begin{cases} 1, & \text{if } PDC \in [0, 0.25) \\ 2, & \text{if } PDC \in [0.25, 0.50) \\ 3, & \text{if } PDC \in [0.50, 0.75) \\ 4, & \text{if } PDC \in [0.75, 1]. \end{cases} \tag{2.2}$$

These two auxiliary variables are defined at patient's level as his/her binary/multilevels indicators of adherence to treatment during the one-year observation period and added to the relative drug subdatasets. Note also that the two indicators are highly collinear, so a choice is needed when considering them as possible covariates in a regression framework.

The adherence binary indicator is then extended to be included in the recurrent events framework, defining a time dependent variable as follows

$$\textbf{Adherent}(t_{event}) = \begin{cases} 0, & \text{if } PDC(t_{event}) \in [0, 0.8) \\ 1, & \text{if } PDC(t_{event}) \in [0.8, 0.1] \end{cases} \tag{2.3}$$

where t_{event} is the time instant relative to a hospitalization and $PDC(t_{event})$ is computed as specified in Equation 1.2. In this case, notice that the adherence variable is no more characteristic of a patient, as it may change its value from an event to another.

2.4. ACE Inhibitors dataset

To obtain the final dataset considered in our work, the ACE inhibitors subdataset is further refined, in order to express some of the available information in a suitable format. Since during our analysis we consider at first a classical survival approach and, then, a recurrent and terminal events approach (see Chapter 3), the dataset comprehends variables which serve in one or the other framework.

First of all, we need to keep track of time. The variable **timeEvent** expresses the number of days between the patient's study entry (i.e., index date) and each recorded hospitalization. It is computed as

$$\textbf{timeEvent} = \text{data\_prest} - \text{data\_rif\_ev}. \tag{2.4}$$

Notice that the first hospitalization discharge of each patient coincides with its study entry, so is characterized by a null value of this variable. To account for the terminal event (i.e., death or censoring) we compute time from the end of the observation period to the last observation or censoring, i.e., variable **timeOUT**, which make an amends for the observation period in order to not introduce bias in the classical survival analysis

$$\textbf{timeOUT} = \text{data\_studio\_out} - \text{data\_rif\_ev} - 365. \tag{2.5}$$

To distinguish a terminal event's nature, we introduce a dummy variable, **death**, which specifies whether the terminal event coincide with a patient's death or with censoring (see Section 3.1)

$$\textbf{death} = \begin{cases} 1, & \text{if terminal event corresponds to patient's death} \\ 0, & \text{if terminal event corresponds to censoring.} \end{cases} \tag{2.6}$$

Moreover, some auxiliary variables are then defined specifically to be considered in the classical survival framework. Variable **TotHosp1Y** counts the number of hospitalizations registered for a patient during its observation period (i.e. its first year of follow up). Variable **Comorbidity1st** represents the number of comorbidities at first hospitalization, whereas **MaxComorbidity** is the maximum number of comorbidities registered in the patient's first year of follow up. Variable **AgeMin** is the patient's age at the index date. In addition, other patient level variables included in the dataset are the gender (variable **Sex**) and the two previously defined adherence quantifiers (**Adherent1Y** and **AdhLev1Y**). Notice that these variables share the same value for all the recordings related to the same patient.

As far as the variables involved in the recurrent events analysis are concerned, variable **GapEvent** encodes the time between subsequent hospitalizations of a patient; for this reason, it is not defined for the first hospitalization record. The variable **event** specifies whether an entry is related to an hospitalization or a terminal event

$$\textbf{event} = \begin{cases} 1, & \text{if the entry corresponds to an hospitalization} \\ 0, & \text{if the entry corresponds to an terminal event.} \end{cases} \tag{2.7}$$

In addition, we also added a variable named **hosp** which counts the subsequent hospitalizations of each patient. For the recurrent events analysis, both terminal and hospitalization events are associated to the time dependent adherence binary indicator **Adherent**, computed as specified in Equation 2.3, in addition to two auxiliary variables, **AgeEvent** and **Comorbidity**, indicating respectively the age and the number of comorbidities (see Table 2.3) of the patient at each gap time.

As an example, Tables 2.5, 2.6 and 2.7 reports the data table corresponding to patient 10003004 in the described ACE inhibitors dataset, divided in

- Time variables and Event labels;
- Variables involved in the classical survival analysis;
- Variables involved in the recurrent and terminal events analysis.

A recap of the involved variables and further formatting choices are provided, if needed, in Chapter 4 before each analysis.

COD_REG	TimeEvent	TimeOUT	GapTime	event	death	hosp
10003004	0			1		1
10003004	229		229	1		2
10003004	360		131	1		3
10003004	528		168	1		4
10003004	881		353	1		5
10003004	2,024	1,659	1,143	0	1	//

Table 2.5: Time Variables and Event labels in the final ACE inhibitors dataset corresponding to patient 10003004.

COD_REG	Sex	Adherent1Y	AdhLev1Y	TotHosp1Y	Comorbidity1st	MaxComorbidity	AgeMin	TimeOUT	death
10003004	F	1	4	3	5	6	74	1,659	1

Table 2.6: Variables involved in the classical survival analysis in the final ACE inhibitors dataset corresponding to patient 10003004.

COD_REG	Sex	Adherent	Comorbidity	AgeEvent	hosp	GapEvent	event	death
10003004	F				1		1	
10003004	F	0	5	75	2	229	1	
10003004	F	1	6	75	3	131	1	
10003004	F	0	6	76	4	168	1	
10003004	F	0	7	77	5	353	1	
10003004	F	1	7	79	//	1,153	0	1

Table 2.7: Variables involved in the recurrent and terminal events analysis in the final ACE inhibitors dataset corresponding to patient 10003004.

Once described the selected cohort of patients and the final ACE subdatasets, in the next Chapter we will introduce the statistical methodologies we will use for our analyses.

3 | Methodologies

In this Chapter we present the statistical methodologies studied and developed for the analysis of the problem described in the previous Chapters. In particular, in Section 3.1 an introduction to Survival Analysis is presented, with a particular emphasis on the Cox Proportional Hazard model. Section 3.2 proposes an introduction to the framework of recurrent events, with a focus about Frailty models. Section 3.3 deals with jointly estimation of terminal and recurrent events through Cox-like models. Finally, Section 3.4 introduces our proposal, i.e. an original extension of state of the art methods, where a non parametric distribution for the multivariate frailty is assumed.

3.1. Survival Analysis in a nutshell

Survival Analysis is a collection of statistical procedures for data analysis in which the variable of interest is time until a specific event occurs. Such variable is known as *survival time*. As it measures the gap from a starting time, the origin, to an endpoint, the event of interest, such variable is often described as a *time-to-event* variable. The time unit of measure depends on the problem considered.
The term 'Survival Analysis' arises from the fact that, initially, the event of interest was death, but nowadays applications comprehends a variety of fields, such as industry, economic and social sciences. In this work we focus on a clinical application, modelling hospitalizations and deaths of patients subject to heart failure who undergo specific pharmacological treatments. The time origin coincides with the discharge of a patient after the first hospitalization due to heart failure, the endpoint is death of the subject and time is measured in days.
The rest of this section is meant to be a brief recall of fundamental concepts and techniques necessary to understand this work; for a complete and extensive treatment of the matter, see [27] or [21].

3.1.1. Censoring

Censoring is a key problem involved in time-to-event data. It occurs when some information is known about the survival time, but not its exact value. Different kind of censoring exists

1. *left censoring*: the event is known to happen before a time, but not exactly when;
2. *gap censoring*: the event is known to happen in a time interval, but not exactly when;
3. *right censoring*: the event is known to happen after a certain time, but not exactly when.

We will focus on *right censoring*, since it is the most common one and that of interest in this case. It usually occurs due to three reasons: when a subject does not experience the event before the study ends; when a subject is lost to follow up; when a subject withdraws from the study due to a different reason than that of interest [45]. This actually constitutes a big difference of time-to-event data with respect to most other statistical data, as specific techniques are needed to cope with the presence of partially observed instances.

The mathematical formulation of time-to-event data can be written as follows. For each subject i, let T_i^* be the non-negative random variable representing the true failure time and C_i the non-negative random variable denoting the time at which the censoring mechanism activates. The actual time-to-event datum observed in a right censoring context is whichever of these two variables is smaller:

$$T_i = min\{T_i^*, C_i\} \tag{3.1}$$

Of course, keeping track of the nature of each datum is necessary, thus each survival time is accompanied by a binary variable, known as *censoring variable*, defined as follows:

$$\delta_i = \begin{cases} 1 & if\ T_i^* \leq C_i \\ 0 & if\ T_i^* > C_i. \end{cases} \tag{3.2}$$

Thus, the simplest survival dataset is composed by a pair $O_i = (T_i, \delta_i)$ per each subject in the study.

The aforementioned formulation of censoring is very general, but it is important to stress that the vast majority of classical survival analysis applications assumes *independent* and *non informative censoring*. The former refers to the fact that censoring is independent from patients' characteristics (C_i is identically distributed for each subject i in the considered cohort), while the latter expresses the idea that the true survival time gives no information about censoring and viceversa (T_i^* and C_i are independent).

3.1.2. Survival and Hazard Functions

Let T be a non negative random variable characterized by density $f_T(t)$ and distribution function $F_T(t)$.

Definition 3.1. *(Survival Function). The survival function expresses the probability of a subject, whose survival time is distributed like T, to survive at least until time t:*

$$S_T(t) = 1 - \mathbb{P}(T \leq t) = 1 - F_T(t). \tag{3.3}$$

As defined above, the survival function shows complementary characteristics with respect to a distribution function: it assumes value one at $t = 0$, is non-increasing and tends to 0 as time passes.

Definition 3.2. *(Hazard Function). The hazard function quantifies the instantaneous risk of failure at time t, conditionally on surviving until that time:*

$$h_T(t) = \lim_{\Delta t \longrightarrow 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T > t)}{\Delta t}. \tag{3.4}$$

The hazard function is often referred to as hazard or mortality rate.
An expression for $h_T(t)$ is straightforward to derive if T is discrete:

$$\mathbb{P}(T = t_k) = f_T(t_k)\ \forall k = 1..K \implies S_T(t) = \sum_{k:t_k \geq t} f_T(t_k) \tag{3.5}$$

finally implying that:

$$h_T(t_k) = \mathbb{P}(T = t_k | T \geq t_k) = \frac{f_T(t_k)}{S_T(t_k)} \tag{3.6}$$

However, T is usually considered continuous. In this case, the derivation of the instanta-

neous hazard expression is as follows

$$\begin{aligned} h_T(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \cdot \frac{\mathbb{P}(T \in [t, t + \Delta t) \cap T \geq t)}{\mathbb{P}(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{1}{\mathbb{P}(T \geq t)} \cdot \frac{T \in [t, t + \Delta t)}{\Delta t} \\ &= \frac{1}{\mathbb{P}(T \geq t)} \cdot \frac{\int_t^{t+\Delta t} f_T(u) du}{\Delta t} \\ &= \frac{f_T(t)}{S_T(t)} \\ &= -\frac{d}{dt} \log(S_T(t)) \end{aligned} \tag{3.7}$$

where: third equality comes from the fact that $T \in [t, t + \Delta t) \implies T \geq t$; fourth equality is possible assuming T admits density $f_T(t)$; fifth equality is obtained using the Fundamental Theorem of Calculus; last passage to logarithmic derivative can done noticing that $-f_T(t)$ is the derivative of $S_T(t)$.
Finally, integrating from 0 to t and recalling that $S_T(0) = 1$

$$S_T(t) = e^{-H_T(t)} = e^{-\int_0^t h_T(u) du} \tag{3.8}$$

where $H_T(t)$ is the cumulative hazard function defined as

Definition 3.3. *(Cumulative Hazard Function). The cumulative hazard function is defined as the integral function of the hazard rate.*

$$H_T(t) = \int_0^t h(u) du \tag{3.9}$$

Equation 3.8 describes the relationship between the survival function and the instantaneous hazard rate: being the latter related to the inverse of the derivative of the former, it states the simple fact that a sharp decrease in the survival probability implies higher mortality rates.

The simplest approach to approximate the survival function is, given a set of lifetime data, to build the *empiric survival function*. It looks at the percentage of individuals in the cohort who are still event free at time t; in this case, however, partially observed data

cannot be encoded properly. Notice that the empiric survival function is not guaranteed to reach zero (it only happens if every subject in the cohort experiences failure).

3.1.3. Kaplan-Meier Estimator & Nelson-Aelen Estimator

The Kaplan-Meier estimator takes his name from the two mathematicians who proposed it in 1958 [25]. Also known as *product limit estimator*, it is a nonparametric statistic which approximates the survival curve in case of censored lifetime data. Its validity depends on three assumption: non-informative and non-interval censoring and survival probabilities stationarity over the study period.
Consider the sequence of true, unique event times $t_1 < t2 < .. < t_J$ (do no account for censored observations times). Accordingly, define d_i as the number of observed events at t_i and n_i the number of patients at risk at t_i. The KM estimator is expressed by:

$$\hat{S}_{KM}(t) = \prod_{i:t_i \geq t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.10}$$

Notice that the term in the multiplication, fixed i, is an estimate for the conditional probability of surviving at time t_i. This estimator is basically a step function with jumps at observed death times and, if no censored observations are present in the considered dataset, it reduces to the empirical survival function. Asymptotic Gaussian confidence intervals can be built once an estimate for the variance is known. It can be retrieved through Greenwood's formula:

$$\mathrm{Var}(\hat{S}_{KM}(t)) = (\hat{S}_{KM}(t))^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{3.11}$$

The Nelson-Aelen estimator was initially proposed by Nelson in 1969 [36], and later applied by Aelen in the field of survival data [1]. It is a non parametric estimator for the cumulative hazard rate which can deal with partially observed data. Within the same framework described above:

$$\hat{H}_{NA}(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i} \tag{3.12}$$

where each term in the summation can be interpreted as the ratio of deaths over the number of exposed patients at t.

In 1972 Breslow proposed an estimator for the survival function plugging 3.12 in 3.8 [7]

$$\hat{S}_B(t) = \exp\{-\hat{H}_{NA}(t)\} = \prod_{i:t_i \leq t} \frac{d_i}{n_i}. \tag{3.13}$$

An estimate for the variance of Equation 3.12 is represented by $\mathrm{Var}(-\log(S_B(t)))$, which can be obtained with Greenwood's formula.

Nelson-Aelen and Kaplan-Meier estimators are asymptotically convergent to the same quantity and quite close to each other, especially when the deaths-exposed ratio is small. However, the first shows generally lower variance and higher bias (above all when S(t) approaches zero) with respect to the second one.

3.1.4. Cox Proportional Hazard Model

In the proposed context, a popular and established approach to model the effect of multiple risk factors, both categorical and continuous, over a (possibly censored) time-to-event outcome is the Cox proportional hazard model [21, 27].

Let us consider a set of n patients, subject to p different risk factors. Let $\boldsymbol{X} \in \mathbb{R}^p$ be the random vector encoding these risk factors, which can be both continuous and categorical. For each patient $i = 1, ..., N$, collected data are constituted by observed failure (or censoring) times, censoring variables and realizations of $\boldsymbol{X}_i$. The Cox model aims at quantifying the effect of the risk factors on the hazard function, which, for each patient, is expressed as

$$h_i(t|\boldsymbol{X}_i) = h_0(t)e^{\boldsymbol{X}_i^T \boldsymbol{\beta}} \tag{3.14}$$

where $h_0(t)$ is a unspecified non-negative function, known as *baseline hazard function*, and $\boldsymbol{\beta}$ is the coefficient vector which need to be estimated.

Cox regression model is referred to as *proportional hazard* model since the Hazard Ratio (HR) for two subjects with covariate vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ (formula 3.15) is constant over time

$$\begin{aligned} HR_{ij} &= \frac{h_i(t|\boldsymbol{X}_i)}{h_j(t|\boldsymbol{X}_j)} \\ &= \frac{h_0(t)\exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}}{h_0(t)\exp\{X_j^T\boldsymbol{\beta}\}} \\ &= \exp\{(\boldsymbol{X}_i - \boldsymbol{X}_j)^T\boldsymbol{\beta}\} \end{aligned} \tag{3.15}$$

and as *semiparametric*, since it makes no assumption regarding the nature of the baseline hazard function itself. Therefore, an approximation is needed in order to recover an estimate of the survival function.

The Cox model put an emphasis on the effect of covariates considering the so-called hazard ratio quantity[57], defined for each of them as

Definition 3.4. *(Hazard Ratio of a covariate)*

$$HR_l = \exp\{\boldsymbol{\beta}_l\} \quad \forall l = 1, ..., p \tag{3.16}$$

In particular, covariates with a hazard ratio significantly less than one are believed to have a decreasing effect on the hazard (or an increasing effect on the survival probabilities), and are called *protective factors*. On the contrary, hazard ratios significantly greater than one identify *risk factors*. The interpretation of hazard ratios slightly differs if it refers to a categorical or to a numerical variable. In the first case, it expresses the grouping effect with respect to a level chosen as baseline. Multilevel factors work the same way, as each level is compared to the chosen baseline. In the second case, the hazard ratio refers to a unitary increment of the variable.

Coefficients estimation is based on the partial likelihood procedure introduced by Cox [13], which involves only probabilities related to subjects which actually experience a failure. In particular, considering

- $\boldsymbol{O}_i = (t_i, \delta_i, \boldsymbol{x}_i)$ as the observed data for subject i;
- $R(t) = \{i : t_i < t\}$ as the risk set at time t (i.e., the set of individuals at risk at time t);
- $t_1 < t_2 < .. < t_K$ as the sequence of ordered (distinct) failure time with K = total number of observed events;

the *partial likelihood* can be written as the product, over the observed failure times, of the conditional probabilities of the observed individual being chosen from the risk set to fail

Definition 3.5. *(Cox partial likelihood)*

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{j=1}^{K} \mathbb{P}(subject\ j\ fails\ at\ t_j \mid one\ failure\ in\ R(t)\ at\ t_j) \\ &= \prod_{j=1}^{K} \frac{exp\{\boldsymbol{X}_j^T \boldsymbol{\beta}\}}{\sum_{i \in R(t_j)} exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}\}} \end{aligned} \tag{3.17}$$

The complete derivation of the full and partial likelihood can be found in any survival text, such as [27], or in the original work of Cox [13]. The estimation is then performed maximizing the *log partial likelihood*

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{K} \left[\boldsymbol{X}_j^T \boldsymbol{\beta} - ln \left(\sum_{l \in R(t_j)} exp\{\boldsymbol{X}_l^t \boldsymbol{\beta}\} \right) \right] \tag{3.18}$$

from which the estimates $\hat{\boldsymbol{\beta}}$ are retrieved through iterative/gradient based optimization algorithms. Notice that the assumption of distinct failure times is often violated, but ties are easily handled with slightly correction in the likelihood formulation (see the works of Breslow [7] and Efron [15] for further details).

Given $\hat{\boldsymbol{\beta}}$, an estimator for the baseline cumulative hazard was proposed by Breslow

$$\hat{H}_0(t) = \sum_{j:t_j<t} \left(\frac{1}{\sum_{l \in R(t_j)} exp\{\boldsymbol{X}_l^T \hat{\boldsymbol{\beta}}\}} \right) \tag{3.19}$$

where j is the index of the usual sequence of observed ordered failure times and $R(t_j)$ the set of individuals at risk at instant t_j. Then, an estimate of the baseline survival function can be obtained using Equation 3.8

$$\hat{S}_0(t) = e^{-\hat{H}_0(t)} = \prod_{j:t_j<t} \left(1 - \frac{1}{\sum_{l \in R(t_j)} exp\{\boldsymbol{X}_l^T \hat{\boldsymbol{\beta}}\}} \right) \tag{3.20}$$

and an estimate for the survival function relative to patient i, $S_i(t)$, can be obtained as

follows

$$\hat{S}_i(t|\boldsymbol{X}_i) = \hat{S}_0(t)^{\exp\{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}\}}. \tag{3.21}$$

Finally, goodness of fit and the proportional hazard assumption must be checked in order to ensure the validity of the presented approach. That is done usually looking at the behaviour of martingale or deviance residuals for the former and Schoenfeld residuals for the latter (see [55] and [17] for further details).

3.1.5. Cox models and time dependent covariates

Time dependent covariates are covariates that changes over time. As an example in our context, we can think of the number of known comorbidities (with respect to Heart Failure) of a patient along his/her follow up time. Cox model can deal natively with such kind of data, and the reason why lies in the theory of martingale and counting processes upon which the model is formulated. At this stage, we will not deepen into such theory (see [24]), but instead focus on a programming trick used to encode time-dependent data which will be useful when dealing with recurrent events (Section 3.2).

Such trick uses intervals of time to replicate different time-to-outcome instances from the time-dependent measurements available for a single subject. Consider a patient i with follow up time T_i, for which K measurements of the time-dependent variable are known. The first measurement is supposed to be taken at $t_i^{(0)} = 0$, when the patient enters the study, while the last at $t_i^{(K-1)} \leq T$. Considering the sequence $0 = t_i^{(0)} < t_i^{(1)} < ... < t_i^{(K-1)} \leq T_i$, time intervals can be defined breaking the subject's follow up time into $K-1$ intervals open on the left and closed on the right, encoding the available data for patient i in $K-1$ different instances. For each instance, a binary variable, the *status variable*, defines whether or not the subject experiences the event at the end of the relative time interval. Notice that all the instances related to the same patient are characterized by the same values of the fixed-time covariates. Table 3.1 shows a simple example.

The estimation process does not change. Even if it may seem that we have repeated measurements per each patient, we do not need to account for correlated data, since the likelihood equations at any time point use only one copy of any subject.

	tstart	*tstop*	*status*	*FT*	*TD*
subject A	$t^{(0)}$	$t^{(1)}$	0	1	1
	$t^{(1)}$	$t^{(2)}$	0	1	2
	$t^{(2)}$	$t^{(3)}$	1	1	3

Table 3.1: Time intervals encoding for a generic subject A for which a fixed time (FT) and a time dependent (TD) covariate are available. Columns *tstart* and *tstop* define the time intervals, while *status* expresses whether the subject experience(1) or not an event at the end of the related time interval.

3.2. Recurrent Events Framework

It is often the case that the events of interest related to each statistical unit may occur more than once over time. Such kind of events are referred to as *recurrent events*. Examples in the epidemiological and medical fields comprehend important phenomena like re-hospitalizations, tumour relapses and post-surgery infections.

Survival Analysis, in particular Cox proportional hazard models, are a popular choice to assess the characteristics of a recurrent events process, but the classical formulation (see Section 3.1.4) is suitable for treating time-to-first-event only. Its application in a context of recurrent events would lead to a loss of information and power due to the discarding of all occurrences subsequent to the first one for each subject. Figure 3.1 provides a comparative scheme to visualize a simple example of recurrent data of four different subjects which highlights this problem: top panel shows that subjects 1 and 3 experiences, respectively, 1 and 3 events and are then censored, while subjects 2 and 4 are only censored (right censoring in considered, see Section 3.1 for a recall on censoring). Lower panel shows the same data as seen by a classical Cox model, which would neglect second event for subject 1 and second and third events for subject 3, as well as their censoring times.

From the above example we can also notice that, in a recurrent events framework, each subject is characterized by a terminal censoring event, which may be due to loss of follow-up, death (subject 1 and 4 in the above example) or administrative censoring (subject 2 and 3). However, their distinction is not accounted for and censoring is always considered as non-informative. This can represent a drawback in our context, since a patient's death is an event of primary importance in biomedical studies: this implies that treating time-to-death data like censored observations is too restrictive. For this reason, often

two different processes (one generating repeated events and one related to deaths) are modelled starting from the same set of observations. In first approximation, the modelling of terminal events can be done independently from the recurrent events one using classical survival techniques. This can result still too simplistic, since the recurrent events processes considered are often informative of death. Because of that, the joint modelling of the two processes is an interesting solution encompassing these limitations, which will be considered in Section 3.3.

Table 3.2 provides a visualization of the data encoding related to subject 1 and 3 of previous example, useful to keep in mind when presenting the models and to point out some important characteristics. Multiple events are ordered and can be labeled accordingly (column *counts*). Column *status* is the usual binary variable to identify proper events (1) or censoring (0); in addition, *death* helps to keep track of the nature of censoring, assigning to censoring times 1 for deaths and 0 otherwise. Accordingly to the application considered, two different timescales can be considered: times-to-events, also known as *total times* or *calendar times*, obtained dividing the follow up time in intervals using the repeated events similarly as seen in Section 3.1.5, or times-between-events (*gap times*), computed as differences between subsequent times-to-event. In Table 3.2, calendar times are encoded through columns *tstart* and *tstop*, while gap times through column *gap*. Thanks to this particular formatting of data, time dependent variables are encoded naturally in the framework of recurrent events: in this example, a unique covariate (*treatment*) is considered, specifying whether a patient in under treatment or not when experiencing an event.

A wide literature about recurrent events modelling has flourished in past years and several different approaches have been proposed, each one of them based on different premises and with a particular focus. In the following, two historically important models are presented, the one proposed by Andersen and Gill [3] and the one proposed by Prentice, Williams and Peterson [43]. This is done to provide context and justify the introduction of frailty models (Section 3.2.1), which constitute the framework chosen to develop joint models for our application. Examples of comparative studies to deepen the understanding of the recurrent events framework, involving the cited approaches and others more, are [2], [64] and [39].

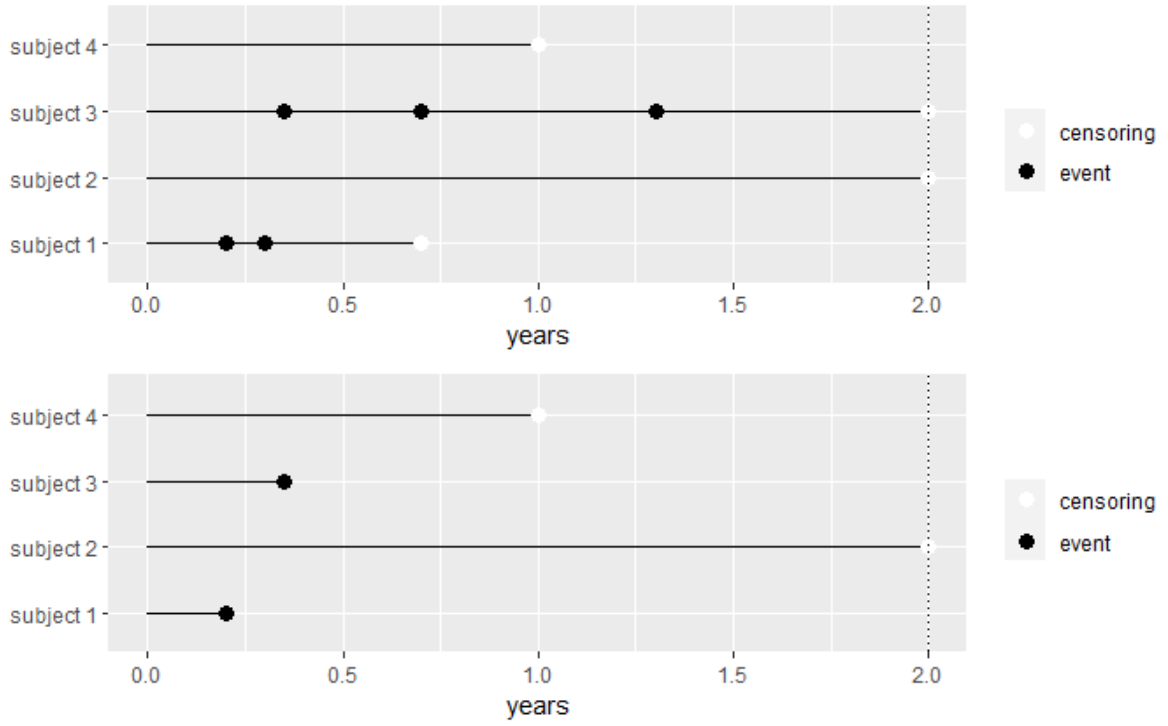

Figure 3.1: Comparative scheme for recurrent events data relative to four different subjects (expressed in years). Top panel shows occurrences of the event of interest as black dots, while white dots represent (right) censoring. Lower panel shows the same data as seen by classical Survival Analysis methods presented in Section 3.1. The vertical dotted line represents administrative censoring.

	gap	*tstart*	*tstop*	*count*	*status*	*death*	*treatment*
subject 1	0.2	0.0	0.2	1	1		0
	0.1	0.2	0.3	2	1		0
	0.4	0.3	0.7		0	1	1
subject 3	0.4	0.0	0.4	1	1		0
	0.3	0.4	0.7	2	1		1
	0.6	0.7	1.3	3	1		1
	0.7	1.3	2.0		0	0	1

Figure 3.2: Encoding of recurrent data of subjects 1 and 3 of the example shown in Figure 3.1. Variables are defined as follows: *time* represents total times of events, while *tstart* and *tstop* define gap times; *count* traces the number of events experienced by the subject; *status* indicates whether the occurrence refers to a proper event(1) or censoring(0), while *death* specifies if censoring is due to death(1) or not(0); *treatment* is a binary variable expressing if the subject is under treatment at event time(1) or not(0).

Andersen & Gill Model

In 1982 Andersen and Gill [3] proposed one of the first models able to deal with recurrent events. It aims at modelling the intensity function, which is a modification of the instantaneous hazard function (Section 3.1)

Definition 3.6. *(Intensity Function). The intensity function quantifies the instantaneous risk of failure at time t, without any assumption about the process history:*

$$i(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t)}{\Delta t} \tag{3.22}$$

where T is the random variable expressing the event time. A&G proposed to model the intensity function of each patient in a Cox like manner

$$i_i(t) = i_0(t) e^{\boldsymbol{X}_i^T \boldsymbol{\beta}} \tag{3.23}$$

where $i_0(t)$ is an unspecified non-negative function, known as *baseline intensity function*, $\boldsymbol{\beta}$ is the coefficient vector which needs to be estimated and $\boldsymbol{X}_i$ is the covariates vector related to subject i. Time T is computed with respect to the study entry, i.e. calendar times are considered in the following. The fundamental assumption of the model, known as *proportional intensity assumption*, is that the instantaneous risk of experiencing an event at time t (expressed by the intensity function) remains the same whether or not a previous event has occurred. It implies that times-to-event are conditionally independent, given the covariates vector. Sometimes this property is often formulated saying that, in the A&G model, the correlation between multiple observation is completely explained by the time dependent variables used as regression covariates. If such hypothesis holds, each recorded event time contributes to the estimation procedure, without accounting if it was experienced by a subject or another. The estimation procedure is the same as in the Cox model, but the risk set definition needs to be modified as follows

Definition 3.7. *Risk set in the Andersen and Gill model is defined as follows*

$$R^{AG}(t) = \{i \in \{1, ..., N\} : \exists j \in \{1, ..., k_i\} \ st \ T_{ij} \geq t\} \tag{3.24}$$

where T_{ij} are distinct event times for each subject i, N is the total number of subjects and k_i identifies the a priori unknown number of observed events by subject i. If the proportional intensity assumption is not fulfilled, the application of the A&G model may

be misleading, since the estimated coefficients' standard deviation is likely to be underestimated [2]. In that case, possible solutions are variance robust methods (see [12] and [29]) or the introduction of random effects (see Section 3.2.1).

Prentice, Williams & Peterson Model

Prentice, Williams and Peterson [43] proposed in 1981 two alternative Cox-based approaches to account for recurrent events, one dealing with total times and another with gap times, based on the procedure of *stratification* according to the order number of repeated events. The intuition behind them consists in modelling a different Cox-like hazard function per each stratum of recurrent events: stratum 1 consists in all first event times, stratum 2 in second event times, etc. Thus, for subject i the hazard for the j-th recurrent event is given by

$$h_{ij}^{Tot}(t) = h_{0j}^{Tot}(t) \exp\{\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_j^{Tot}\} \tag{3.25}$$

where $h_{0j}^{Tot}(t)$ and $\boldsymbol{\beta}_j^{Tot}$ are, respectively, the baseline hazard function and the estimated coefficients vector specific for the jth recurrent event. Tot refers to the fact that total times are considered. The formulation in case of gap times is analogous:

$$h_{ij}^{Gap}(t) = h_{0j}^{Gap}(t - t_{j-1}) \exp\{\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_j^{Gap}\}. \tag{3.26}$$

Assuming that each subject i in the considered cohort experiences k_i events, the Prentice-Williams-Peterson (PWP) approach basically consists in fitting k different Cox models, where, in principle, $k = max_i(k_i)$. However, in practice the parameter k is often set to a lower number, since available data are likely to be too few to provide good estimates for high order number events. The main difference with the A&G model is that this allows the hazards for a recurrent event to change after a previous event: being an individual at risk for the j-th event only if it experienced the previous one, the hazard for a certain recurrence is computed conditionally on the entire previous events. The estimation procedure is again similar to the standard Cox one: the partial likelihood can be obtained by the product of each stratum partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} L_j(\boldsymbol{\beta}_j) \tag{3.27}$$

$$L_j(\boldsymbol{\beta}_j) = \prod_{i=1}^{M} \frac{exp\{\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_j\}}{\sum_{l \in R_j^{PWP}(t_{ij})} \exp\{\boldsymbol{X}_l^T \boldsymbol{\beta}_j\}} \tag{3.28}$$

where $\boldsymbol{\beta}$ is a matrix whose rows contains each stratum coefficients vector ($\boldsymbol{\beta}_j$), t_{ij} are the distinct ordered failure times, with i=1,...,M and $M \leq N$ (number of subjects). The PWP risk set is defined as

$$R_j^{PWP}(t) = \{l, l = 1, ..., N | T_{l(j-1)} \leq t \leq T_l\} \tag{3.29}$$

for total times, or

$$R_j^{PWP}(t) = \{l, l = 1, ..., N | (T_{lj} - T_{l(j-1)}) \geq t\} \tag{3.30}$$

for gap times.

When events of interest are subsequent re-hospitalizations of patients subject to a particular disease, these two types of extensions of the Cox model are appealing, due to their simple characterization of risk factors. However:

- gap times are often preferred to characterize the recurrent process;
- a comprehensive model (focused on the intensity of the recurrent process) is preferred, rather than a stratified one.

The ideal choice would be an Andersen and Gill formulation for gap times. However, it would require conditional (on the covariate vector) independence of the gap times, a condition which is almost never met in practice. Actually, when passing from calendar to gap times, a major problem is represented by correlation among times at patient level due to induced dependent censoring. This phenomenon is intrinsic in the gap times formulation and happens even when the corresponding calendar times are independent. To provide a very simple explanation in our case, we can think that a longer first event time would imply a greater probability of censoring for the following events, as an older patient is likely to be more fragile. Induced dependent censoring is a complicate matter and has been extensively studied (see, for example, [28] and [23]), but for the scope of our work we can simply notice that it produces within-subject gap times correlation which is likely to cannot be explained by observed covariates only.

3.2.1. Frailty Models

As we have seen, the main assumption in the Cox model is independence of survival times from one observational unit to another, given the observed values of covariates. The same concept is extended to multivariate failure times in the Andersen and Gill model. However, often this assumption is not plausible: in the univariate case, hidden structures in the sampling scheme may be present, or units may be clustered leading to unobserved covariates (e.g. groups of patients may have unobserved genetic or environmental factors in common); in the repeated events case, within-individual correlation between events times is common, and, as mentioned, when gap times are involved it is also enhanced by induced dependent censoring.

Random effects survival models [22], often called *frailty models*, exploit the introduction of a random unobserved covariate (the frailty) that describes the excess risk specific of each distinct patient, accounting for within-subject correlated times and, in general, for heterogeneity unexplained by the observed set of covariates. In this framework, observed times are assumed to be independent conditionally on the covariate vector and on the unobserved random effects. *Proportional hazard frailty models* assume a shape for the conditional hazards similar to the Cox model, including moreover the action of the patient-specific frailties. These random variables are assumed to follow a probability distribution, the shape of which is described with a few parameters, and can act multiplicatively or additively on hazards depending on the specific model.

The shared frailty model [22] is one of the most famous proportional frailty models. Consider subject i ($i = 1, ..., N$), t_{ij} his/her j-th observed times (which can be both total or gap times) and δ_{ij} the binary indicator for recurrent events which is 0 if the observation is censored and 1 if is properly observed. Let $\boldsymbol{X}_{ij}$ be the vector of covariates (fixed or time-dependent) for individual i at event j and h_0 the common baseline hazard function. The hazard function for subject i at observed time t_{ij} is given by

$$h_i(t_{ij}|\boldsymbol{X}_{ij}, w_i) = h_0(t_{ij}) w_i \exp\{\boldsymbol{X}_{ij}^T \boldsymbol{\beta}\} \tag{3.31}$$

where $\boldsymbol{\beta}$ is the vector of estimated coefficients and w_i is the frailty related to patient i, shared among all the i-th patient's observed times. Initially, frailties are assumed to be independent and identically distributed according to some distribution, characterized by a parameter vector $\boldsymbol{\theta}$. One common example is the gamma distributed shared frailty

model, where the expectation is set to 1 and the variance is the inverse of the parameter to be estimated

$$w_i \overset{iid}{\sim} \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right) \tag{3.32}$$

The characteristics of the gamma distributed shared frailty model have been widely investigated (see, for example, [4]), also considering other frailties distributions. It also represents the basis upon which more complex models have been built, like the nested frailty model, which can deal with hierarchically clustered data through nested random effects [50], or cure models, which admit the possibility of being cured after experiencing a recurrent event [44].

An alternative approach is proposed in [48], where a model with multivariate normally distributed random effects and its estimation procedure are presented. Let t_{ij} be the collection of observed times for units $i, i = 1, ..., N$ and event $j, j = 1, ..., k_i$, and δ_{ij} the relative censoring variables. To simplify the notation, consider the times grouped by i and ordered by j as indexed by a unique variable l going from 1 to $\sum_i k_i = M$. Then, considering $\boldsymbol{b}$ a vector of frailties of dimension $q \leq N$, the conditional hazard for each observation (indexed by l according to the new notation) can be written as

$$h_l(t|\boldsymbol{b}, \boldsymbol{X}_l) = h_0(t) \exp\{\boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b}\} \tag{3.33}$$

where $\boldsymbol{\beta}$ is the usual vector of estimated coefficients, $\boldsymbol{X}_l$ is the l-th row of the fixed effects design matrix (of dimension $M \times p$)) and $\boldsymbol{Z}_l$ the l-th row of the random effects design matrix (of dimension $M \times q$), which assign to each subject (previously indexed by i) the proper effect. The model assumes that $\boldsymbol{b}$ follows a zero-mean multivariate Normal distribution, characterized by covariance matrix $\boldsymbol{E}$

$$p(\boldsymbol{b}; \boldsymbol{E}) = \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{E}) \tag{3.34}$$

with $\boldsymbol{0}$ the q-dimensional zero vector. In principle, the simplest shape of $\boldsymbol{E}$ is given by θI_N (where I_N is the $N \cdot N$ identity matrix and θ the only parameters which need to be estimated) in the case of independent, identically distributed random effects per each subject.

Assuming independent and non-informative censoring conditionally on $\boldsymbol{b}$, the marginal

full likelihood can be written integrating out the random effects

$$L(h_0(t), \boldsymbol{\beta}, \boldsymbol{\theta}) = \int \prod_{l=1}^{M} h_l(t|\boldsymbol{b})^{\delta_l} S_l(t|\boldsymbol{b}) p(\boldsymbol{b}; \boldsymbol{E}) d\boldsymbol{b} \tag{3.35}$$

$$\begin{aligned} &= \int \prod_{l=1}^{M} \left[h_0(t) \exp\{\boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b}\}\right]^{\delta_l} \\ &\quad \text{x } \exp[-H_0(t) \exp\{\boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b}\} \\ &\quad \text{x } \frac{(2\pi)^{-\frac{q}{2}}}{\sqrt{det(\boldsymbol{E})}} \exp\{-\frac{1}{2} \boldsymbol{b}^T \boldsymbol{E}^{-1} \boldsymbol{b}\} d\boldsymbol{b} \end{aligned} \tag{3.36}$$

where $H_0(t) = \int_0^t h_0(u)du$ is the usual cumulative hazard baseline function. This integral does not admit an analytical solution. In their work [48], Ripatti and Palmgren used a Laplace approximation to obtain an approximate log marginal likelihood

$$\begin{aligned} l(h_0(t), \boldsymbol{\beta}, \boldsymbol{\theta}) \approx &- \frac{1}{2} \log(|\boldsymbol{E}|) \\ &- \frac{1}{2} \log \left(\left| \sum_{l=1}^{M} H_0(t) \exp\{\boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b}\} \boldsymbol{Z}_l \boldsymbol{Z}_l^T - \boldsymbol{E}^{-1} \right| \right) \\ &+ \sum_{l=1}^{M} \delta_l \{\log(h_0(t)) + \boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b}\} \\ &- H_0(t) \exp\{\boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b}\} - \frac{1}{2} \boldsymbol{b}^T \boldsymbol{E}^{-1} \boldsymbol{b} \end{aligned} \tag{3.37}$$

where $|E|$ stands for the matrix determinant. They noticed that rows 3 and 4 of Equation 3.37 represent the full likelihood of a Cox model plus a penalization term for extreme values of $\boldsymbol{b}$. Assuming that this term encloses almost all the information about the regression coefficients, it is possible to obtain estimates for $(\boldsymbol{\beta}, \boldsymbol{b})$ maximizing the relative penalized partial likelihood

$$PPL(\boldsymbol{\beta}, \boldsymbol{b}) = \sum_{l=1}^{M} \delta_l \left\{ \boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b} - \log \sum_{k \in R(t_l)} \exp\{\boldsymbol{X}_k^T \boldsymbol{\beta} + \boldsymbol{Z}_k^T \boldsymbol{b}\} \right\} - \frac{1}{2} \boldsymbol{b}^T \boldsymbol{E}^{-1} \boldsymbol{b} \tag{3.38}$$

Partially differentiating in $\boldsymbol{\beta}$ and $\boldsymbol{b}$ we obtain

$$\sum_{l=1}^{M} \delta_l \left\{ \boldsymbol{X}_l - \frac{\boldsymbol{X}_l \exp\{\boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b}\}}{\sum_{k \in R(t_l)} \exp\{\boldsymbol{X}^T \boldsymbol{\beta} + \boldsymbol{Z}_k^T \boldsymbol{b}\}} \right\} = 0 \tag{3.39}$$

$$\sum_{l=1}^{M} \delta_l \left\{ \boldsymbol{Z}_l - \frac{\boldsymbol{Z}_l \exp\{\boldsymbol{X}_l^T \boldsymbol{\beta} + \boldsymbol{Z}_l^T \boldsymbol{b}\}}{\sum_{k \in R(t_l)} \exp\{\boldsymbol{X}_k^T \boldsymbol{\beta} + \boldsymbol{Z}_k^T \boldsymbol{b}\}} \right\} = \boldsymbol{E}^{-1} \boldsymbol{b} \tag{3.40}$$

which, given an initial estimate for $\boldsymbol{\theta}$, can be solved through iterative methods. Once obtained, the coefficients estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}})$ can be plugged into Equation 3.37 to retrieve an approximate profile likelihood [10] for $\boldsymbol{\theta}$, which can be maximized in closed form solving the following system of equations

$$-\frac{1}{2} \left\{ tr\big(\boldsymbol{E}^{-1} \frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\theta}}\big) + tr\big(K''_{PPL}(\hat{\boldsymbol{b}})^{-1} \frac{\partial \boldsymbol{E}^{-1}}{\partial \boldsymbol{\theta}}\big) + tr\big(\hat{\boldsymbol{b}}^T \boldsymbol{E}^{-1} \frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\theta}} \boldsymbol{E}^{-1} \hat{\boldsymbol{b}}\big) \right\} = \boldsymbol{0} \tag{3.41}$$

where $K''_{PPL}(\hat{\boldsymbol{b}}) = \frac{\partial^2 PPL}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T}|_{\boldsymbol{b}=\hat{\boldsymbol{b}}}$ and tr is the trace operator (see [48] for further details and variance formula of the coefficients estimates). The complete estimation routine is sketched in Algorithm 3.1.

Algorithm 3.1 Ripatti and Palmgren Estimation procedure pseudo-code

1: Given initial estimate $\hat{\boldsymbol{\beta}}$,$\hat{\boldsymbol{b}}$,$\hat{\boldsymbol{\theta}}$
2: Set $converged = False$, $MaxIt = 10000$, $it = 0$
3: **while** !$converged$ and $it < MaxIt$ **do**
4: Given actual $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{\theta}}$, solve Eq 3.39 via Newton-Raphson
5: Given actual $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, solve Eq 3.40 via Newton-Raphson
6: Given actual $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$, solve Eq 3.41
7: compute $converged$ and update it
8: **end while**
9: Compute variance of obtained estimates

The discussed model represents an assessed tool in the study of recurrent events and has been implemented in the well-optimized R package `coxme`[59]. For this reason we decide to apply the model presented in Equation 3.33 to both the recurrent event and the terminal event processes. In this way, the two trained models will serve us as a basis for assessing the effect of joint estimation, where we try to make the recurrent event process informative for the terminal event. Obtained results will be compared with the joint models presented in the following sections.

3.3. Joint Models

As previously stated, successive events across time (e.g. hospitalizations) may be terminated by loss to follow-up, administrative censoring or death. Moreover, in the context of survival analysis death is an event of major interest, thus is required to be modeled. Since it could be correlated with recurrent events, the usual assumption of non-informative censoring of the recurrent event process by censoring due to death may be not satisfied. Joint frailty models allow to study the joint evolution over time of two survival processes by considering the major terminal event (i.e. death) as informative censoring (see [23] for an introduction about this topic), usually under an integrated likelihood approach [30] or in a bayesian setting [40].

In the following are detailed the joint frailty model proposed by Rondeau et al. in 2007[51], which represent an assessed tool in the field, and a joint frailty model recently developed by Ng et al. [37]. Suppose to follow a cohort of N patients, denoted by index i ($i = 1, ..., N$), in which each one experiences k_i events, denoted by index j with $j = 1, ..., k_i$. Let T_{ij}^R be the times of recurrent events and T_i^D the terminal event time for unit i, both subject to right censoring, whose time is denoted by C_i. Define $T_{ij} = \min\{T_{ij}^R, T_i^G, C_i\}$ $\forall j = 1, ..., k_i$ the sequence of random variables defining observed times for subject i and t_{ij} their realizations, δ_{ij}^R the corresponding censoring variables for the recurrent event (1 if $T_{ij} = T_{ij}^R$, 0 otherwise) and δ_i^D the censoring variable for the terminal event related to subject i (1 if $T_{ik_i} = T_i^D$, 0 otherwise). $\boldsymbol{X}_{ij}^R \in \mathbb{R}^{p_1}$ denotes the vector of covariates (either fixed or time dependent) associated to recurrent events for subject i, while $\boldsymbol{X}_i^D \in \mathbb{R}^{p_2}$ the one associated to its terminal events. Superscript T stands as usual for vector transposition.

3.3.1. Joint Frailty Model by Rondeau et al.

Rondeau et al. [51] proposed to model the hazard rates of the two process separately, but admitting a common frailty term (w_i) related to individuals (like in the shared frailty model) to account for heterogeneity. Such term acts differently on the hazard rates by means of the parameter α

$$\begin{cases} h_{ij}^R(t|w_i, \boldsymbol{X}_{ij}^R) = h_0^R(t) w_i \exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R\} \\ h_i^D(t|w_i, \boldsymbol{X}_i^D) = h_0^D(t) w_i^\alpha \exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D\} \end{cases} \tag{3.42}$$

where $h_0^R(t)$ and $h_0^D(t)$ are the baseline common hazard for recurrent events and death,

respectively, while $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ the vectors of estimated coefficients. Time t can either refers to calendar times or gap times. The random effects w_i are supposed, initially, to be independent and identically distributed according to a Gamma of unknown parameter θ:

$$w_i \overset{iid}{\sim} \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right). \tag{3.43}$$

An alternative distribution of the frailty proposed by Rondeau et al. is the log-Normal distribution. In this case, the model is reparametrized as follows:

$$\begin{cases} h_{ij}^R(t|\eta_i, \boldsymbol{X}_{ij}^R) = h_0^R(t)\exp\{\eta_i + \boldsymbol{\beta}^T\boldsymbol{X}_{ij}^R\} \\ h_i^D(t|\eta_i, \boldsymbol{X}_i^D) = h_0^D(t)\exp\{\alpha\eta_i + \boldsymbol{\gamma}^T\boldsymbol{X}_i^D\} \end{cases} \tag{3.44}$$

$$\eta_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{3.45}$$

The estimation approach starts considering the marginal full log likelihood with respect to random effects. It can be obtained as

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_0^R(t), h_0^D(t), \tilde{\boldsymbol{\theta}}, \alpha) &= \log\{L(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_0^R(t), h_0^D(t), \tilde{\boldsymbol{\theta}}, \alpha)\} \\ &= \log\left\{\prod_{i=1}^{N} L_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_0^R(t), h_0^D(t), \tilde{\boldsymbol{\theta}}, \alpha)\right\} \end{aligned} \tag{3.46}$$

where $\tilde{\boldsymbol{\theta}}$ is the vector of parameters defining the chosen distribution of random effects and L_i is the contribution to the full likelihood related to individual i, $i = 1, ..., N$

$$\begin{aligned} L_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_0^R(t), h_0^D(t), \tilde{\boldsymbol{\theta}}, \alpha) = \int \prod_{j=1}^{k_i} & h_{ij}^R(t|w_i)^{\delta_{ij}^R} S_{ij}^R(t|w_i) \text{ x} \\ & h_i^D(t|w_i)^{\delta_i^D} S_i^D(t|w_i) p(w_i|\tilde{\boldsymbol{\theta}}) dw_i. \end{aligned} \tag{3.47}$$

In Equation 3.47:

- $p(w_i|\tilde{\boldsymbol{\theta}})$ refers to the density function of the chosen distribution for the random effects;
- $h_{ij}^R(t|w_i)$ and $h_i^D(t|w_i)$ to the hazards of recurrent event and death, respectively, considering the random effects as fixed;
- $S_{ij}^R(t|w_i)$ and $S_i^D(t|w_i)$ to the survival functions of recurrent events and death, re-

spectively, considering the random effects as fixed.

The exact form depend from the chosen random effects distribution, the model parametrization and the choice of the timescale. As an example, considering calendar times and the model specified by Equation 3.42 and Equation 3.43 it results

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, h_0^R(t), h_0^D(t), \theta, \alpha) = \sum_{i=1}^{N} \left\{ \left[\sum_{j=1}^{k_i - 1} \delta_{ij}^R \log \left[h_0^R(t) w_i \exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R\} \right] \right] + \right. \\ \left. + \delta_i^D \log \left[h_0^D(t) w_i^\alpha \exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D\} \right] - \log \Gamma \left(\frac{1}{\theta} \right) - \frac{1}{\theta} \log \theta + \log I \right\} \tag{3.48}$$

with

$$I = \int_0^\infty w_i^{(k_i - 1) + \alpha \delta_i^D + \frac{1}{\theta} - 1} \exp \left\{ - w_i H_0^R(t) \exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R\} - w_i^\alpha H_0^D(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D\} - \frac{w_i}{\theta} \right\} dw_i$$

where $H_0^R(t) = \int_0^t h_0^R(u)du$ and $H_0^D(t) = \int_0^t h_0^D(u)du$ are the cumulative baseline functions of the two processes. Moreover, the authors argued that smooth baseline hazard functions have to be preferred, and thus proposed to add a term to penalize roughness in the estimation process:

$$pl(\Phi) = l(\Phi) - k_1 \int_0^\infty h_0^{R''}(t)dt - k_2 \int_0^\infty h_0^{D''}(t)dt \tag{3.49}$$

where $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, h_0^R(t), h_0^D(t), \boldsymbol{\theta}, \alpha)$, k_1 and k_2 are two hyperparameters which governs the degree of penalization, $h_0^{R''}(t)$ and $h_0^{D''}(t)$ are, respectively, the second derivative of the baseline hazard function for recurrent events and death. The optimization procedure is carried out by means of the iterative method of Levenberg and Marquardt [42]. Since baseline hazard functions and their derivatives are not known, an approximation through M-splines is considered, with the number of knots Q representing another hyperparameter of the model. First and second derivatives are approximated using finite differences. Finally, since integrals in the full log likelihood are often not analytically tractable, they are evaluated through Gaussian quadrature using Laguerre polynomials with 20 points (see [51] and [52] for further details).

The described model represents one of the first attempts to link the recurrent events process to the terminal events one, but is still used in practice, having been implemented

in the R package `frailtypack` [53]. However, its characterization of the random effects is a bit too simplistic and it is hard to train, given the high number of hyperparameters to tune during the estimation process.

3.3.2. Joint Frailty Model by Ng et al.

In [37], Ng et al. propose a different approach for jointly modelling the recurrent and terminal events processes, considering two correlated random effects for the hazard rates for recurrent and terminal events, modeled using a multivariate Gaussian. In particular, the authors denote u_i and v_i the frailty related to the patient i to account for correlation within recurrent event times and for heterogeneity in the death process, respectively. Their joint distribution is

$$p\big([u_i, v_i]|\boldsymbol{\mathcal{E}}\big) = \mathcal{N}_2(\mathbf{0}, \boldsymbol{\mathcal{E}}) \tag{3.50}$$

where $\boldsymbol{\mathcal{E}}$ is expressed as

$$\boldsymbol{\mathcal{E}} = \begin{bmatrix} \theta_u^2 & \rho\theta_u\theta_v \\ \rho\theta_u\theta_v & \theta_v^2 \end{bmatrix} \tag{3.51}$$

This formulation allows a well-defined, straightforward interpretation of all the parameters involved, being θ_u^2 and θ_u^2 the quantifiers of unobserved heterogeneity in the two processes, while ρ models their dependence. In this framework, the two processes are simply modeled as

$$\begin{cases} h_{ij}^R(t|u_i, \boldsymbol{X}_{ij}^R) = h_0^R(t)\exp\{\eta_{ij}\} = h_0^R(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}^R + u_i\} \\ h_i^D(t|w_i, \boldsymbol{X}_i^D) = h_0^D(t)\exp\{\zeta_i\} = h_0^D(t)\exp\{\boldsymbol{\gamma}^T\boldsymbol{X}_i^D + v_i\} \end{cases} \tag{3.52}$$

where η_{ij} and ζ_i are the linear predictors related to patient i and event j, while t refers to a gap times timescale. The estimation procedure is developed in the Generalized Linear Mixed Model (GLMM) framework: starting from BLUP (Best Linear Unbiased Predictor) estimation [49], they argued that estimators for $\boldsymbol{\Omega} = [\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{u}, \boldsymbol{v}] \in \mathbb{R}^{p_1+p_2+2N}$ can be obtained maximizing a likelihood with shape

$$L(\boldsymbol{\Omega}) = L_1(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{u}, \boldsymbol{v})L_2(\boldsymbol{u}, \boldsymbol{v}|\theta_u, \theta_v, \rho) \tag{3.53}$$

where L_1 is the usual Cox partial likelihood assuming random effects as fixed intercepts, while L_2 is the random effects density, given its characterizing parameters. The relative

log likelihood l can be expressed as the sum of two components, l_1 and l_2

$$l_1 = \sum_{k=1}^{K} \delta_k^R \Big\{ \eta_k - \sum_{l \in R^R(t_k)} \exp\{\eta_k\} \Big\} + \sum_{n=1}^{N} \delta_n^D \Big\{ \zeta_n - \sum_{l \in R^D(t_n)} \exp\{\eta_l\} \Big\} \tag{3.54}$$

$$l_2 = -\frac{1}{2} \Big\{ N \log(2\pi |\boldsymbol{E}|) + \boldsymbol{q}^T \boldsymbol{E}^{-1} \boldsymbol{q} \Big\} \tag{3.55}$$

where: K is the total number of recurrent events experienced by subjects in the considered cohort ($K = \sum_{i=1}^{N} k_i$); N is the total number of subjects; $R^R(t)$ and $R^D(t)$ are, respectively, the risk sets at time t for recurrent event and terminal event (where t denotes a gap time, similarly as in Equation 3.30); $\boldsymbol{q} = [\boldsymbol{u}, \boldsymbol{v}]$ is a $2 \cdot N$ column vector, built stacking recurrent events process random effects of each subject on top of terminal event process random effects for each subject; $\boldsymbol{E}$ is the variance-covariance matrix (with determinant $|\boldsymbol{E}|$) characterizing the distribution of $\boldsymbol{q}$, which is

$$p(\boldsymbol{q}|\boldsymbol{E}) = \mathcal{N}_{2 \cdot N}(\boldsymbol{0}, \boldsymbol{E}) \tag{3.56}$$

with $\boldsymbol{E} = \boldsymbol{\mathcal{E}} \bigotimes \boldsymbol{I_N}$ and $\bigotimes$ standing for the kronecker product. Given estimates for $\boldsymbol{\Phi} = [\theta_u, \theta_v, \rho]$, the defined log likelihood is maximized through the Newton-Raphson iterative method

$$\hat{\boldsymbol{\Omega}}^{r+1} = \hat{\boldsymbol{\Omega}}^r + \boldsymbol{G}^{-1} \frac{\partial l}{\partial \boldsymbol{\Omega}} \Big|_{\boldsymbol{\Omega} = \hat{\boldsymbol{\Omega}}^r, \boldsymbol{\Phi} = \hat{\boldsymbol{\Phi}}} \tag{3.57}$$

where $\frac{\partial l}{\partial \boldsymbol{\Omega}} = [\frac{\partial l}{\partial \boldsymbol{\beta}}, \frac{\partial l}{\partial \boldsymbol{\gamma}}, \frac{\partial l}{\partial \boldsymbol{u}}, \frac{\partial l}{\partial \boldsymbol{v}}]$ and G is the information matrix corresponding to log likelihood l, i.e.

$$\boldsymbol{G} = \begin{bmatrix} -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{u}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{v}^T} \\ -\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{u}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{v}^T} \\ -\frac{\partial^2 l}{\partial \boldsymbol{u} \partial \boldsymbol{\beta}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{u} \partial \boldsymbol{\gamma}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{u} \partial \boldsymbol{v}^T} \\ -\frac{\partial^2 l}{\partial \boldsymbol{v} \partial \boldsymbol{\beta}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{v} \partial \boldsymbol{\gamma}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{v} \partial \boldsymbol{u}^T} & -\frac{\partial^2 l}{\partial \boldsymbol{v} \partial \boldsymbol{v}^T} \end{bmatrix} \tag{3.58}$$

Similarly, its inverse can be expressed in a convenient block form

$$\boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{\beta},\boldsymbol{\beta}} & \boldsymbol{B}_{\boldsymbol{\beta},\boldsymbol{\gamma}} & \boldsymbol{B}_{\boldsymbol{\beta},\boldsymbol{u}} & \boldsymbol{B}_{\boldsymbol{\beta},\boldsymbol{v}} \\ \boldsymbol{B}_{\boldsymbol{\gamma},\boldsymbol{\beta}} & \boldsymbol{B}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} & \boldsymbol{B}_{\boldsymbol{\gamma},\boldsymbol{u}} & \boldsymbol{B}_{\boldsymbol{\gamma},\boldsymbol{v}} \\ \boldsymbol{B}_{\boldsymbol{u},\boldsymbol{\beta}} & \boldsymbol{B}_{\boldsymbol{u},\boldsymbol{\gamma}} & \boldsymbol{B}_{\boldsymbol{u},\boldsymbol{u}} & \boldsymbol{B}_{\boldsymbol{u},\boldsymbol{v}} \\ \boldsymbol{B}_{\boldsymbol{v},\boldsymbol{\beta}} & \boldsymbol{B}_{\boldsymbol{v},\boldsymbol{\gamma}} & \boldsymbol{B}_{\boldsymbol{v},\boldsymbol{u}} & \boldsymbol{B}_{\boldsymbol{v},\boldsymbol{v}} \end{bmatrix} \tag{3.59}$$

It is worth to notice that in this estimation procedure, like in the classical Cox model and in Ripatti and Palmgren model, the baseline functions can be left unspecified. Once obtained $\hat{\boldsymbol{\Omega}}$, Ng et al. proposed to follow a REML (Restricted Maximum Likelihood, [8, 18]) approach to update estimates for $\boldsymbol{\Phi}$. This is done solving the equation of 1st order derivative of the REML log likelihood with respect to $\boldsymbol{\Phi}$, which yields

$$\operatorname{tr}\left(\boldsymbol{E}^{-1}\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\Phi}}\right) + \operatorname{tr}\left(\left(\boldsymbol{B}_{\boldsymbol{q},\boldsymbol{q}} + \boldsymbol{q}\boldsymbol{q}^T\right)\frac{\partial \boldsymbol{E}^{-1}}{\partial \boldsymbol{\Phi}}\right) = \mathbf{0} \tag{3.60}$$

with tr denoting the trace operator, $\boldsymbol{q} = [\boldsymbol{u}, \boldsymbol{v}]$ and $\boldsymbol{B}_{\boldsymbol{q},\boldsymbol{q}}$ the matrix composed by the four blocks related to random effects of $\boldsymbol{G}^{-1}$. Algorithm 3.2 sketches the complete estimation routine. Once convergence is reached, the four blocks related to linear predictors coefficients of $\boldsymbol{G}^{-1}$ represents an asymptotic estimator of the variance-covariance matrix of the coefficients estimates

$$\operatorname{var}\left(\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\gamma}}\end{bmatrix}\right) = \begin{bmatrix}\boldsymbol{B}_{\boldsymbol{\beta},\boldsymbol{\beta}} & \boldsymbol{B}_{\boldsymbol{\beta},\boldsymbol{\gamma}}\\ \boldsymbol{B}_{\boldsymbol{\gamma},\boldsymbol{\beta}} & \boldsymbol{B}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}\end{bmatrix} \tag{3.61}$$

An asymptotic estimator for the variance-covariance matrix of $\boldsymbol{\Phi}$ is instead obtained inverting the REML loglikelihood information matrix, as specified in [37].

Algorithm 3.2 Model Estimation procedure by Ng et al.

1: Given initial estimate $\hat{\boldsymbol{\Omega}}$,$\hat{\boldsymbol{\Phi}}$
2: Set $converged = False$, $MaxIt = 10000$, $it = 0$
3: **while** !$converged$ and $it < MaxIt$ **do**
4: Maximize $l = l_1 + l_2$ (Eq 3.54) via Newton-Raphson to update $\hat{\boldsymbol{\Omega}}$
5: Given actual $\hat{\boldsymbol{\Omega}}$, solve Equation 3.60 to update estimates of $\boldsymbol{\Phi}$
6: compute $converged$ and update it
7: **end while**
8: Compute standard errors of obtained estimates from relative asymptotic variance-covariance matrices

The described model represents a state of the art tool to deal with recurrent and terminal events. As a matter of fact, a dedicated R package is yet not available, reason why we had to implement the proposed estimation routine by ourselves. One of the main strengths of the model is its elegant characterization of random effects, described through a bivariate Normal distribution. However, in practice (e.g. in a provider's assessment),

an easier characterization (e.g. a discrete distribution) may outperform it, resulting more insightful. We decided to investigate in this direction, adopting for the random effects a non parametric distribution with discrete support.

3.4. Discrete Nonparametric Frailty

In this section we propose our original model, in which the hazard rates are modeled similarly as in Section 3.3.2, but assuming a bivariate non parametric discrete distribution for the frailties. This choice stems from the fact that a discrete distribution of frailties allows a further level of interpretation in the medical practice, in addition to the opportunity of uncovering and analyze a latent partition in the considered cohort of patients. The model is presented alongside a specific EM algorithm for its training, which was inspired by [20].

3.4.1. Notation

Consider a cohort of N patients for which is known the clinical history of recurrent and terminal events. In particular, for each patient i $(i = 1, ..., N)$ we know:

- $\boldsymbol{t}_i$ the vector of observed gap times, which in turn can be split in
  - $t_{ij}^R,\ j = 1, ..., k_i$ the recurrent events gap times composing clinical history of patient i;
  - t_i^D the last gap time in clinical history of patient i before the terminal event.
- $\boldsymbol{\delta}_i$ the vector of censoring variables related to $\boldsymbol{t}_i$, which in turn can be split in
  - $\delta_{ij}^R,\ j = 1, ..., k_i$ the recurrent events gap times censoring variables;
  - δ_i^D the terminal event gap time censoring variable.
- $\boldsymbol{\mathcal{O}}_i$ the collection of observed covariates for patient i at times $\boldsymbol{t}_i$, which in turn can be split in
  - $\boldsymbol{X}_{ij}^R,\ j = 1, ..., k_i$ the vectors of covariates observed at each recurrent events gap time;
  - $\boldsymbol{X}_i^D$ the vector of covariates observed at terminal event gap time.

  Notice that $\boldsymbol{X}_{ij}^R$ and $\boldsymbol{X}_i^D$ can be composed by different kind of variables.

3.4.2. Model formulation

As mentioned, hazard rates for the recurrent events process and terminal event process are modeled similarly as in Ng et al. (Section 3.3.2) and Ripatti and Palmgren (Section 3.2.1)

$$\begin{cases} h_{ij}^R(t_{ij}^R) = h_0^R(t_{ij}^R)\exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R + u_i\} \\ h_i^D(t_i^D) = h_0^D(t_i^D)\exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D + v_i\} \end{cases} \tag{3.62}$$

where, as usual: $h_0^R(t)$ and $h_0^D(t)$ are, respectively, the recurrent events and terminal events hazard baselines; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are, respectively, the coefficients vectors for the recurrent and terminal events processes; T denotes vector transposition. In this formulation, random effects u_i and v_i are distributed according to P^*, which is an unknown measure on $\mathbb{R}^2$

$$[u, v]_i \overset{iid}{\sim} P^* \quad \forall i = 1, ..., N. \tag{3.63}$$

Such measure in supposed to be discrete and with a finite support, thus it can be characterized by a vector of L points in $\mathbb{R}^2$, $\boldsymbol{P}$, and a vector of relative weights, $\boldsymbol{w}$

Support	$\boldsymbol{P}$	P_1	P_2	...	P_L
Weight	$\boldsymbol{w}$	w_1	w_2	...	w_L

Notice that each weight expresses the probability of a patient to be assigned to a certain point l, $l = 1, ..., L$

$$w_l = \mathcal{P}([u, v]_i = P_l) \quad \forall i = 1, ..., N \tag{3.64}$$

for this reason, the sum of the weights is constrained to be unitary. Moreover, notice that the number of points constituting the support of the distribution (i.e. L) is assumed to be unknown a priori.

3.4.3. Likelihood

Initially, let us consider L as fixed. In order to define the likelihood of the proposed model, it is useful to introduce a set of auxiliary random variables. In particular, for each subject i, define a random vector $\boldsymbol{z}_i$ as follows

$$\boldsymbol{z}_i = [z_{i1}\ z_{i2}\ z_{i3}\ ...\ z_{iL}] \tag{3.65}$$

where

$$z_{il} = \begin{cases} 1 & if\ [u,v]_i = P_l \\ 0 & otherwise \end{cases}$$

Thus, each auxiliary vector is distributed according to a multivariate Bernoulli distribution of parameters $\boldsymbol{w}$. If we suppose to having observed the realizations of such auxiliary random vectors (collected in the random matrix $\mathcal{Z}$), we can write the likelihood of the model as

$$L^{full/par}(\boldsymbol{\Omega}; data|\mathcal{Z}) = \prod_{i=1}^{N}\prod_{l=1}^{L}\left[L_i^{full/par}(\boldsymbol{\Omega}; data|[u,v]_i = P_l)\right]^{z_{il}} \tag{3.66}$$

where $\boldsymbol{\Omega}$ denotes the set of all parameters we need to estimate. An important model design choice is whether to consider a full likelihood or a partial likelihood in the estimation process. Following some considerations reported in Appendix A, we decide to adopt a full likelihood approach; thus, our aim is to maximize

$$L(\boldsymbol{\Omega}; data|\mathcal{Z}) = \prod_{i=1}^{N} L_i(\boldsymbol{\Omega}; \boldsymbol{t_i}, \boldsymbol{O_i}, \boldsymbol{\delta_i}|\boldsymbol{z}_i) \tag{3.67}$$

where $\boldsymbol{\Omega} = [\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{w}, \boldsymbol{P}, H_0^R(t), H_0^D(t)]$ are the quantities to be estimated, while *data* refers to the overall set of observed gap times, censoring variables and covariates. Recall that $H_0^R(t)$ and $H_0^D(t)$ are the cumulative baseline hazards related to the two processes. Each individual contribution to the likelihood can be written as the following product

$$L_i(\boldsymbol{\Omega}; \boldsymbol{t}_i, \boldsymbol{O}_i, \boldsymbol{\delta}_i|\boldsymbol{z}_i) = \prod_{j=1}^{k_i}\left[L_{ij}^R(\boldsymbol{\Omega}; t_{ij}^R, O_{ij}^R, \delta_{ij}^R)\right]\ \text{x}\ L_i^D(\boldsymbol{\Omega}; t_i^D, O_i^D, \delta_i^D) \tag{3.68}$$

where, respectively

$$L_{ij}^R(\boldsymbol{\Omega}; t_{ij}^R, O_{ij}^R, \delta_{ij}^R|\boldsymbol{z}_i) = \prod_{l=1}^{L}\Big\{ w_l\left[h_0^R(t_{ij}^R)\exp\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}^R + P_l^{(1)}\}\right]^{\delta_{ij}^R}\ \text{x} \\ \exp\left[-H_0^R(t_{ij}^R)\exp\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}^R + P_l^{(1)}\}\right]\Big\}^{z_{il}} \tag{3.69}$$

$$L_i^D(\boldsymbol{\Omega}; t_i^D, O_i^D, \delta_i^D | \boldsymbol{z}_i) = \prod_{l=1}^{L} \Big\{ w_l \big[h_0^D(t_i^D) \exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D + P_l^{(2)}\} \big]^{\delta_i^D} \text{ x} \\ \exp\big[- H_0^D(t_i^D) \exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D + P_l^{(2)}\} \big] \Big\}^{z_{il}}. \tag{3.70}$$

We can then express the loglikelihood, always assuming $\mathcal{Z}$ as known, in a convenient way as the sum of three terms

$$l(\boldsymbol{\Omega}; data | \mathcal{Z}) = l_w(\boldsymbol{\Omega}_w; data | \mathcal{Z}) + l_R(\boldsymbol{\Omega}_R; data | \mathcal{Z}) + l_D(\boldsymbol{\Omega}_D; data) | \mathcal{Z}) \tag{3.71}$$

defined as

$$l_w(\boldsymbol{\Omega}_w; data | \mathcal{Z}) = \sum_{l=1}^{L} \sum_{i=1}^{N} z_{il} \log(w_l) \tag{3.72}$$

$$l_R(\boldsymbol{\Omega}_R; data | \mathcal{Z}) = \sum_{l=1}^{L} \sum_{i=1}^{N} z_{il} \Bigg[\sum_{j=1}^{k_i} \delta_{ij}^R \big[log(h_0^R(t_{ij}^R)) + \boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R + P_l^{(1)} \big] \\ - H_0^R(t_{ij}^R) \exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R + P_l^{(1)}\} \Bigg] \tag{3.73}$$

$$l_D(\boldsymbol{\Omega}_D; data | \mathcal{Z}) = \sum_{l=1}^{L} \sum_{i=1}^{N} z_{il} \Bigg[\delta_i^D \big[log(h_0^D(t_i^D)) + \boldsymbol{\gamma}^T \boldsymbol{X}_i^D + P_l^{(2)} \big] \\ - H_0^D(t_i^D) \exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D + P_l^{(2)}\} \Bigg] \tag{3.74}$$

where, respectively, $P_l^{(1)}$ and $P_l^{(2)}$ stand for the abscissa and the ordinata of the chosen support point $\boldsymbol{P}_l$.

3.4.4. Expectation-Maximization algorithm

Since the loglikelihood in Equation 3.71 depends on a set of random variables (z_{il}) it can not be directly maximized. The standard procedure in such cases is to resort to an Expectation-Maximization strategy [6]. EM algorithms are iterative methods consisting in two steps: in order to obtain estimates for the quantities of interest (here represented by $\boldsymbol{\Omega}$), the loglikelihood is firstly averaged with respect to the random variables, considering the parameters of interest fixed to their previous estimates, and then maximized.

Expectation step

In our case, the E step consists in computing at each iteration

$$\mathbb{E}_{\mathcal{Z}|\hat{\mathbf{\Omega}}}\left[l(\mathbf{\Omega}; data)\right] \tag{3.75}$$

which can then be maximized in order to update estimates for $\hat{\mathbf{\Omega}}$. Looking at Equations 3.72-3.74, we can notice that they involve the auxiliary random variables z_{il} multiplied by quantities independent from them. Thus, the expectation step coincides with computing

$$\mathbb{Z}_{il} = \mathbb{E}\left[z_{il}|\hat{\mathbf{\Omega}}, data\right] \tag{3.76}$$

and plugging them into loglikelihood Equation 3.71. Given current estimates for Ω, the $\mathbb{Z}_{il}$ can be computed in closed form using Bayes theorem

$$\begin{aligned} \mathbb{Z}_{il} &= \mathbb{E}\left[z_{il}|\hat{\mathbf{\Omega}}, \boldsymbol{t}_i, \mathcal{O}_i\right] \\ &= \mathcal{P}(z_{il} = 1||\hat{\mathbf{\Omega}}, \boldsymbol{t}_i, \mathcal{O}_i) \\ &= \frac{\mathcal{P}(\boldsymbol{t}_i|z_{il} = 1, \hat{\mathbf{\Omega}}, \mathcal{O}_i)\mathcal{P}(z_{il} = 1)}{\sum_{k=1}^{L} \mathcal{P}(\boldsymbol{t}_i|z_{ik} = 1, \hat{\mathbf{\Omega}}, \mathcal{O}_i)\mathcal{P}(z_{ik} = 1)}. \end{aligned} \tag{3.77}$$

Recall that, fixed L and for each i, the auxiliary variables z_{il} are Bernoulli distributed of parameters w_l and the probability of the observed gap times for subject i is

$$\mathcal{P}(\boldsymbol{t}_i|z_{il} = 1, \hat{\mathbf{\Omega}}, \mathcal{O}_i) = \prod_{j=1}^{k_i} \left[h_i^R(t_{ij}^R)^{\delta_{ij}^R} S_i^R(t_{ij}^R)\right] \text{ x } h_i^D(t_i^D)^{\delta_i^D} S_i^D(t_{ij}^D) \tag{3.78}$$

where $h^R(t)$ and $h^D(t)$ are the recurrent events and terminal event hazards in Equations 3.62, while $S^R(t)$ and $S^D(t)$ are the relative survival functions (see Section 3.1). Given that, after some simplifications, Equation 3.77 finally yields

$$\mathbb{Z}_{il} = \frac{w_l \exp\left\{\sum_{j=1}^{k_i}\left(\delta_{ij}^R P_l^{(1)} - H_0^R(t_{ij}^R)\exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R + P_l^{(1)}\}\right) + \delta_i^D P_l^{(2)} - H_0^D(t_i^D)\exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D + P_l^{(2)}\}\right\}}{\sum_{k=1}^{\hat{k}} w_k \exp\left\{\sum_{j=1}^{L}\left(\delta_{ij}^R P_k^{(1)} - H_0^R(t_{ij}^R)\exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R + P_k^{(1)}\}\right) + \delta_i^D P_k^{(2)} - H_0^D(t_i^D)\exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D + P_k^{(2)}\}\right\}}. \tag{3.79}$$

It is worth to notice that, since the $\mathbb{Z}_{il}$ represent the probabilities of assigning patient i to

point l given the current state of parameters, from them we can extract a latent partition of subjects.

Maximization step

The update rules for $\hat{\boldsymbol{\Omega}}$ can now be obtained. At this stage, it is useful to notice that each one of the three terms involved in Equation 3.71 depends on a disjoint subset of parameters composing $\boldsymbol{\Omega}$. In particular

- l_w involves $\boldsymbol{w}$;
- l_R involves $\boldsymbol{\beta}$, $H_0^R(t)$ and the abscissa of the points composing the support of the discrete distribution;
- l_D involves $\boldsymbol{\gamma}$, $H_0^D(t)$ and the ordinata of the points composing the support of the discrete distribution.

Thus, the maximization of the averaged loglikelihood can be carried out separately with respect to these three terms. The first one yields the update rule for $\boldsymbol{w}$. Recalling that the sum of weights must be unitary, using Lagrangian optimization we obtain

$$\hat{w}_l^{(up)} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{Z}_{il} \quad \forall l = 1, ..., L \tag{3.80}$$

Since l_R involves multiple parameters we adopt a multi-step approach in the optimization. In particular, keeping $\hat{\boldsymbol{\beta}}$ and $\hat{H}_0^R(t)$ fixed to their previous estimates, we obtain the update rule for the abscissa of the support points solving the relative partial derivative equation

$$\hat{P}_l^{(1)(up)} = \log\left[\frac{\sum_{i=1}^{N}\mathbb{Z}_{il}\sum_{j=1}^{k_i}\delta_{ij}^R}{\sum_{i=1}^{N}\mathbb{Z}_{il}\sum_{j=1}^{k_i}\hat{H}_0^R(t_{ij}^R)\exp\{\hat{\boldsymbol{\beta}}^T\boldsymbol{X}_{ij}^R\}}\right] \quad \forall l = 1, ..., L \tag{3.81}$$

Then, keeping fixed $\hat{\boldsymbol{P}}^{(1)}$ (i.e. the vector of abscissa of the support points) to the updated values, noticing that for each fixed i the sum of $\mathbb{Z}_{il}$ is unitary, we can rewrite l_R as

$$\begin{aligned} l_R(\boldsymbol{\beta}, H_0^R(t)|\hat{P}^{(1)}) = \sum_{i=1}^{N}\sum_{j=1}^{k_i}\Big[\delta_{ij}^R\log(h_0^R(t_{ij}^R)) + \boldsymbol{\beta}^T\boldsymbol{X}_{ij}^R + (\sum_{l=1}^{L}\mathbb{Z}_{il}\hat{P}_l^{(1)}) + \\ - H_0^R(t_{ij}^R)\exp\{\boldsymbol{\beta}^T\boldsymbol{X}_{ij}^R + \log(\sum_{l=1}^{L}\mathbb{Z}_{il}\exp(\hat{P}_l^{(1)}))\}\Big] \end{aligned} \tag{3.82}$$

which is the usual full loglikelihood of a Cox model with known offset represented by

$$\text{offset} = \log(\sum_{l=1}^{\hat{k}} \mathbb{Z}_{il} \exp(\hat{P}_l^{(1)}))\}. \tag{3.83}$$

For this reason, we propose an extension of the Breslow estimator for the cumulative baseline hazard for recurrent events

$$\hat{H}_0^R(t) = \sum_{ab:t_{ab}^R<t} \frac{m_{ab}}{\sum_{cd\in\mathcal{R}(t_{ab})} \exp\{\hat{\boldsymbol{\beta}}^T \boldsymbol{X}_{cd}^R + \log(\sum_{l=1}^L \mathbb{Z}_{il} \exp(\hat{P}_l^{(1)}))\}} \tag{3.84}$$

where m_{ab} is the number of events happened at t_{ab} (correction factor for tied times) and $\mathcal{R}(t_{ab})$ is the risk set at time t_{ab}. In order to retrieve an update rule for β we can then substitute $\hat{H}_0^R(t)$ and its derivative in Equation 3.82 to obtain a profile loglikelihood for $\boldsymbol{\beta}$, which not surprisingly coincides with the partial loglikelihood of a Cox model with the same offset specified in Equation 3.83

$$l_R^{prof}(\boldsymbol{\beta}) = \sum_{i=1}^N \sum_{j=1}^{k_i} \delta_{ij}^R \Big[\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R - m_{ij}^R \log \sum_{ab\in\mathcal{R}(t_{ij}^R)} \exp\{\boldsymbol{\beta}^T \boldsymbol{X}_{ij}^R + \log \sum_{l=1}^L \mathbb{Z}_{il} \exp(P_l^{(1)}\}\Big] \tag{3.85}$$

which can be maximized through iterative methods (usually Newton-Raphson) in order to retrieve $\hat{\boldsymbol{\beta}}^{(up)}$. Notice moreover that, since this is the usual formulation adopted in Cox model estimation, the retrieval of $\hat{\boldsymbol{\beta}}^{(up)}$ can be done through standard R survival package `survival`.

The maximization of l_D can be performed following the same procedure designed for l_R. The update rules are given by

$$\hat{P}_l^{(2)(up)} = \log\left[\frac{\sum_{i=1}^N \mathbb{Z}_{il}\delta_i^D}{\sum_{i=1}^N \mathbb{Z}_{il}\hat{H}_0^D(t_i^D)\exp\{\hat{\boldsymbol{\gamma}}^T \boldsymbol{X}_i^D\}}\right] \quad \forall l = 1, ..., L \tag{3.86}$$

$$\hat{H}_0^{D(up)}(t) = \sum_{a:t_a^D<t} \frac{m_a}{\sum_{c\in\mathcal{R}(t_a)} \exp\{\hat{\boldsymbol{\gamma}}^T \boldsymbol{X}_c^D + \log(\sum_{l=1}^L \mathbb{Z}_{il} \exp(\hat{P}_l^{(2)}))\}} \tag{3.87}$$

$$l_D^{prof}(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \delta_i^D \Big[\boldsymbol{\gamma}^T \boldsymbol{X}_i^D - m_i^D \log \sum_{a \in \mathcal{R}(t_i^D)} \exp\{\boldsymbol{\gamma}^T \boldsymbol{X}_i^D + \log \sum_{l=1}^{L} \mathbb{Z}_{il} \exp(P_l^{(2)})\} \Big]. \quad (3.88)$$

Due to the fact that the maximization step is composed by separate phases with respect to parameters of the recurrent events and of the terminal event processes, it may seems that the proposed model estimation algorithm is not joint; however, it is important to stress the fact that, in the expectation step, the computation of $\mathbb{Z}_{il}$ takes into account all the current values of the parameters in $\boldsymbol{\Omega}$. Since all the above update rules depend on such quantities the estimation procedure is actually joint.

3.4.5. A priori unknown number of support points

Up to now, we have detailed the steps of an EM algorithm which consider the total number of support points of the discrete distribution as known. However, as specified above, we want to consider a generic discrete distribution on $\mathbb{R}^2$. In order to generalize, we propose a wrapper method that, given an initial grid, performs a support reduction, which is inspired by the one presented in [31].

Grid Initialization

The first step consists in the definition of a grid of points in $\mathbb{R}^2$, which ideally covers the region in which the (unknown) true support of the discrete distribution is believed to lie. This is done accordingly to previously available knowledge, which may come from general exploratory analysis, medical knowledge of the phenomenon or previously fitted models. An alternative may be represented, for example, by the sampling from a bivariate Normal distribution characterizing the random effects in a model like the one proposed by Ripatti and Palmgren (Section 3.2.1). In this case, the weights can be initialized according to the corresponding Normal density and then normalized to be unitary. The main drawback is represented by the fact that, due to its nature, the algorithm is very likely to be sensitive to the grid initialization. In such case, it is better to provide an initialization strategy as general and non-informative as possible. An example can be represented by a uniform distribution of points over a rectangle in $\mathbb{R}^2$, whose boundaries are defined, as mentioned, to cover the supposed area of the true support. Again, a possible way to define them is to use some quantitative characterization coming from previous models (e.g. consider the standard deviation estimates of a Gaussian characterization like in Ripatti and Palmgren). In both cases, the number of points composing the initial grid

must ensure a proper exploration of the considered region.

Support Reduction

At the start of each iteration of the EM algorithm the current support is reduced, repeatedly merging the minimum distance couple of points, until no couple with distance less than a user-defined threshold exists. Further insights on the threshold are given in Section 3.4.7. The rules followed in the merging are

$$\boldsymbol{P}^{(new)} = \left[\frac{P_1^{(1)(old)} + P_2^{(1)(old)}}{2}, \frac{P_1^{(2)(old)} + P_2^{(2)(old)}}{2} \right] \tag{3.89}$$

$$w^{(new)} = w_1^{(old)} + w_2^{(old)}. \tag{3.90}$$

Once the grid and the weights have been updated, the expectation step detailed in the previous Section can be performed. Before passing to the maximization step, the current latent partition of subjects (constituted by $\text{label}_i \ \forall i = 1, ..., N$) is extracted, assigning patients to points with maximum $\mathbb{Z}_{il}$

$$\text{label}_i = \text{argmax}_l \mathbb{Z}_{il} \quad \forall i = 1, ..., N \tag{3.91}$$

and deleting eventual points in the support to which no patient is longer assigned. Then, the maximization step can be carried out as described above. The algorithm stops when a given number of iterations is reached or the number of masses in the discrete distribution is stable (no reduction happens in the current iteration) and the difference between old and updated weights, computed in maximum norm, is less than a threshold (set to 1e-03). In the panel Algorithm 3.3 the overall procedure is summarized.

3.4.6. Standard Errors computation

We decided to adopt a simple approach to provide estimates for the two processes coefficients standard errors. We rely on the R built-in functions, used during the process to compute the coefficients estimates, to obtain cheap estimates of the corresponding standard deviations. The reason of this choice consists in following the usual approach of Cox models, consisting in the inversion of the Hessian matrix evaluated at the found coefficients estimates, applied separately to the last iteration, fixed intercept Cox models,

Algorithm 3.3 Support Reduction Estimation Procedure

Set K
Grid Initialization (Gaussian, Uniform, ...)
Choose distance (Euclidean, Manhattan,...)
Set $MinDist$
Set $epsw = 1$
Set $converged = False$, $MaxIt = 200$,$eps = 1e-03$,$it = 0$
while !$converged$ and $it < MaxIt$ **do**
    Support Reduction: merge points nearer than MinDist
    Expectation step: compute $\mathbb{Z}_{il}$
    Extract the latent partition, delete unassigned masses
    Update $\hat{\boldsymbol{w}}$
    Update $\hat{\boldsymbol{P}}$
    Update $\hat{\boldsymbol{\beta}}$,$\hat{\boldsymbol{\gamma}}$,$\hat{H}_0^R$,$\hat{H}_0^D$
    if Reduction in the current iteration **then**
        $converged = FALSE$
    else
        compute $epsw$
        if $epsw < eps$ **then**
            $converged = TRUE$
        end if
    end if
end while

which are a byproduct of our estimation routine. This allow us to estimate the statistical significance of the parameters computing the corresponding Wald statistic [11]. The same reasoning can be applied to obtain standard errors of the two cumulative baseline hazard functions from the corresponding R functions used in the estimation routine.

This approach, which is now implemented in the estimation procedure, is cheap and allows us to obtain plausible estimates without much effort, however it is very restrictive. Actually, besides not providing standard errors for the parameters involved in the random effects discrete distribution, it does not consider the mutual influence of parameters in the computation. We are basically considering the involved elements (i.e. the two fixed intercept Cox models obtained during the last iteration) as separate entities. For this reasons, the design of a comprehensive method to obtain proper estimates of standard errors

represents the first further development of this model. A way to obtain a comprehensive variance-covariance matrix is represented by inverting the observable Information matrix

$$I(\mathbf{\Omega}) = -\frac{\partial^2 l_{obs}(\mathbf{\Omega})}{\partial \mathbf{\Omega}^2} \tag{3.92}$$

where $\mathbf{\Omega}$ comprehends the two processes coefficients, the two processes baseline hazards, the support and the weights of the random effect discrete distribution, and $l_{obs}(\mathbf{\Omega})$ is the observable loglikelihood of our joint model, which can be obtained from Equation 3.71 integrating out the random effects. As usual, the obtained matrix should be then evaluated in correspondence of the estimated values of the parameters.

3.4.7. Model design choices

Besides the grid initialization, which was already discussed, the algorithm is sensitive to two main design choices. The first is represented by the kind of distance used in the support reduction procedure. The most common one is the Euclidean distance, which provide a good interpretability of the criterion used for merging points. Another possible choice is represented by the Manhattan distance, which may result useful in case of particular patterns in the hidden discrete distribution of random effects. The most important choice which must be done by the user, which is also correlated to the chosen kind of distance for the merging of points, is the threshold *MinDist* discriminating which points will be collapsed. It represents a critical part of the estimation procedure, since it leads to different results in term of the final discrete distribution discovered. As a general rule, if the available medical knowledge of the matter permits this kind of reasoning, it must be set to the minimal value for which the belonging to a mass or another of a patient describes a significant difference in the considered application. From the practical point of view, it is recommended to perform a sensitivity analysis, looking at the evolution of some fitting criterion (e.g. loglikelihood, AIC, etc) with respect to different values of the threshold, in order to identify the most promising candidates. Of course, the choice itself of the fitting criterion will influence the sensitivity analysis. We choose to resort to the most intuitive one (i.e. the AIC), as, in general, the designing of a robust procedure to identify the best candidate for the distance threshold is a complicate matter which is beyond the scope of this thesis.

We now introduced statistical methodologies we will use for our analyses. In the next Chapter we will present applications and results of our work.

4 | Results

In this chapter we present the results obtained through the application of the methodologies proposed in Chapter 3 on the ACE Inhibitors Dataset, presented in Chapter 2. In Section 4.1, classical survival tools are used to grasp insights from available data about the mortality process in heart failure patients. Section 4.2 moves the focus to hospitalizations, modelling separately the recurrent events and terminal events processes through frailty models. Section 4.3 and Section 4.4 show the results of a comparative study, where our joint modelling approach is fitted to data, together with the ones of Rondeau et al. [51] and Ng et al. [37].

4.1. Classical Survival Analysis

As first step, we perform some classical survival techniques (Section 4.1.1) and a Cox regression (Section 4.1.2), in order to gain insights about the effect of covariates on the overall survival time. The purpose of such introductory analysis is to highlight the strenghts and weaknesses of the classical Cox analysis in presence of recurrent events, in our dataset represented by hospitalizations. As mentioned, we presents the results concerning the ACE Inhibitors dataset, but the same methodological framework can be applied to the others drug datasets which compose the Heart Failure dataset. All statistical analyses are performed using `survival` package [60] in the R software environment [46].

4.1.1. Descriptive Analysis

In this part we would like to introduce a statistical description of the HF database cohort related to the 3,232 patients who underwent ACE-Inhibitors therapy, selected as described in Section 2.2 and Section 2.4.

The classical survival analysis is carried out considering death as time-to-event outcome for all the patients. In particular, **death** and **timeOUT** covariates respectively represent the binary variable that indicates whether a patient is dead or not at the end of the

study and the overall survival time (expressed in days) of the patient. Since during the cohort selection procedure we only consider patient alive at the end of the first year of the observation period, in order to compute adherence-related variables (see Sections 1.3 and 2.3), overall survival time is computed from the end of the observation period until death or censoring. Initially, the set of potential covariates is represented by

- **Sex**: gender of the patient. Factor with levels "M" for males and "F" for females;
- **Adherent1Y**: binary flag indicating whether a patient is adherent(1) or not(0) to ACE at the end of the 1-year observation period, computed as specified in Section 1.3;
- **AdhLev1Y**: categorical variable. The adherence class of a patient during the 1-year observation period (1: [0,0.25); 2: [0.25,0.5); 3: [0.5,0.75); 4: [0.75:1)), computed as specified in Section 1.3;
- **AgeMin**: age (in years) at first hospitalization;
- **Comorbidity1st**: number of comorbidities registered at first hospitalization;
- **MaxComorbidity**: maximum number of comorbidities registered within the 1-year observation period.
- **totHosp1Y**: number of hospitalizations within the first 1-year observation period.

Some of the considered covariates are highly collinear, so we will need to perform a selection before fitting our models.

Table 4.1 shows the summaries of the aforementioned characteristics. Median follow-up time is 1,198 (IQR=[695;1,729]) days (about 3 years) and 2,428 (75.12%) patients are alive at the end of the follow-up. Among 3,232 patients, 1,840 (56.9%) are males and 1,502 (45.48%) are adherent to ACE therapy. At index hospitalization, median age and number of comorbidities are 74 (IQR = [66;80]) years and 2 (IQR=[1,3]), respectively. At the end of the 1-year observation period, the median number of hospitalizations is 2 (IQR = [1;3] with a maximum of 14 re-hospitalization events), while the median of maximum number of comorbidities within the first year is 2 (IQR = [1;3]).

To describe the general survival behaviour of the population and the effect of categorical factors on time-to-death, we use Kaplan-Meier estimators (see Sections 3.1.3) and log-rank test approaches [27]. Figure 4.1 shows the Kaplan-Meier estimate for the overall survival curve. From the figure we observe that median survival time is not reached during the

Variable		**Value**
death	0 (%)	2,428 (75.12%)
	1 (%)	804 (24.88%)
timeOUT [days]	median (Q1;Q3)	1,238 (695;1,729)
	mean(s.d.)	1,198 (608.059)
	min – max	1 – 2,190
Sex	F (%)	1,392 (43.06%)
	M (%)	1,840 (56.94%)
Adherent1Y	0 (%)	1,730 (53.52%)
	1 (%)	1,502 (46.48%)
AdhLev1Y	[0.00,0.25) (%)	691 (21.38%)
	[0.25,0.50) (%)	386 (11.94%)
	[0.50,0.75) (%)	511 (15.81%)
	[0.75,1.00] (%)	1,644 (50.87%)
AgeMin [years]	median (Q1;Q3)	74 (66;80)
	mean(s.d.)	72.22 (11.307)
	min – max	18 – 98
Comorbidity1st	median (Q1;Q3)	2 (1;3)
	mean(s.d.)	3 (1.079)
	min - max	0 – 7
MaxComorbidity	median (Q1;Q3)	2 (2;3)
	mean(s.d.)	2.059 (1.302)
	min - max	0 – 10
TotHosp1Y	median (Q1;Q3)	2 (1;3)
	mean(s.d.)	2.333 (1.596)
	min - max	1 – 14

Table 4.1: Summary statistics related to the 3,232 HF patients who underwent ACE-Inhibitors therapy.

first 6 years of follow-up, since the curves is always above 0.50. No difference is observed in terms of survival curves stratified by gender: p-value of log-rank test is 0.11 (see Figure 4.2). Figures 4.3 shows that being adherent to ACE increases the survival probability (p-value of log rank test is 2e-07). Statistically significant difference for overall survival curves is also observed for adherence classes (see Figure 4.4, p-value of log-rank test is 9e-06), where patients in level 4 with PDC $\in$ [0.75,1] (purple curve) show the higher survival. However, specification of adherence through levels is assessed to be significant only due to

Variable	**Strata**	**Value**	**N**	**pvalue**
AgeMin	*junior*	AgeMin ≤66	826	<2e-16
	senior	66<AgeMin≤80	1,656	
	old senior	AgeMin > 80	750	
Comorbidity1st	*Very Few*	0≤Comorbidity1st≤1	1,065	7e-15
	Few	2≤ Comorbidity1st≤3	1,877	
	Many	Comorbidity1st>3	290	
MaxComorbidity	*Very Few*	0≤MaxComorbidity≤1	667	<2e-16
	Few	2≤ MaxComorbidity≤3	1,887	
	Many	MaxComorbidity>3	678	
TotHosp1Y	*Low*	0≤TotHosp1Y≤3	2,638	1e-05
	High	TotHosp1Y>3	594	

Table 4.2: P-values of stratified log-rank tests related to quantitative characteristics.

the 4th level versus the rest, as confirmed by pairwise log-rank test (p-value of 3e-06). For a preliminary investigation of the effect of numerical variables on time-to-death, we use a discretization strategy in order to implement the above approaches also to the remaining quantitative characteristics. From the p-values related to the stratified log-rank tests in Table 4.2 we observe that:

- statistically significant difference (pvalue <2e-16) is observed in terms of age at first hospitalization, that is investigated through binning using interquartile range and the two tails (classes *junior*, *senior* and *old senior*, see Table 4.2).
- Number of comorbidities at index hospitalization and maximum number of comorbidities registered in the observation period are investigated using a classification in *Very Few* (0 or 1), *Few* (2 or 3) and *Many* (outlying subjects beyond Q3=3). Both covariates are statistically significant, reporting a p-value of, respectively, 7e-15 and less than 2e-16.
- Number of hospitalization registered within the 1-year observation period shows significant statistical difference (pvalue of 1e-05) between subject with *Low* and *High* number of hospitalizations.

In the following section, we applied a Cox PH model to study the simultaneous effects of multiple characteristics on overall survival, proposing a continuous modelling for the numerical variables.

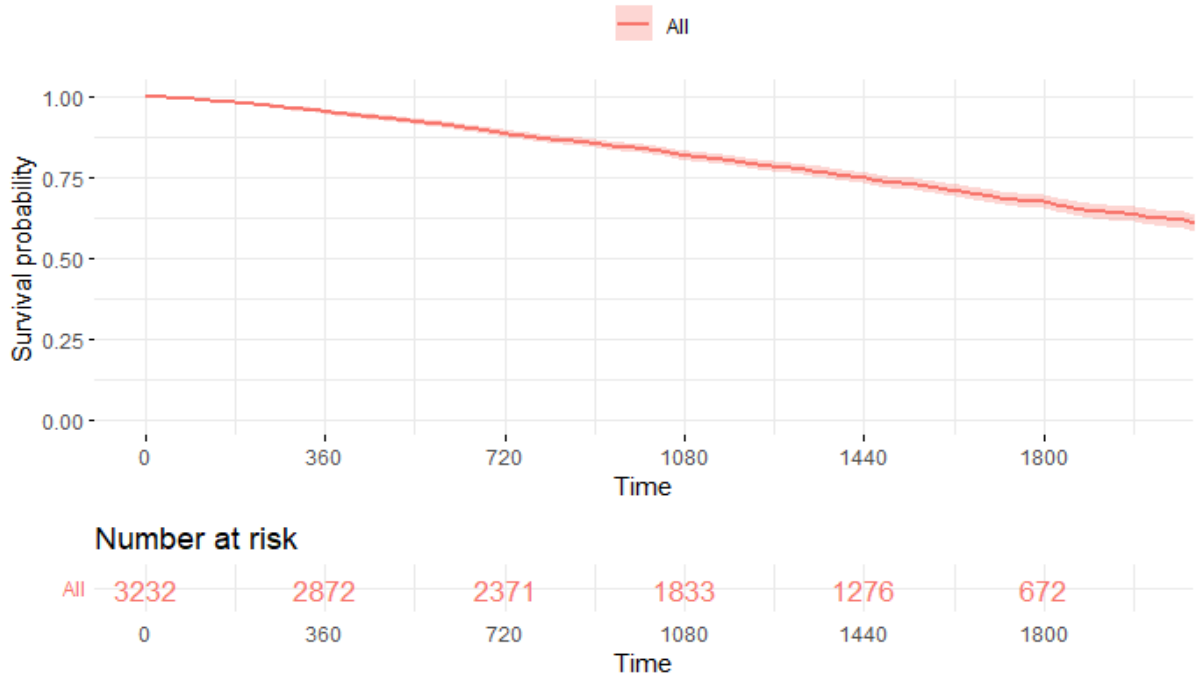

Figure 4.1: Kaplan-Meier estimate for overall survival

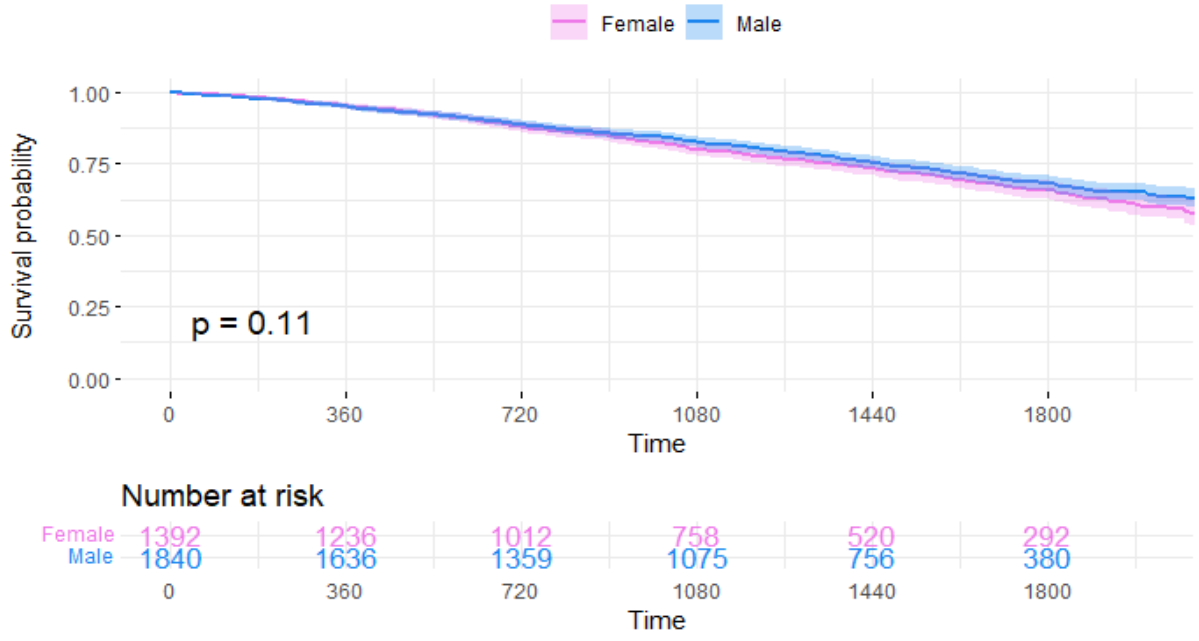

Figure 4.2: Kaplan-Meier estimates for overall survival stratified by gender along with the relative risk table (blu: male; pink: female). Value p refers to the p-value of log-rank test to compare the survival distributions stratified by gender.

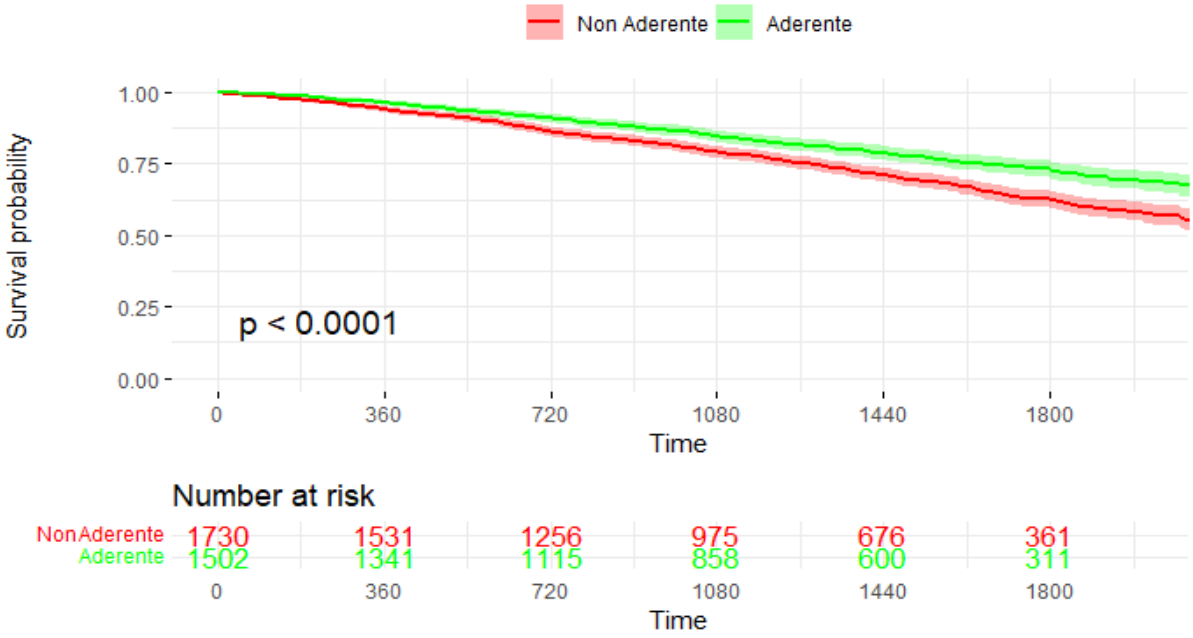

Figure 4.3: Kaplan-Meier estimates for overall survival stratified by binary adherence in the first year of follow up, along with the relative risk table (red: non adherent 0; green: adherent 1). Value p refers to the p-value of log-rank test to compare the survival distributions stratified by adherence.

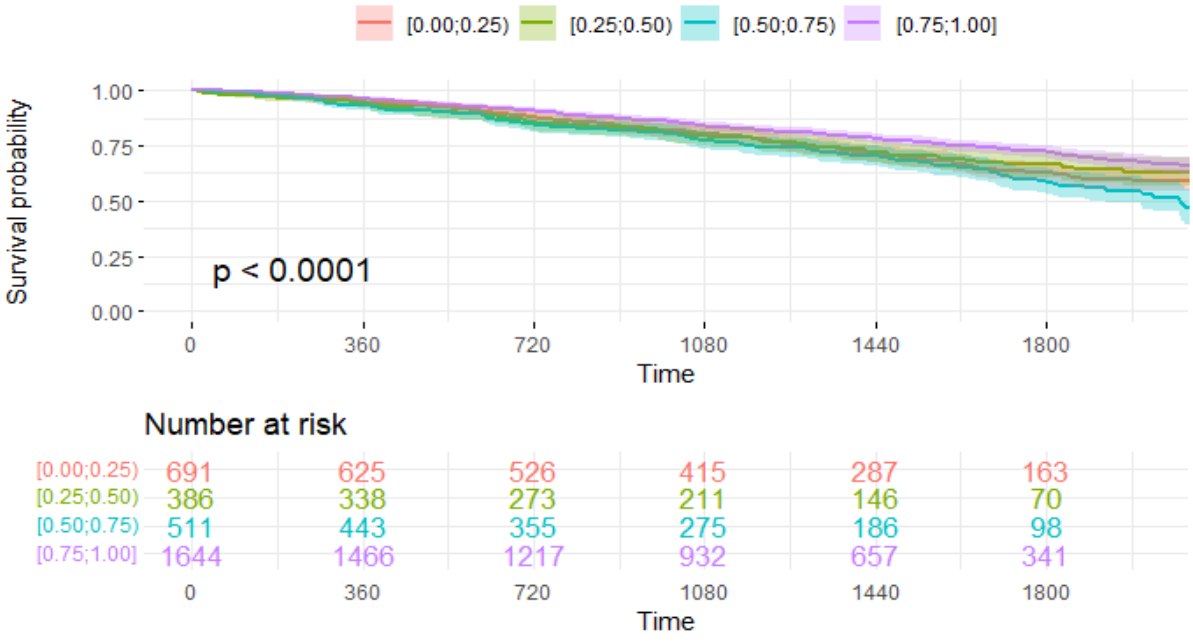

Figure 4.4: Kaplan-Meier estimates for overall survival stratified by adherence levelsin the first year of follow up, along with the relative risk table (red: [0.00,0.25); green: [0.25,0.50); blue: [0.50,0.75); purple: [0.75,1.00]. Value p refers to the p-value of log-rank test to compare the survival distributions stratified by adherence levels.

4.1.2. Cox Proportional Hazard model

In the previous section we have assessed the presence of significant effects on the overall survival due to the proposed characteristics. The Cox proportional hazard model (see Section 3.1.4) gives us the chance of further refining our analysis, encoding the numerical variables in a continuous manner. However, some of the proposed covariates (binary adherence and adherence classes; number of comorbidities at index event and maximum number of comorbidities within the 1-year observation period) encodes similar information and are likely to be correlated. Thus, in order to avoid introduction of bias in the estimation process we must perform a selection procedure.

As far as adherence is concerned, we decided to choose binary adherence. This choice is done having observed (see previous Section) that the pairwise comparison of the 4th level against others is as significant as the full specification in four levels. In addition, the binary encoding is straightforward and thus more easily understandable.
In order to choose the comorbidity-related variable, instead, two different Cox models are trained. The former involves **Sex**, binary **Adherent1Y**, **AgeMin**, **TotHosp1Y** and the number of comorbidities at index hospitalization (**Comorbidity1st**); in the latter, instead, the maximum number of comorbidities registered in the observation period (**MaxComorbidity**) is chosen as last regressor. The two models are then compared through Akaike Information Criterion (AIC) and Concordance Index (CI): the performances are very similar (AIC=11735.03, CI=0.697 against AIC=11754.19, CI=0.693), suggesting that both models are valid and can be used for inference. In the following are presented the results related to the second one.

To sum up, the final Cox model chosen for inference is represented by

$$h_i(t|\boldsymbol{X}_i) = h_0(t)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}} \ \forall i = 1, ..., N \tag{4.1}$$

where T denotes the transposition operator, $\boldsymbol{\beta}$ the vector of estimated coefficients, i is the index of the patient and $\boldsymbol{X}_i$ its covariate vector

$$\boldsymbol{X}_i = \{Sex_i, Adherent1Y_i, AgeMin_i, MaxComorbidity_i, TotHosp1Y_i\}$$

Variables	Estimate	StdDev	HR	CI95	pvalue
Sex [M]	0.147	0.073	1.159	[1.003,1.338]	0.044
Adherent1Y [1]	-0.258	0.073	0.772	[0.669,0.892]	0.0004
EtaMin	0.065	0.004	1.068	[1.058,1.077]	<2e-16
MaxComorbidity	0.195	0.028	1.216	[1.151,1.283]	1.6e-12
TotHosp1Y	0.064	0.023	1.066	[1.020,1.114]	0.004

Table 4.3: Cox model summary. For categorical variables, the considered stratum is indicated between brackets.

Table 4.3 reports the summary of the final Cox model. In addition, comparative survival probability plots (see Figures 4.5-4.9) are provided to visualize the effect of each covariate on the survival probabilities. The approach followed is to focus on a covariate at time and keep fixed the other ones.

The covariate **Sex** results statistically significant at 5%. As expected, it expresses the trend of Male patients to have a lower life expectation (HR=1.159), as shown in Figure 4.5 that represents the survival curves stratified by gender. Categories are selected according to the following criteria: we consider male(female) patients who are adherent, are firstly hospitalized at the age of 74 and shows, during their 1-year observation period, both a maximum number of comorbidities and a number of hospitalizations of 2.

The covariate **Adherent1Y** assumes great significance (p-value of 0.0004), assessing a decreased risk in mortality of about 23% (HR=0.772). Survival probability plot stratified by adherence in Figure 4.6 also confirms this results. Categories are selected similarly as before: we consider adherent versus non adherent female patients who are firstly hospitalized at the age of 74 and shows, during their 1-year observation period, both a maximum number of comorbidities and a number of hospitalizations of 2.

The covariate **AgeMin** is statistically significant at any level (p-value less than 2e-16) and a 1-year increase in age at index hospitalization reflects a risk increase of 6.76%. Figure 4.7 reports nine different survival curves stratified by age at index hospitalization: considered values of **AgeMin** are chosen to span uniformly the variable range (18-28-38-48-58-68-78-88-98 years) of female adherent patients with 2 comorbidities and 2 hospitalizations within the 1-year observation period.

The covariate **MaxComorbidity** is significant at any level (p-value of 1.6e-12). In clinical literature, comorbidity is a condition known to deeply influence the severity of a patient clinical picture; also in this case, it shows a high effect in increasing the mortality risk (21% increase per comorbidity registered within the 1-year observation period). Figure 4.8 reports a comparative survival probability plot composed of nine curves: it consider female adherent patients, hospitalized at 74 years of age and with 2 hospitalizations during the observation period, who showed, respectively, 0-1-2-3-5-6-7-8-10 comorbidities during the observation period.

The covariate **TotHosp1Y** is statistically significant at less than 1% (p-value of 0.0004) and shows a that each hospitalization in the observation period increases the risk of 6.6%. This is confirmed by the related comparative plot, which is reported in Figure 4.9. It considers (female, adherent) patients with 1-2-4-5-7-9-10-12-14 hospitalization during the observation period, who entered the study at 74 years of age and showed 2 comorbidities during the observation period. However, 75% of patients shows up to three hospitalizations and, as assessed by the relative log rank test (see Section 4.1.1), the effect of such variable is intended to be relevant when outlying subjects (with a very high number of hospitalizations) are involved. According to the previous considerations, probably the proportional hazard framework is a bit restrictive for modeling the information enclosed in hospitalizations.

Finally, goodness of fit and proportional hazard assumption are assessed. Figure 4.10 shows the deviance residuals for the final Cox model in Equation 4.1, which present a good behaviour around 0 confirming the goodness of fit. Figure 4.11 displays the Schoenfeld residuals for the different covariates (each panel is related to a different covariate). From the figure we can conclude that the proportional hazard hypothesis are satisfied for all the variables.

In conclusion, the trained Cox model is an effective tool to investigate the effect of the proposed variables (above all incidence of adherence to the considered drug) in the considered context of heart failure patients. However, in the present Cox framework, the hospitalizations can only be considered as the number of re-admissions within the first one year, losing information about the dynamics of the repeated events. Therefore, a more appropriate modeling of re-hospitalization events able to incorporate additional information regarding the dynamics of these processes is needed. Moving from the classical survival analysis, in the following sections we introduce alternative approaches: hospital-

izations will be considered as dynamic repeated events and modeled, alongside deaths, through different and increasingly complex Cox-based structures.

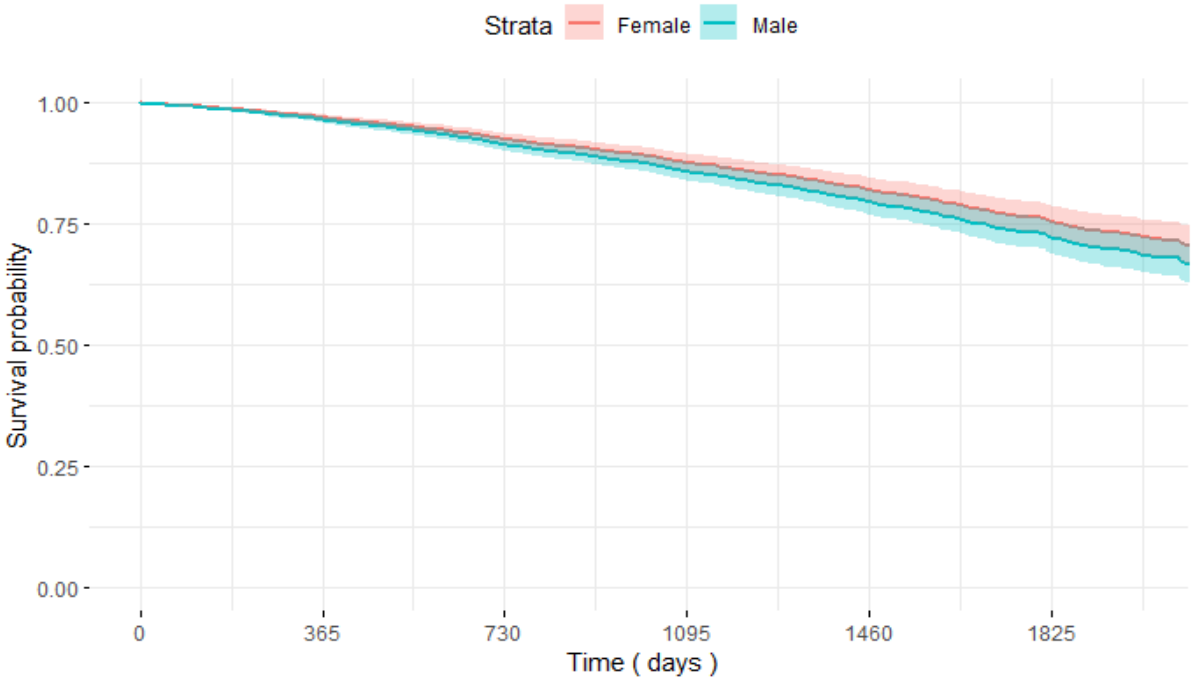

Figure 4.5: Survival probability plot stratified by sex (red: female; blue: male) related to the final Cox model in Equation 4.1. Other characteristics are selected according to the following criteria: we consider male(female) patients who are adherent, are firstly hospitalized at the age of 74 and shows, during their 1-year observation period, both a maximum number of comorbidities and a number of hospitalizations of 2.

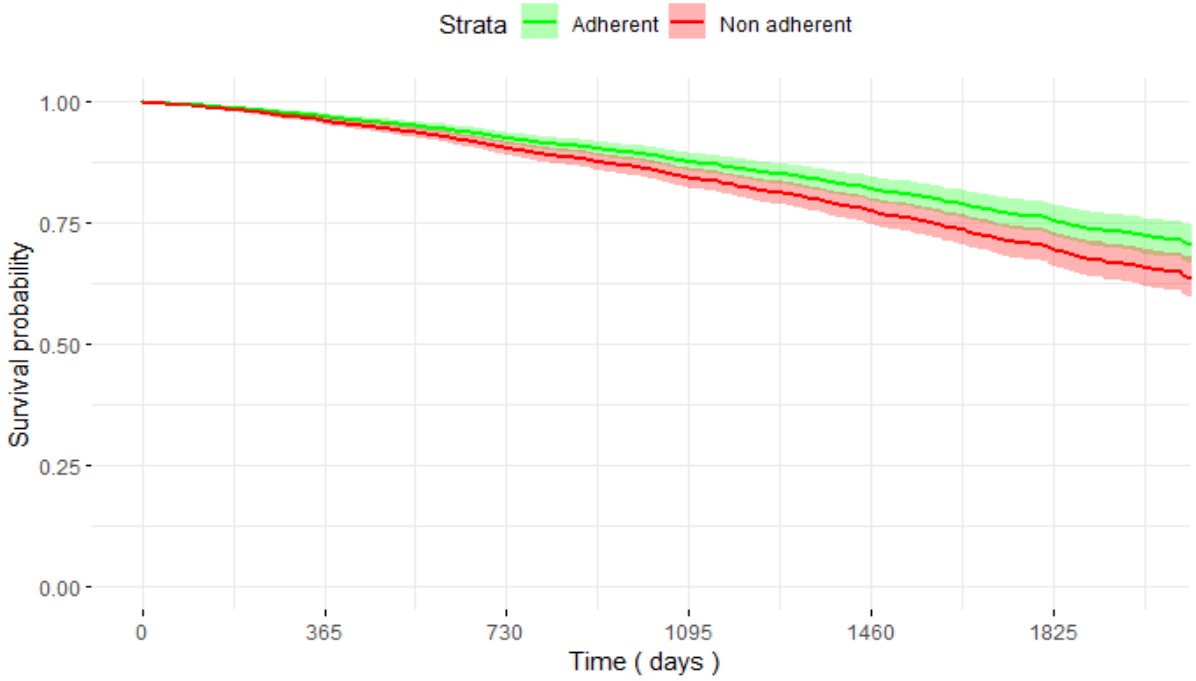

Figure 4.6: Survival probability plot stratified by adherence (red: non-adherent; green: adherent) relative to the final Cox model in Equation 4.1. Other characteristics: female, firstly hospitalized at the age of 74, maximum number of comorbidities of 2, number of hospitalizations of 2.

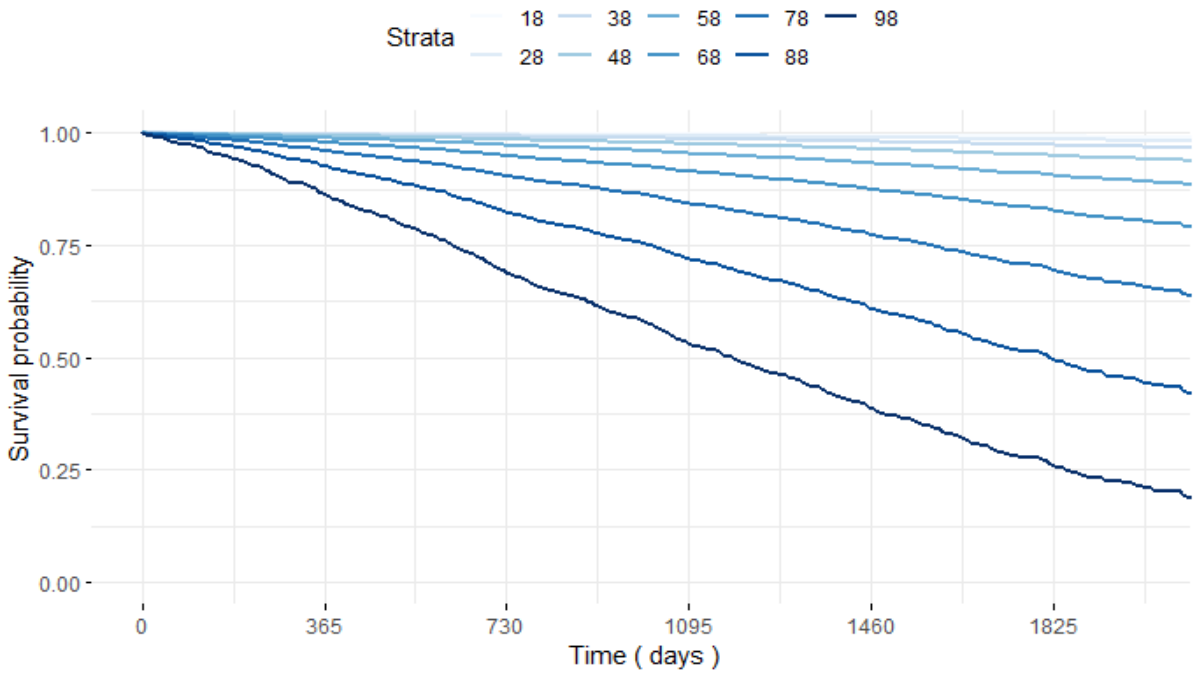

Figure 4.7: Survival probability plot stratified by age at first hospitalization relative to the final Cox model in Equation 4.1. Nine evenly spaced values covering the whole range are considered (18-28-38-48-58-68-78-88-98). Other characteristics: female, adherent, maximum number of comorbidities of 2, number of hospitalizations of 2.

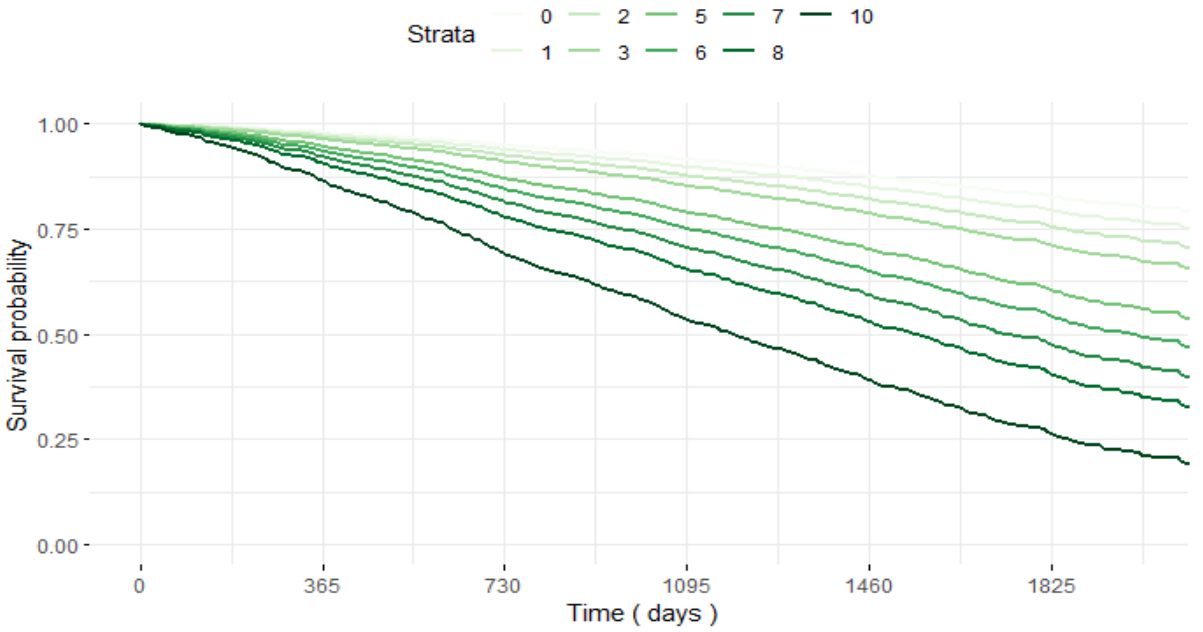

Figure 4.8: Survival probability plot stratified by maximum number of registered comorbidities within the observation period relative to the final Cox model in Equation 4.1. Nine increasing values covering the whole range are considered (0-1-2-3-5-6-7-8-10). Other characteristics: female, adherent, firstly hospitalized at the age of 74, number of hospitalizations of 2.

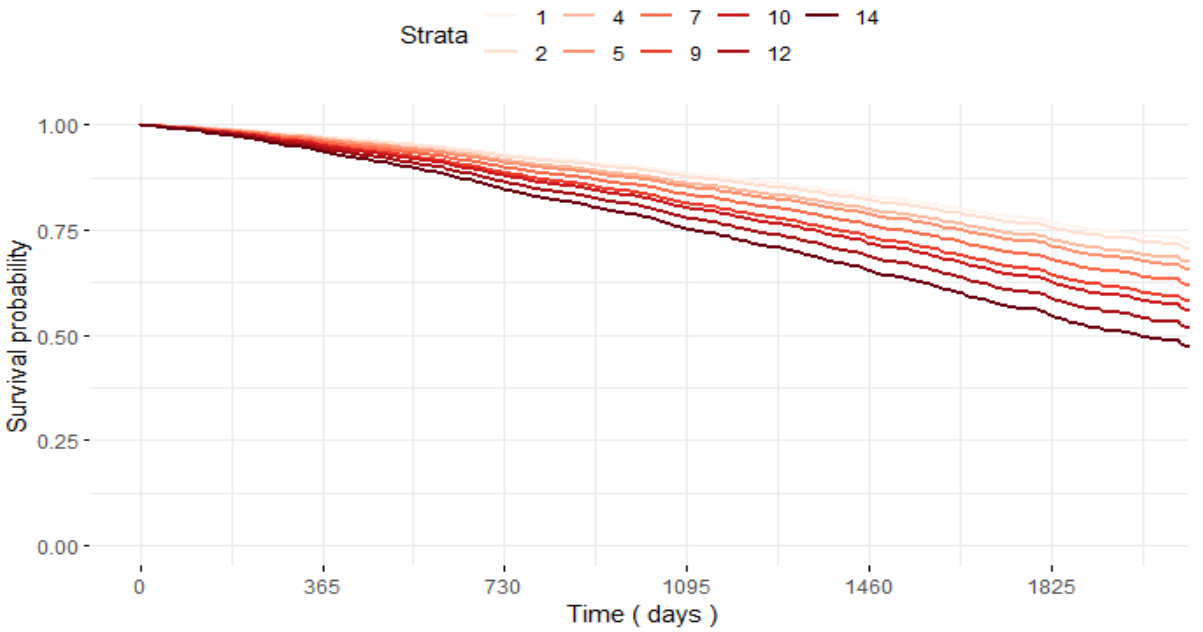

Figure 4.9: Survival probability plot stratified by number of hospitalizations within the observation period relative to the final Cox model in Equation 4.1. Nine increasing values covering the whole range are considered (1-2-4-5-7-9-10-12-14). Other characteristics: female, adherent, firstly hospitalized at the age of 74, maximum number of registered comorbidities of 2.

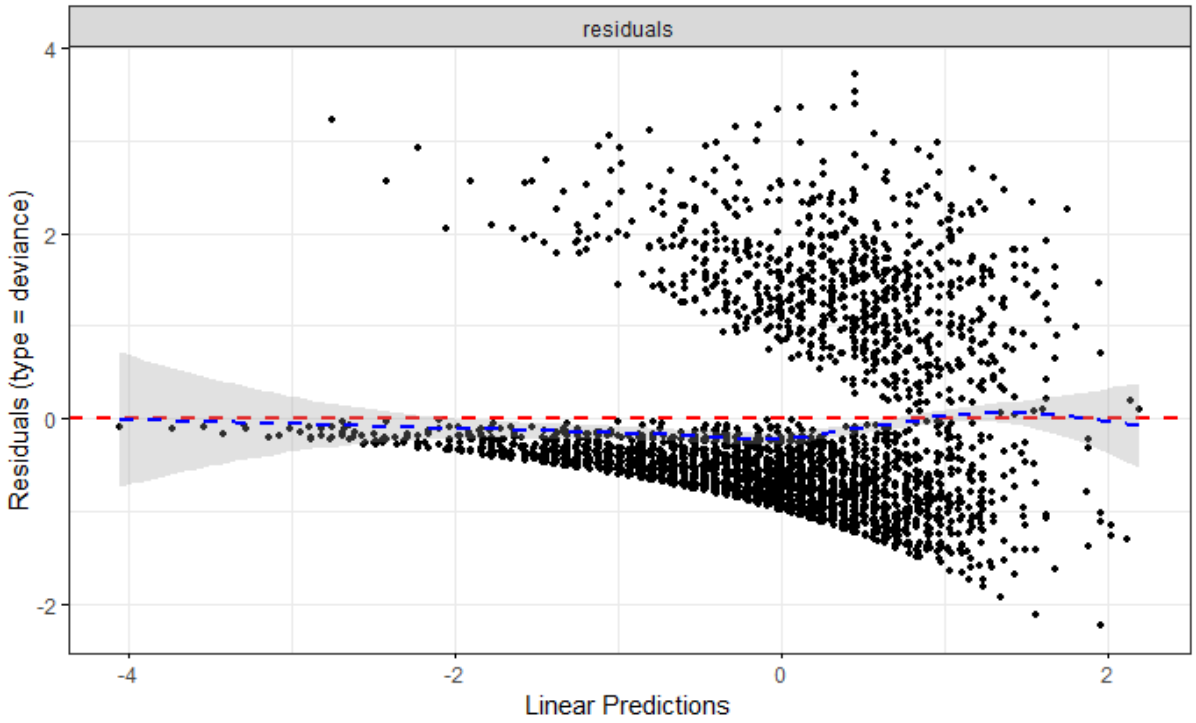

Figure 4.10: Deviance residuals plot for the final Cox model in Equation 4.1

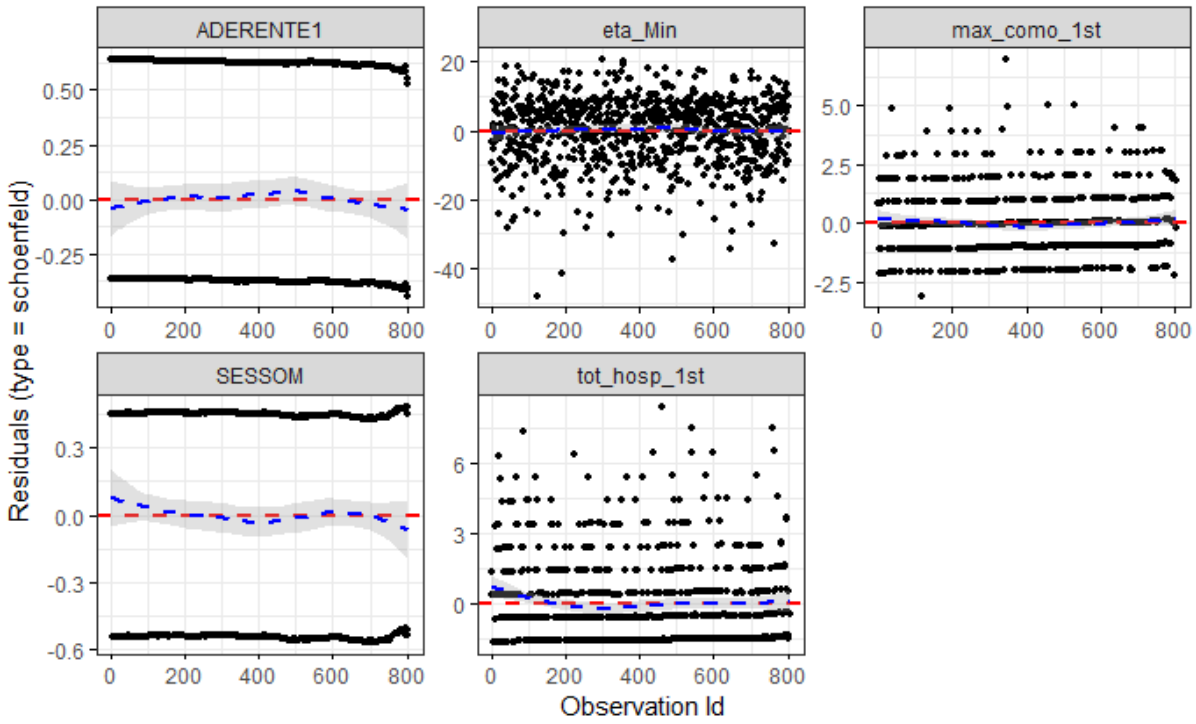

Figure 4.11: Schoenfeld residuals plots for the final Cox model in Equation 4.1. Each panel is related to a different covariate.

4.2. Recurrent Events Framework

In this Section we move to a recurrent events framework (Section 3.2), presenting the results obtained modeling the hospitalization and death processes independently. In Subsection 4.2.1 is detailed the encoding of the Heart Failure Dataset in the context of recurrent and terminal events, as well as a brief description of the variables involved. In Subsection 4.2.2 we fit two independent frailty models, one for recurrent events and one for terminal events, following the approach of Ripatti and Palmgren (Section 3.2.1). This is done in order to provide a baseline when fitting joint models in Section 4.3. Finally, we perform a comparison with two classical (i.e. without frailty) Cox models, fitted on the same data, to highlight the effect of frailties when dealing with correlated data. In the following sections, results concerning the Heart Failure dataset subset related to ACE Inhibitors are presented; similar analyses can be performed on the other types of drugs mentioned in Chapter 1. Frailty models are fitted using the `coxme` package [59] in the R software environment [46].

4.2.1. Descriptive Analysis

The HF database cohort considered is the same of Section 4.1.1, composed by patients who underwent ACE-Inhibitors therapy, selected as described in Section 2.2 and 2.4. We recall that, within the recurrent and terminal events framework, adherence to the treatment is quantified by the time dependent binary variable defined at the end of Section 2.3. In this case, the selection of patients alive at least until the end of the first year of follow up is no longer needed, however we decide to consider the same cohort of patients of the classical survival analysis, in order to be able to make a proper comparison.

Our aim is to fit two different frailty models, one describing the recurrent events process of hospitalizations subsequent to first discharge due to heart failure and one describing the major terminal event process (i.e. death). For this reason, our data consists of the clinical histories of the 3,232 patients in the considered cohort. Figure 4.12 provide a visualization of the clinical histories of four patients. Blue dots represents hospitalizations, red squares deaths and red triangles terminal events different from death (loss to follow up, withdrawal from the study, end of study). The time unit considered (horizontal axis) is years. Overall, the ACE Inhibitors dataset comprehends 12,746 hospitalization recordings plus 3,232 terminal events, which 804 (24.9%) are deaths. The smallest number of re-hospitalizations experienced by a subject is 1, while the highest is 42; however, 90.4% of the patients experience up to 10 hospitalizations before the terminal event.

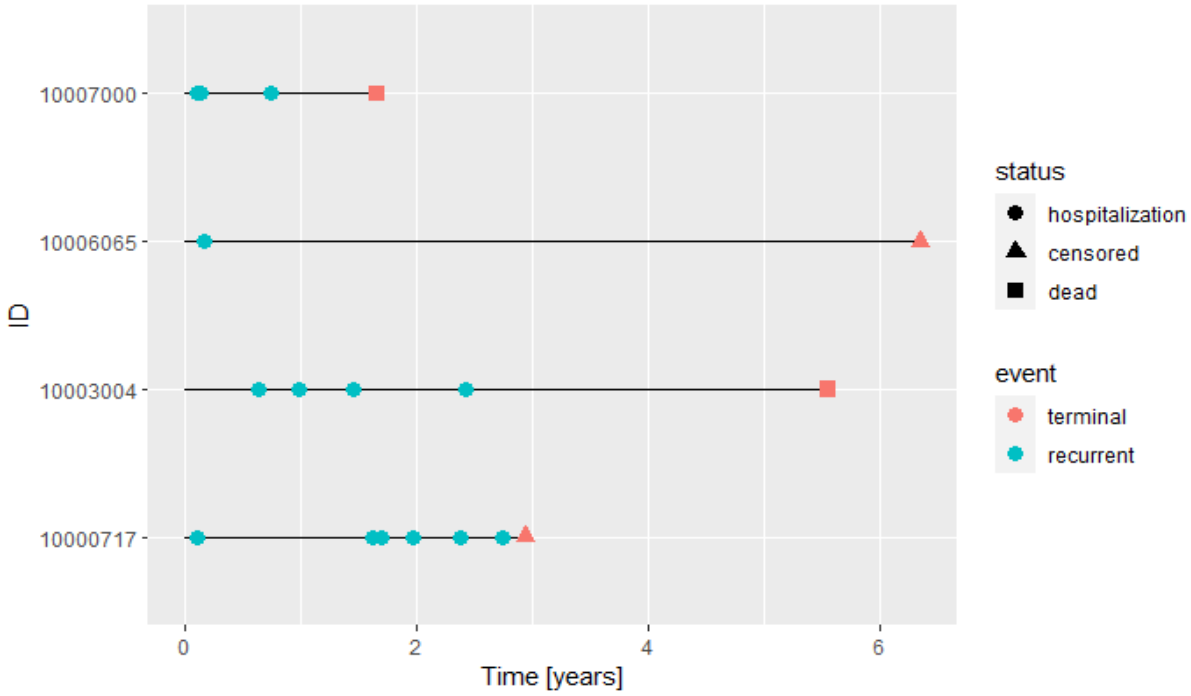

Figure 4.12: Clinical Histories of HF hospitalizations of patients 10007000, 10006065, 10003004 and 10000717. Blued dots represents hospitalizations, red squares deaths and red triangles censoring due to secondary events. Time is expressed in years, starting from each patients' study entry.

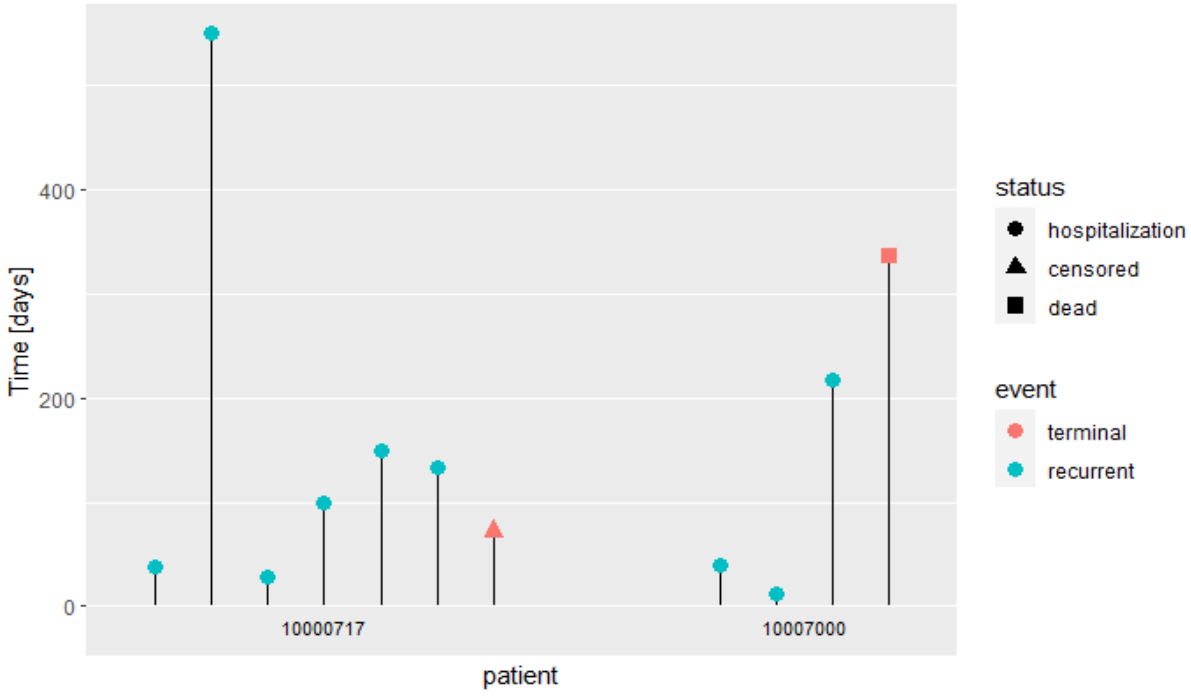

Figure 4.13: Clinical histories of patients 10007000 and 10000717, decomposed in gap times. Gap Times of each patient are ordered from left to right according to the event number. Time (vertical axis) is expressed in days. Notation follows the same rules of Figure 4.12.

In our work, we choose to focus on a time-between-events (i.e. *gap times*) approach (Section 3.2). As mentioned, each patient in our dataset is characterized by a clinical history of hospitalizations due to heart failure, plus a terminal event. Our time outcome of interest, expressed in days, describes the time elapsed from an event to the next one in each patients history: gap time 1 is the time elapsed from study entry to first re-hospitalization, gap time 2 is the time elapsed from re-hospitalization 1 to re-hospitalization 2, until the terminal gap time, defined as the time elapsed from the last recorded hospitalization and the terminal event. Gap times are encoded in the dataset through variable **GapEvent**. The binary variable **hospitalization** describes whether a gap time ends in a hospitalization (1) or a terminal event (0). Binary variable **death**, instead, is defined only for the last gap time of each patient history, and states if the terminal gap time ends with the patients' death(1) or with a secondary terminal event (0). Figure 4.13 displays, as an example, the gap times formulation of the clinical histories of two different patients. Each patients' gap times are ordered, from left to right, according to the event number (gap time 1, gap time 2, ..), the time unit considered is days and each gap time event is encoded following the same notation of Figure 4.12.

In our approach, all the gap times composing a patient's clinical history are equally considered to define the hazard of experiencing a new hospitalization, without taking into account a stratification based on the event number (gap time 1 is considered like gap time 2, etc in the estimation procedure). This is possible thanks to the inclusion in the model of time dependent covariates and random effects (see Section 3.2.1). Figure 4.14 reports gap times stratified by event number. In this plot, gap times for event numbers from 1 to 9 are considered, since they comprehend 90% of data. As expected, the number of subjects experiencing a gap time decreases as the event number increases; the distribution of gap times in each stratum, however, is quite similar, suggesting that the eventual stratification effect would be small.

The instantaneous hazard of death is modeled considering terminal gap times without taking into account the number of hospitalizations experienced in the patients' clinical history. Comparing Figure 4.12 and 4.13, we notice that, in this formulation, some information about the terminal event process may be lost when passing from overall survival times to gap times. Actually, long overall survival times may result in small terminal gap times (patient 10000717) and viceversa. This problem is tackled considering explanatory variables which summarizes the patient condition at the last known hospitalization of each patient.

A common characteristic of the recurrent and terminal gap times distribution can be noticed from the histograms reported in Figure 4.15, which are peaked near zero. That suggests an increased risk for both the two processes in the period following an hospitalization, when the patient is ideally recovering from the previous heart failure event.

The set of potential covariates chosen for modeling the re-hospitalization and death hazards are

- **Sex**: sex of the patient. Factor with levels "M" for males and "F" for females;
- **Adherent**: time dependent binary flag indicating whether a patient is adherent or not to the ACE Inhibitors treatment, computed considering the time window from patient's index date up to the corresponding gap time (see Section 2.3);
- **AgeEvent**: age (in years) at the last known hospitalization;
- **Comorbidity**: number of comorbidities registered at last known hospitalization;

Table 4.4 finally reports an example of the complete data table related to subject 10003004.

ID	Sex	Adherent	AgeEvent	Comorbidity	GapEvent	Event	Death
10003004	F	0	75	5	229	1	
10003004	F	1	75	6	131	1	
10003004	F	0	76	6	168	1	
10003004	F	0	77	7	353	0	
10003004	F	1	79	7	1,143	0	1

Table 4.4: Data table of patient 10003004.

4.2.2. Frailty Proportional Hazard model

As mentioned, the hazards of the hospitalizations and death processes for a patient are modeled through two independent *frailty proportional hazard* models (following the approach of Ripatti and Palmgren as in Equation 3.35)

$$h_i^R(t|\boldsymbol{x}_i^R(t)) = h_0^R(t)\exp\{\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}_i^R(t) + u_i\} \tag{4.2}$$

$$h_i^D(t|\boldsymbol{x}_i^D(t)) = h_0^D(t)\exp\{\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{x}_i^D(t) + v_i\} \tag{4.3}$$

where: R and D refers to hospitalizations and death, respectively; i refers to the patient ID; t refers to a gap time with respect to the last known hospitalization event; $\boldsymbol{x}_i^R(t)$

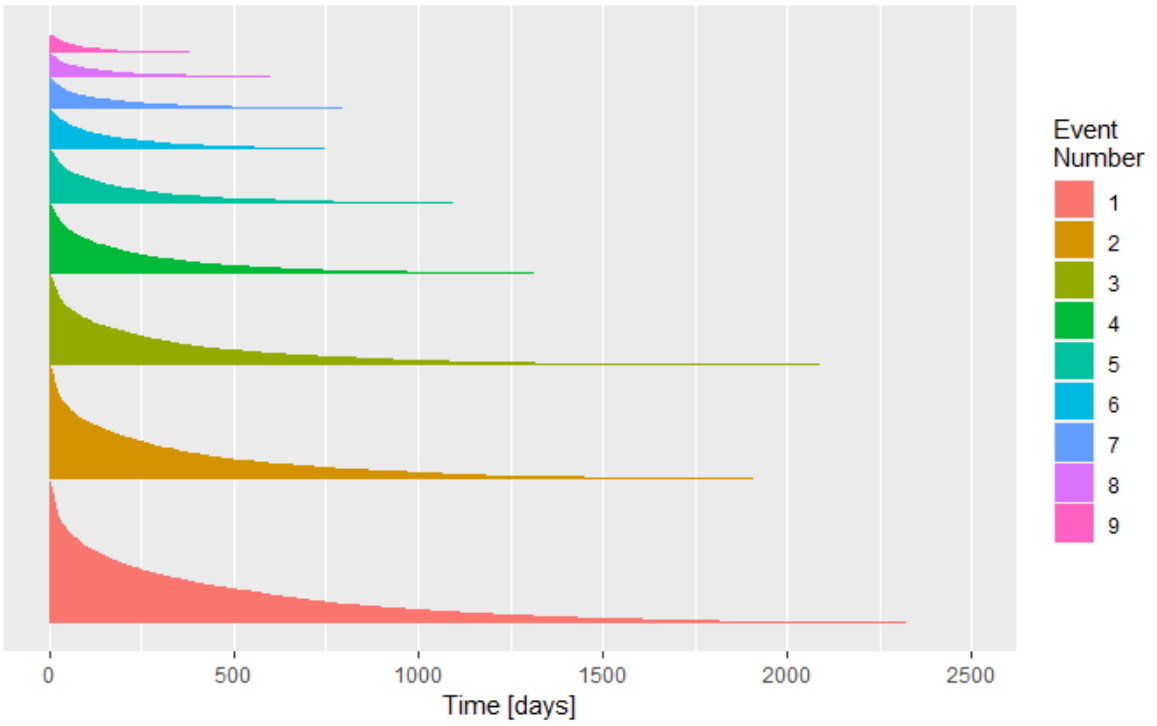

Figure 4.14: Gap times stratified by event number. Gap times for event numbers from 1 to 9 are considered and ordered from the bottom to the top. In each stratum, gap times are ordered from shortest to longest to provide a profile of their distribution. Time is expressed in days.

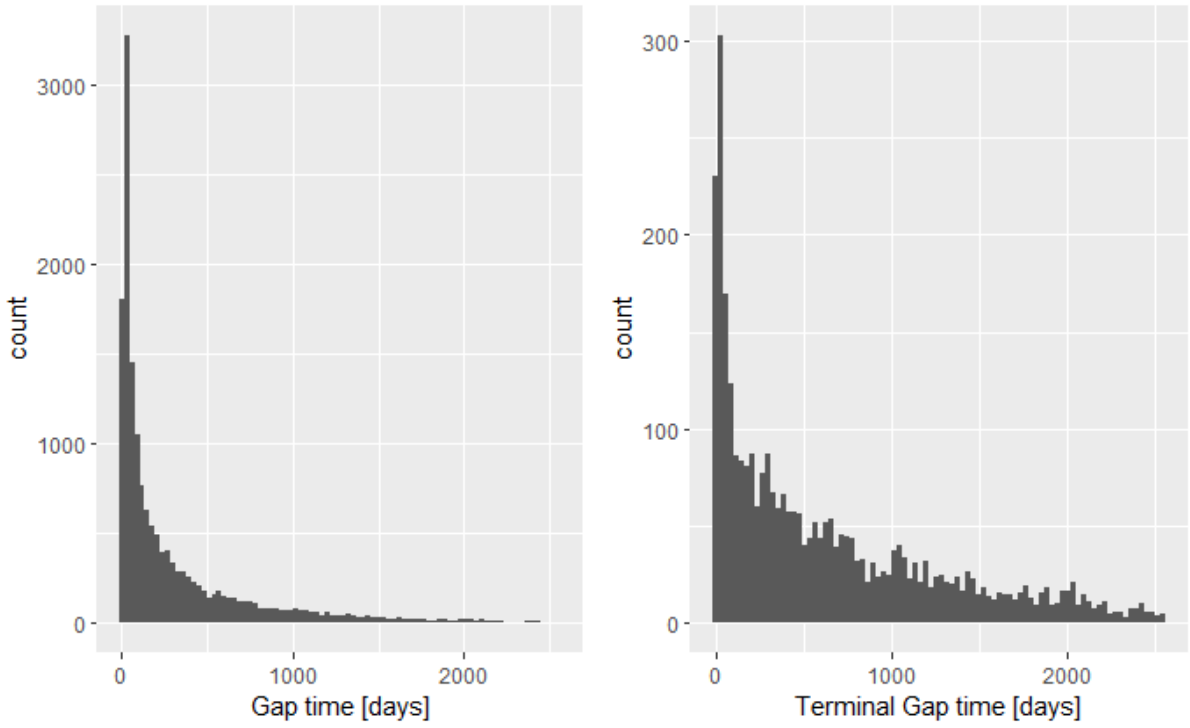

Figure 4.15: Left panel shows the histogram of recurrent events gap times. Right panel shows the histogram of terminal events gap times. Both shows a peaked behaviour near zero. Time is expressed in days.

and $\boldsymbol{x}_i^D(t)$ are the observed covariates at time t; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the estimated coefficients of the two models; u_i and v_i are the patient-specific frailties. We recall that the random intercepts are assumed to be independent, normally distributed and characterized by their variances

$$\begin{aligned} p(u) &= N(0, \theta_u^2) \\ p(v) &= N(0, \theta_v^2). \end{aligned} \tag{4.4}$$

The two considered sets of covariates are

$$\begin{aligned} \boldsymbol{X}_i^R(t) &= \{Sex_i, Adherent_i(t), AgeEvent_i(t), Comorbidity_i(t)\} \\ \boldsymbol{X}_i^D(t) &= \{Sex_i, Adherent_i(t), AgeEvent_i(t), Comorbidity_i(t)\} \end{aligned}$$

Table 4.5 reports the summary of the two models, fitted using the R package `coxme`, which implements the estimation procedure of Ripatti and Palmgren (Section 3.2.1).

Variables	**Estimate**	**StdDev**	**HR**	**CI95**	**pvalue**
Recurrent Events					
Sex [M]	0.046	0.023	1.047	[1.002,1.095]	0.041
Adherent [1]	-0.258	0.021	0.772	[0.741,0.805]	<2e-16
AgeEvent	-0.015	0.001	0.986	[0.983,0.988]	<2e-16
Comorbidity	0.116	0.006	1.123	[1.109,1.138]	<2e-16
θ_u^2 (θ_u)	0.093 (0.304)				
Terminal Event					
Sex [M]	0.141	0.080	1.151	[0.984,1.346]	0.079
Adherent [1]	-0.234	0.084	0.791	[0.671,0.933]	0.005
AgeEvent	0.041	0.004	1.042	[1.033,1.051]	<2e-16
Comorbidity	0.446	0.021	1.497	[1.497,1.629]	<2e-16
θ_v^2 (θ_v)	0.428 (0.654)				

Table 4.5: Summary of the Recurrent and Terminal events frailty models specified by Equations 4.2 and 4.3. For categorical variables, the considered stratum is indicated between brackets.

The covariate **Sex** is included in both the models to assess if a different trend in hospitalizations exists between males and females and assess if the effect on death is almost negligible, like in the classical Cox model in Section 4.1.2. It results in the first case statistically significant at 5%, while not significant in the second (pvalues, respectively, of 0.041 and 0.079) and suggests that male subjects are slightly more prone to risk of hospitalization (HR=1.044) and death (HR=1.164).

The covariate **Adherent** is included in both models to assess the effect of adherence to the treatment of ACE Inhibitors both on the hospitalization and death hazards. In both models it results statistically significant at any level, showing pvalues, respectively, less than 2e-16 for the hospitalizations model and 0.005 for the terminal event model. In particular, being adherent to the treatment yields a 22.8% decrease in the risk of a new hospitalization (HR=0.844) and a 20.9% (HR=0.791) decrease of the risk of death. This suggests ACE inhibitors treatment to have a strong positive effect in both reducing hospitalizations and deaths in heart failure patients.

The covariate **AgeEvent** is included initially in both models because it is likely to restore part of the information lost when passing from overall survival times to gap times. It results statistically significant at all levels. Its effect on the hospitalization hazard is a 1.4% reduction of the risk of hospitalization per year (HR=0.986), while its effect on the death hazard is an increase of the risk of the 4.2%. The two effects are likely to be correlated, since part of the risk of experiencing a new hospitalization is replaced by the risk to die when patients get older.

The covariate **Comorbidity** is included in both models, since it is an indicator of the worsening of a patients' conditions, which is likely to influence both the risk of hospitalization and death. It results in both models statistically significant at all levels, and shows an high increase of 11.6% in the risk of hospitalization per comorbidity registered and a very high increase of 44.6% in the risk of death per comorbidity registered.

As discussed in Section 3.2.1, frailties can be interpreted as a measure of the unexplained heterogeneity at patients' level in the the recurrent and terminal events processes. We assume them to be normally distributed with zero mean and, thus, characterized by their variances. In this case, the hospitalization frailty (u in Equation 4.2) variance estimate

is 0.093, while the terminal events frailty (v in Equation 4.3) one is 0.428. The fact that the within-patient variability is very low in the modeling of the hospitalizations process, while is quite high in the death one, may be due to the fact that the two processes shows different timescales (see Figure 4.15), as well as due to the two processes being trained on very different amounts of data (15,978 against 3,232). However, this difference can be explained also from a clinical point of view, as subjectivity is likely to be more relevant on mortality that on hospitalizations, as they are regulated by fixed procedures. At this stage the two frailties are modeled independently, even if a correlation between the two random effect is reasonable. In the next Section we consider joint models of the two processes which relies on correlated frailties.

Moreover, in Section 3.2.1 we also discuss the introduction of random effects as a solution to cope with data correlation, which are likely to cause introduction of bias in the coefficient estimates and artificial shrinkage of their variances (*overdispersion*). For this reason, we fit two classical Cox models, one to model the hospitalizations process and one to model the terminal events process, on the same data and compare the results with the discussed frailty models. Figure 4.16 and 4.17 reports, respectively, an hazard ratios estimates and 95% confidence intervals comparison for the recurrent events process and the terminal events process. Confidence intervals of frailty models are coloured in blue, while classical Cox model ones are coloured in red. Hazard ratios estimates are denoted by dots. The pointwise estimates of the two classes of models are congruous, even if some differences can be appreciated both for the recurrent events model and the terminal events model. Frailty models confidence intervals, instead, are definitely larger then their classical Cox counterparts, primarily, as expected, as far as recurrent events are concerned. This is important, since underestimating a coefficient variance may lead to wrong decisions about is statistical significance.

4.3. Joint Modelling

In this section we show the results obtained when jointly modelling the processes of hospitalizations and deaths. In Subsection 4.3.1 we fit the model proposed by Rondeau et al. and described in Section 3.3.1, while in Subsection 4.3.2 we consider the novel model proposed by Ng et al., discussed in Section 3.3.2. The dataset considered, comprehending the time-between-events outcome, the censoring variables and the explanatory variables, is the same used in Section 4.2.2 to fit the two independent frailty models.

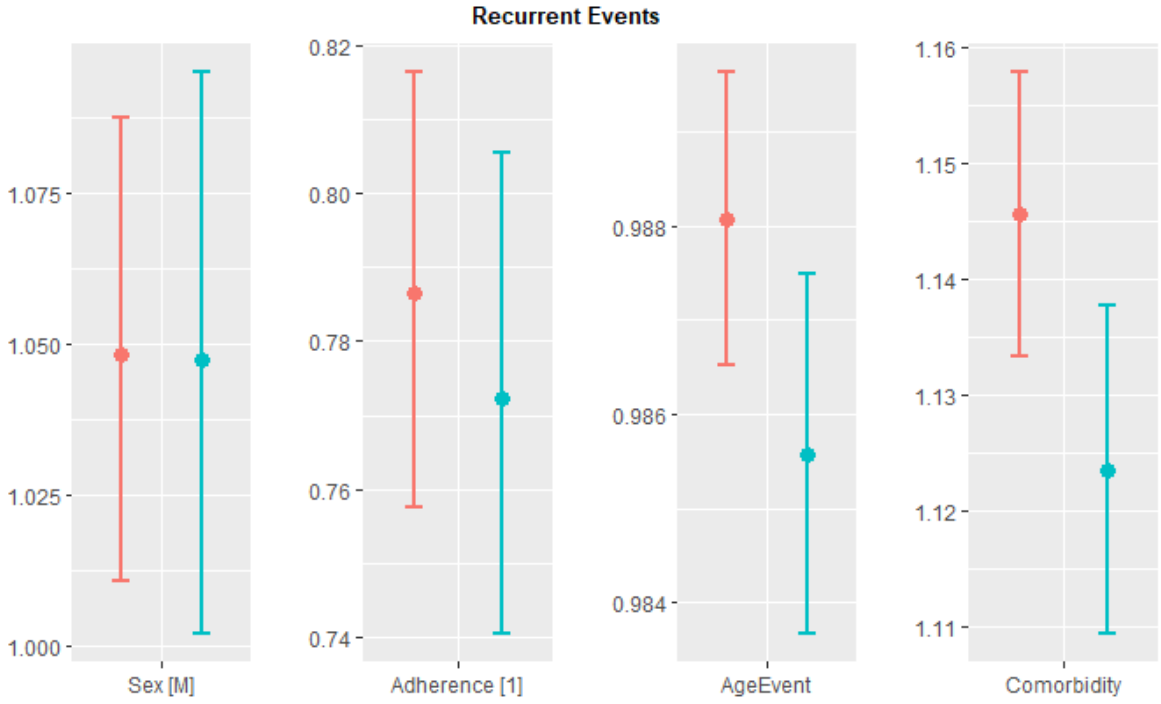

Figure 4.16: Comparison of HR 95% confidence intervals for the recurrent event process. Red intervals corresponds to the classical Cox model, while blue intervals to its frailty counterpart. Dots represents pointwise estimates.

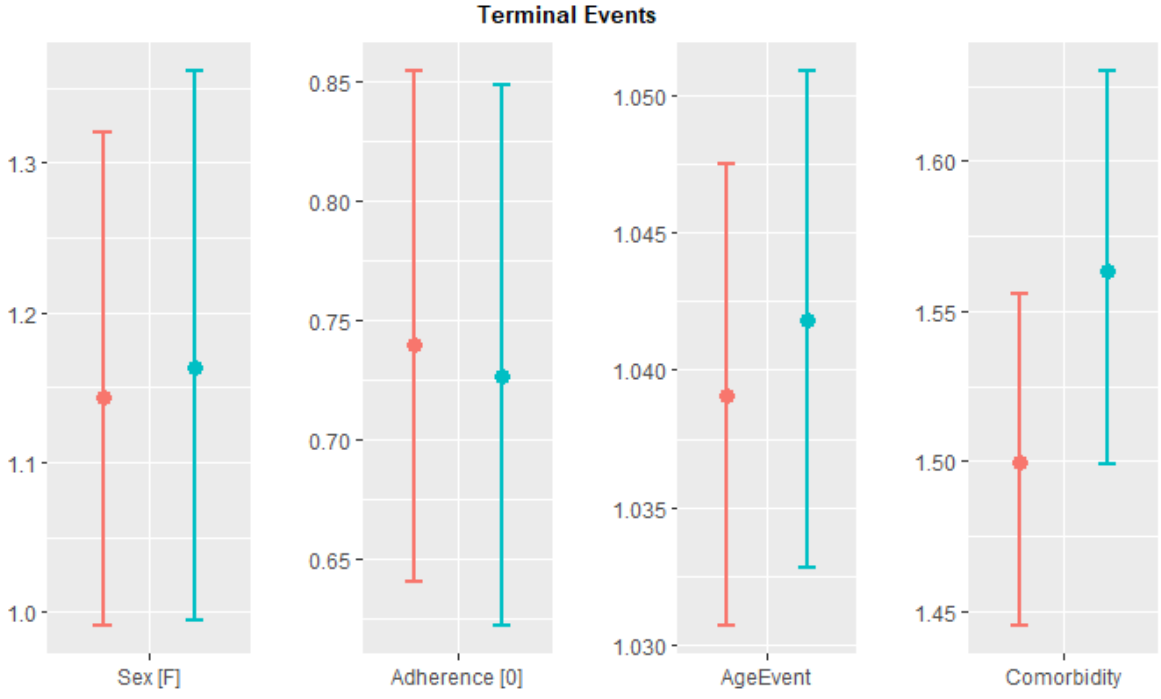

Figure 4.17: Comparison of HR 95% confidence intervals for the terminal event process. Red intervals corresponds to the classical Cox model, while blue intervals to its frailty counterpart. Dots represents pointwise estimates.

4.3.1. Joint Model by Rondeau et al.

In the approach proposed by Rondeau et al. (see Section 3.3.1) the instantaneous hazards of a new hospitalization and death are modeled as in Equation 3.44

$$\begin{cases} h_i^R(t|\eta_i, \boldsymbol{x}_i^R(t)) = h_0^R(t)\exp\{\eta_i + \boldsymbol{\beta}^T\boldsymbol{x}_i^R(t)\} \\ h_i^D(t|\eta_i, \boldsymbol{x}_i^D(t)) = h_0^D(t)\exp\{\alpha\eta_i + \boldsymbol{\gamma}^T\boldsymbol{x}_i^D(t)\} \end{cases} \tag{4.5}$$

where: R and D refers to hospitalization and death, respectively; i refers to the patient ID; t refers to a gap time with respect to the last known hospitalization event; $\boldsymbol{x_i^R}(t)$ and $\boldsymbol{x_i^D}(t)$ are the observed covariates at time t; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the estimated coefficients of the two models; η_i is the patient-specific random intercept in the recurrent events process, while α is the multiplicative parameter which quantify the effect of the patient frailty on the terminal event process. We assume a Gaussian distribution for the random effect

$$p(\eta) = N(0, \sigma^2) \tag{4.6}$$

characterized by the variance parameter σ^2. Moreover, we recall that the considered set of covariates are

$$\begin{aligned} \boldsymbol{X_i^R}(t) &= \{Sex_i, Adherent_i(t), AgeEvent_i(t), Comorbidity_i(t)\} \\ \boldsymbol{X_i^D}(t) &= \{Sex_i, Adherent_i(t), AgeEvent_i(t), Comorbidity_i(t)\} \end{aligned}$$

The model is fitted through the R package `frailtypack`. Several hyperparameters need to be chosen (see Section 3.3.1 for details) in order to carry out the estimation procedure:

- the number of Laguerre points considered when using Gaussian quadrature to approximate integrals is 20;
- the number of knots of M-splines used to model the baseline hazard functions $h_0^R(t)$ and $h_0^D(t)$ is 12.

Moreover, the two smoothing parameter in the penalized log likelihood (k_1 and k_2 in Equation 3.49) are estimated, as suggested in [53], through a cross-validation procedure implemented in the package, which relies on the fitting of two independent frailty models for the two processes. Table 4.6 reports the summary of the estimated model.

Variables	Estimate	StdDev	HR	CI95	pvalue
Recurrent Events					
Sex [M]	0.044	0.023	1.044	[0.995,1.093]	0.060
Adherent [1]	-0.263	0.022	0.769	[0.736,0.802]	<2e-16
AgeEvent	-0.014	0.001	0.986	[0.984,0.988]	<2e-16
Comorbidity	0.118	0.007	1.125	[1.109,1.141]	<2e-16
Terminal Event					
Sex [M]	0.055	0.081	1.056	[0.901,1.238]	0.497
Adherent [1]	-0.211	0.087	0.810	[0.683,0.960]	0.015
AgeEvent	0.041	0.003	1.042	[1.036,1.048]	<2e-16
Comorbidity	0.137	0.023	1.147	[1.096,1.199]	2.07e-09
Frailty					
σ^2	0.114	0.009			<2e-16
α	2.672	0.318			<2e-16

Table 4.6: Summary of the Joint Recurrent and Terminal events frailty model proposed by Rondeau et al. and specified in Equation 4.5. For categorical variables, the considered stratum is indicated between brackets.

The estimated joint model is overall quite similar to the disjoint model considered in the previous Section. The covariate **Sex** has the same effect on the hospitalization hazard (HR=1.044, pvalue=0.060) as in the independent model, while its effect and statistical significance reduce as far as death hazard is concerned (HR=1.056, pvalue=0.497). The covariate **Adherent** shows the same behaviour as in the independent models, being statistically significant at any levels for the hospitalizations process and at 5% for the terminal events one, yielding respectively a HR of 0.769 for recurrent events and a HR of 0.810 for death. Also the covariate **AgeEvent** shows a very similar effect as before, yielding almost the same estimates (HR=0.986, pvalue<2e-16 for hospitalizations and HR=1.042, pvalue<2e-16 for death). The covariate **Comorbidity**, instead, shows the very same effect on the hospitalization hazard as the independent model (HR=1.125), but a much lower effect on the death hazard (HR=1.147 against 1.563 of the corresponding independent model). This could be due to possible difficulties encountered in the estimation routine adopted, which is likely to be unstable due to the high number of hyperparameters to be tuned.

In general, modeling independently the recurrent and terminal events process should not lead to particular differences in the coefficients' estimates; however, when the frailties correlation is not considered their effect (i.e. their variance) is likely to be underestimated. In our case, the estimated variance of the random effect involved in the recurrent events process is slightly higher than its independent model counterpart (0.114 against 0.093). The estimated α parameter, instead, is 2.672. Since it acts multiplicatively on η_i to define the terminal events random effect, it yields a final variance of 0.799, which is significantly higher than the independent model one (0.428).

In conclusion, the model by Rondeau et al. provides, in our case, a starting tool to model the joint recurrent and terminal events processes through dependent frailties. The main drawbacks are, however, the complexity of the estimation procedure, which is dependent from the choice of the aforementioned hyperparameters and often numerically unstable, and the difficult interpretation of the α parameter.

4.3.2. Joint Model by Ng et al.

The model proposed by Ng et al. (Section 3.3.2) suggests similar shapes for the instantaneous hazards of the recurrent and terminal events processes as the one proposed in the case of disjoint modeling (see Equation 3.52) as follows

$$\begin{cases} h_i^R(t|\boldsymbol{x}_i^R(t)) = h_0^R(t)\exp\{\boldsymbol{\beta}^T\boldsymbol{x}_i^R(t) + u_i\} \\ h_i^D(t|\boldsymbol{x}_i^D(t)) = h_0^D(t)\exp\{\boldsymbol{\gamma}^T\boldsymbol{x}_i^D(t) + v_i\} \end{cases} \tag{4.7}$$

where the followed notation is the same of Equation 4.5. Two distinct frailties (u_i and v_i) are considered at patient's level for the modeling of the two processes, but their distribution is jointly modeled as a bivariate Normal distribution of parameters

$$p\left(\begin{bmatrix} u \\ v \end{bmatrix}\right) = \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \theta_u^2 & \rho\theta_u\theta_v \\ \rho\theta_u\theta_v & \theta_v^2 \end{bmatrix}\right). \tag{4.8}$$

Again, the considered set of covariates related to the ACE Inhibitors dataset are

$$\begin{aligned} \boldsymbol{X}_i^R(t) &= \{Sex_i, Adherent_i(t), AgeEvent_i(t), Comorbidity_i(t)\} \\ \boldsymbol{X}_i^D(t) &= \{Sex_i, Adherent_i(t), AgeEvent_i(t), Comorbidity_i(t)\}. \end{aligned}$$

The main novelty of the approach by Ng et al.[37] is represented by the proposed estima-

tion ruotine, described in Section 3.3.2, which alternates matricial Newton-Raphson steps to update pointwise estimates of $\mathbf{\Omega} = [\beta, \gamma, u, v]$, and updates for the variance parameters (θ_u,θ_v and ρ) based on the closed-form solution of the REML first derivative equation. The estimation is carried out through the use of the R software, but since no package is yet available the routine was custom-implemented as reported in Appendix B. Table 4.7 reports the summary of the obtained model.

Variables	**Estimate**	**StdDev**	**HR**	**CI95**	**pvalue**
Recurrent Events					
Sex [M]	0.042	0.024	1.043	[0.996,1.092]	0.087
Adherent [1]	-0.263	0.022	0.768	[0.736,0.802]	<2e-16
AgeEvent	-0.015	0.001	0.985	[0.983,0.987]	<2e-16
Comorbidity	0.118	0.007	1.126	[1.111,1.141]	<2e-16
Terminal Event					
Sex [M]	0.160	0.090	1.174	[0.984,1.400]	0.132
Adherent [1]	-0.309	0.092	0.734	[0.612,0.879]	0.005
AgeEvent	0.037	0.005	1.038	[1.028,1.048]	1.3e-12
Comorbidity	0.379	0.024	1.461	[1.392,1.533]	<2e-16
Frailty					
θ_u^2	0.124	1.7e-06			<2e-16
θ_v^2	1.389	4.59e-07			<2e-16
ρ	0.880	4.5e-03			<2e-16

Table 4.7: Summary of the Joint Recurrent and Terminal events frailty model proposed by Ng et al. and specified in Equation 4.7. For categorical variables, the considered stratum is indicated between brackets.

From the coefficients' estimates point of view, the model is similar to its disjoint counterpart, confirming the interpretation of factors. However, some little discrepancies can still be found.

The covariate **Sex** results to be in both processes not statistically significant at 5% (pvalue, respectively, of 0.087 and 0.132). It is again confirmed the suggestion that male subjects are slightly more prone to risk of hospitalization (HR=1.043) and death (HR=1.174).

The covariate **Adherent** results to be statistically significant at any level, showing pvalues, respectively, less than 2e-16 for the recurrent events model and 0.005 for the terminal event model. Being adherent to the treatment yields a 23.2% decrease in the risk of a new hospitalization (HR=0.768) and a 26.6% decrease in the risk of death (HR=0-734).
The covariate **AgeEvent** results, as usual, statistically significant at all levels. Its effect on the hospitalization hazard is a 1.5% reduction of the risk of hospitalization per year (HR=0.985), while on the death hazard it yields an increase of the risk of the 3.8% (HR=1.038).
The covariate **Comorbidity** results in both models statistically significant at all levels, and shows an high increase of 12.6% in the risk of hospitalization per comorbidity registered and a very high increase of 46.1% in the risk of death per comorbidity registered.

The small differences noticed between the coefficients estimates of the disjoint model (see Table 4.5) and the one by Ng et al. (see Table 4.7) endorse the validity of the interpretation provided, as well as the custom implementation of the estimation routine for the latter model. In this perspective, the fact that the model by Rondeau et al. provides some slightly more different estimates, above all for the comorbidity hazard ratio of the death process (1.147 against 1.497 of the disjoint model and 1.461 of the model by Ng et al.), backs up the hypothesis that its estimation process could be subject to numerical issues. However, it is worth to notice that most of the covariates coefficients are estimated similarly by all the three models, and in any case the identification of risk against protective factors is concordant.

Besides coefficients estimation, the novelty is represented by the bivariate Normal characterization of frailties distribution. The estimated values are $\theta_u^2 = 0.124$ for the variance of unaccounted heterogeneity in the recurrent events process and $\theta_v^2 = 1.378$ for its terminal events process counterpart. Looking at the estimates of the disjoint model, the difference is quite notable, with a 25% increase in θ_u^2 (from 0.093 in the disjoint case to 0.124 in the model by Ng et al.) and almost a 70% increase (from 0.428 to 1.389) in θ_v^2. This magnification in the random effects variance estimates, when considering disjoint models fitted through the `coxme` package [59], was also mentioned in the paper by Ng et al. [37], who further investigated it through simulation studies. As already mentioned in the previous Section 4.3.1, this magnifying effect can be supposed to be in part due to not considering, in the disjoint models, the correlation between processes, now accounted for by the parameter ρ. In particular, in our case the estimate for the correlation parameter is very

high (0.880), suggesting a very strong dependence between the two processes frailties. This suggests, from a medical point of view, that in our cohort patients which are more prone to the risk of a new hospitalization (due to unaccounted, random characteristics at patients level) are also more prone to the risk of death. A possible explanation for this phenomenon is that the patients in the worst clinic conditions are the ones more at risk both of a new hospitalization and of death. However, even if highly correlated, the strong difference between the two processes random effects variances suggests a far more important role of randomicity in the terminal event process.

In spite of a good stability in coefficients estimation and an elegant and interpretable random effects characterization, the implemented model by Ng et al. presents some drawbacks. In particular, from an estimation procedure point of view, the implemented routine involves complex matricial products in the internal loop related to Newton-Raphson steps, between entities whose dimensions are in the order of the number of patients in the considered cohort and the number of total observed gap times (see Appendix B, rows 120 – 203). This results in high computational power and memory requirements when treating medium-high dimensions datasets, like the ACE Inhibitors dataset (that, we recall, comprehends 15,978 gap times registered per 3,232 patients). A significant gain in speed is obtained parallelizing matricial products through a `Rcpp` [14] dedicated function. Another bottleneck is represented by the inversion of the information matrix $\boldsymbol{G}$ (see Equations 3.59 - 3.60 and Appendix B, rows 205 – 220), necessary to compute the REML parameters update. In this case, the problem was tackled by adopting the R built-in Cholevski factorization strategy to invert the matrix. In our case, the estimation was carried out on the MOX HPC system [35], considering parallel product over 40 processes. The model training required approximately 8 minutes per Newton-Raphson iteration and 2 minutes per REML parameters update. The investigation of alternative or better optimized estimation procedures represents a possible field for future development of this model. Actually, even if the convergence rate is usually high for what it concerns each Newton-Raphson optimization (i.e. the internal loop usually converges in few iterations), the algorithm may requires a high number of outer loop iterations in case of a flat REML loglikelihood. This would result in a overall highly time consuming training of the model. In our case, for example, to match the given convergence threshold of 1e-03 between subsequent values of variance parameters (computed in euclidean norm), almost 100 iterations were necessary. Figure 4.18 visualize the smooth evolution of estimated random effects variance parameters across iterations of the outer loop of the estimation routine.

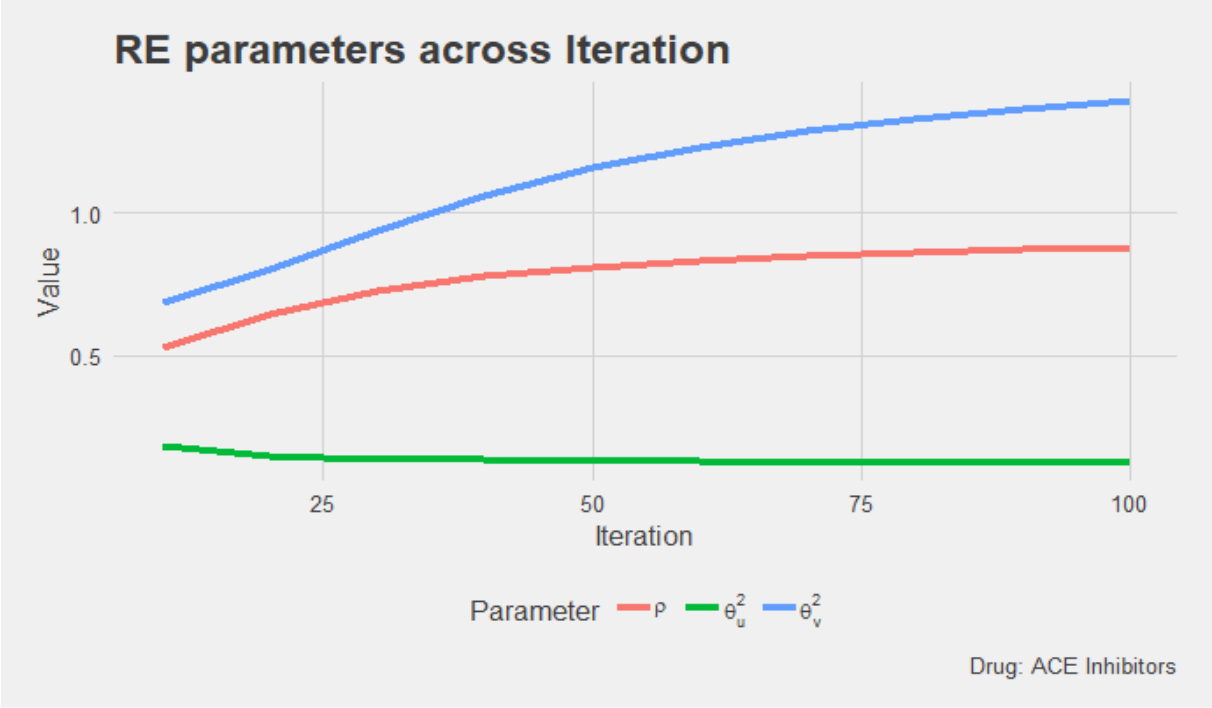

Figure 4.18: Random Effects variance-covariance matrix parameters evolution across outer loop iteration for model in Equation 4.7 using the estimation routine reported in Appendix B.

Besides computational and convergence issues, it is worth to discuss the considered bivariate Normal characterization of the random effects from providers' assesment point of view. In general, frailty models adopt parametric distribution with the aim of quantifying the heterogeneity at patients' level. In our case, the choice to jointly model the two considered processes frailties through a bivariate Gaussian gives an elegant mathematical characterization, providing, along to the usual parameters for heterogeneity (i.e. θ_u and θ_v), a bounded, interpretable parameter for the correlation between the frailties (i.e. ρ). Besides suggesting important insights on the considered phenomenon, this characterization may lack immediacy in practice. As an example in such direction, a discrete distribution of random effects may be able to suggest different frailty profiles to better organize management of patients. Model in Equation 4.7 produces, as a byproduct, pointwise estimates of the random effects of each patients, $\boldsymbol{u}$ and $\boldsymbol{v}$, upon which we can perform an a posteriori analysis. Figure 4.19 visualizes the pointwise estimates on $\mathbb{R}^2$, considering the hospitalizations process frailties u_i as abscissa and the death process frailties v_i as ordinata.

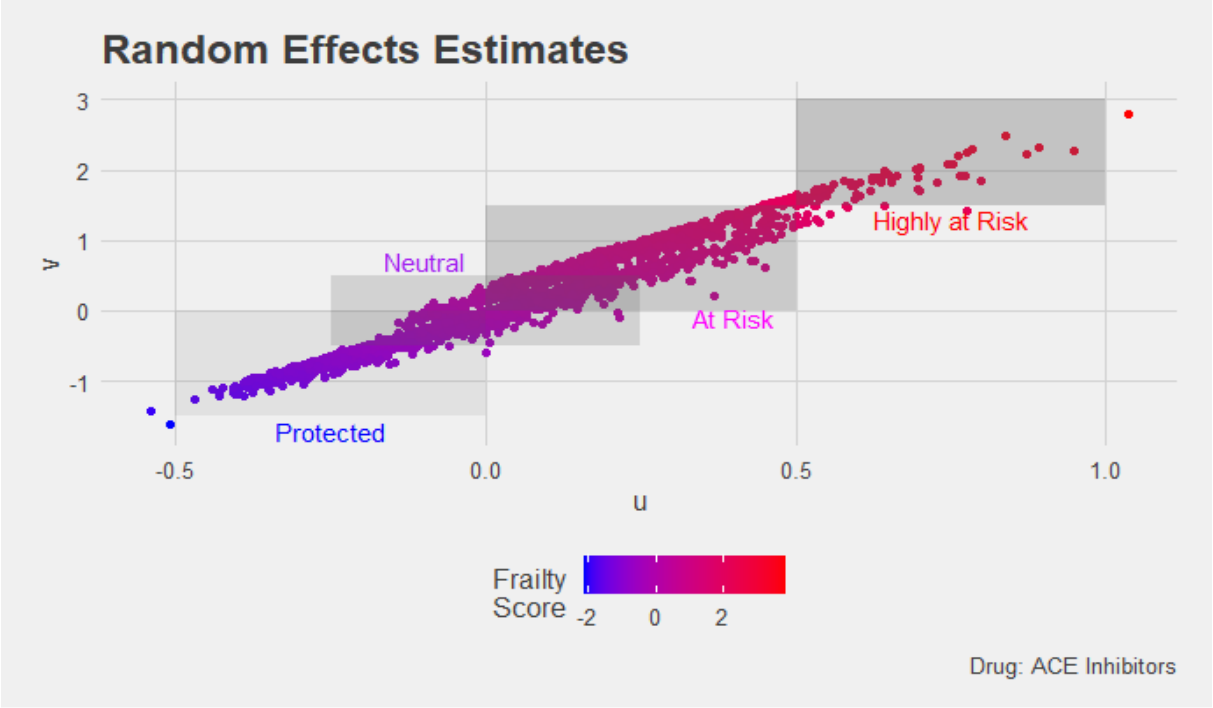

Figure 4.19: Random Effects pointwise estimates for model by Ng et al. in Equation 4.7. Points are visualized in $\mathbb{R}^2$ considering hospitalization frailties as abscissa and death frailties as ordinata. Due to the points pattern, their colouring follows a simple frailty score computed as the sum of the two points' coordinates. The overlapping boxes identify a profiling of patients according to such score.

As we can see, due to the high correlation, the points (u_i, v_i) disposition follows a diagonal line: for this reason, it is easy to define, in principle, a simple frailty score as the sum of a points' coordinates. Such score, which ranges from -2 to 4, is used to color the points in the plot, in order to distinguish in a gradual way more robust patients (in blue) from more fragile ones (in red). In fact, blue points reflect patients with a smaller risk of a new hospitalization and death due to unaccounted characteristics at patient's level (i.e. more robust), whereas red points are characterized by patients with a bigger risk of a new hospitalization and death due to unaccounted characteristics at patient's level (i.e. more fragile). Then, points are grouped: the shaded boxes in the figure are a visual attempt to identify a classification of patients based on added risk in the two processes due to randomicity. In particular, we can notice

- patients neutral to the random added risk, that can considered as a baseline for the population (purple *Neutral* box centered in $O = (0, 0)$);
- patients protected with respect of the baseline (blue *Protected* box);

- patients slightly more at risk with respect to the baseline (magenta *At Risk* box);
- outliying patients highly more at risk both of a new hospitalization and of death (dark red *Highly at Risk* box).

The density of points in the mentioned boxes can act as a (coarse) approximation of the probability to belong to such frailty profiles, allowing the definition of management strategies to cope with the random risk for the two processes. Of course, the described qualitative reasoning serves purely as an example, lacking the necessary rigour to be used as an effective tool. In general, there exist a posteriori techniques which exploits the mentioned byproduct to perform analysis similar to the one described, which may resort to clustering or mixture density estimation techniques. Being aware of the expendability in the practical context of a discrete distribution of frailties, we extended the model proposed by Ng et al. to a novel comprehensive approach which provides each patient with a bivariate nonparametrically and discretely distributed frailty, as deeply described in Section 3.4. This innovative method is now applied to ACE inibithors subdataset.

4.4. Discrete Nonparametric Frailty

In this section, we present the results obtained fitting our new approach proposed in Section 3.4, where the frailties of the two processes are non parametrically and discretely distributed. The model was designed, as mentioned, to be able to discover a characterizing discrete distribution (in $\mathbb{R}^2$) for the bivariate frailties, which may be desirable, in practice, for providers' assesment.

The instantaneous hazards are modelled as in Equation 3.62 as follows

$$\begin{cases} h_i^R(t|\boldsymbol{x}_i^R(t)) = h_0^R(t)\exp\{\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x}_i^R(t) + u_i\} \\ h_i^D(t|\boldsymbol{x}_i^D(t)) = h_0^D(t)\exp\{\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{x}_i^D(t) + v_i\} \end{cases} \tag{4.9}$$

where: $h_0^R(t)$ and $h_0^D(t)$ are, respectively, the baseline hazard functions of the recurrent and terminal events processes; $\boldsymbol{x}_i^R(t)$ and $\boldsymbol{x}_i^D(t)$ are, respectively, the set of possibly time dependent covariates included in the linear predictor of the hospitalizations and death process; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ the vector of to-be-estimated coefficients relative to the mentioned covariates sets; u_i and v_i represent the abscissa and the ordinata of the subject i point, which belongs to the support of the discrete distribution. The bivariate random effects are distributed according to a generic, discrete measure of $\mathbb{R}^2$

$$[u, v]_i \overset{iid}{\sim} P^* \quad \forall i = 1, ..., N \tag{4.10}$$

which can be represented by its support (i.e. a finite set of points $\boldsymbol{P}$ in $\mathbb{R}^2$) and a set of weights (i.e. the probabilities of a patient to be assigned to a point, $\boldsymbol{w}$)

$$[u, v]_i \overset{iid}{\sim} \begin{cases} \boldsymbol{P}_1 & w_1 \\ \boldsymbol{P}_2 & w_2 \\ ... & \\ \boldsymbol{P}_L & w_L \end{cases} \qquad \forall i = 1, ..., N \tag{4.11}$$

The support characteristics are unknown a priori and its estimation is performed, jointly to the weights and the hazards' parameters, by the algorithm. In particular, we adopt a support reduction strategy (see Algorithm 3.3), which considers an initial grid of points, designed to cover the supposed region in which the true hidden support lies, and then updates and refines it according to the proposed EM algorithm (see Section 3.4.4).

As usual, we present the results obtained fitting the model on the ACE Inhibitors dataset. The considered sets of covariates for modeling the two processes' hazards are, as before

$$\boldsymbol{X}_i^R(t) = \{Sex_i, Adherent_i(t), AgeEvent_i(t), Comorbidity_i(t)\}$$
$$\boldsymbol{X}_i^D(t) = \{Sex_i, Adherent_i(t), AgeEvent_i(t), Comorbidity_i(t)\}.$$

As discussed in Section 3.4.5, the proposed estimation routine mainly depends on three model design choices: (i) the considered distance in the support reduction procedure; (ii) the initialization of the grid; (iii) the threshold under which two distinct points of the grid are merged (in the following, *MinDist*). With respect to the choice of the distance, we consider, for all our analysis, the Euclidean distance. This is done due to its simple interpretation when it comes to the identification of significantly distinct points in the random effects distribution support. As far as the grid initialization of the support is concerned, we will consider the two approaches mentioned in the methodological part (see Section 3.4): the bivariate Gaussian initialization and the Uniform distribution over a rectangle initialization. In particular, we compare and provide interpretation of the results obtained in the two cases, considering an a priori chosen value for the *MinDist* threshold, which was supposed to produce a significant distinction of points (Subsection 4.4.1). Then, since the tuning of such threshold parameter is the most relevant model

design choice, we perform a simple sensitivity analysis in such direction, still differentiating with respect to the two initializations (Subsection 4.4.2). Finally, since the estimation routine involves randomization, we perform an analysis to assess the sensitivity of the estimation, still considering the two different initialization cases (Subsection 4.4.3).

4.4.1. Gaussian Initialization & Uniform Initialization

Initially, we choose a *MinDist* value of 0.25, believed to be suitable to identify significantly different latent populations, in order to estimate the two different initialization models and compare them.

The first initialization strategy considered consists in sampling a fixed (high) number of points from a bivariate Gaussian distribution, centered in the origin of $\mathbb{R}^2$ and characterized by a diagonal variance-covariance matrix. The variance parameters are set looking at the estimates obtained fitting the model obtained implementing the approach by Ng et al. in Equation 4.7 ($\theta_u^2 = 0.12, \theta_v^2 = 1.4$). We impose zero correlation in the initial grid definition to ensure a proper exploration of the space and to avoid a too informative initialization. Weights for each points are computed using the considered bivariate Gaussian distribution density and then normalized to sum to the unit. We consider a sample of 1000 points, composing the Gaussian initial grid reported in Figure 4.20. Each point is colored according to a gradient scale from blue to red, which distinguishes points associated with decreased risk (blue) from one associated with increased risk (red). The size of each point reflects its weights in the discrete distribution. Note that this grid undergoes a support reduction step before the estimation procedure starts.

The initialization strategy of a Uniform distribution over a rectangle follows a slightly different procedure: as before, the rectangle is centered in the origin of $\mathbb{R}^2$, while the length of its sides is set to six times the standard deviation of the corresponding variance parameters of the model obtained implementing the approach by Ng et al. in Equation 4.7 ($\theta_u^2 = 0.12, \theta_v^2 = 1.4$). This is done in order to cover the region in which the support is believed to lie, according to Figure 4.19. Then, the defined rectangle area is filled with equispaced points (in both directions), with distance fixed to the value of *MinDist*. Weights are initialized uniformly. Figure 4.21 reports the obtained Uniform initial grid, consisting in 261 points. Points are colored adopting a gradient from blue to red (as in Figure 4.20 related to the Gaussian initial grid), which identifies four main different effects: reduced risk both in hospitalization and death (blue, left-down corner); reduced risk in hospitalization, increased risk in death (magenta, left-high corner); increased risk

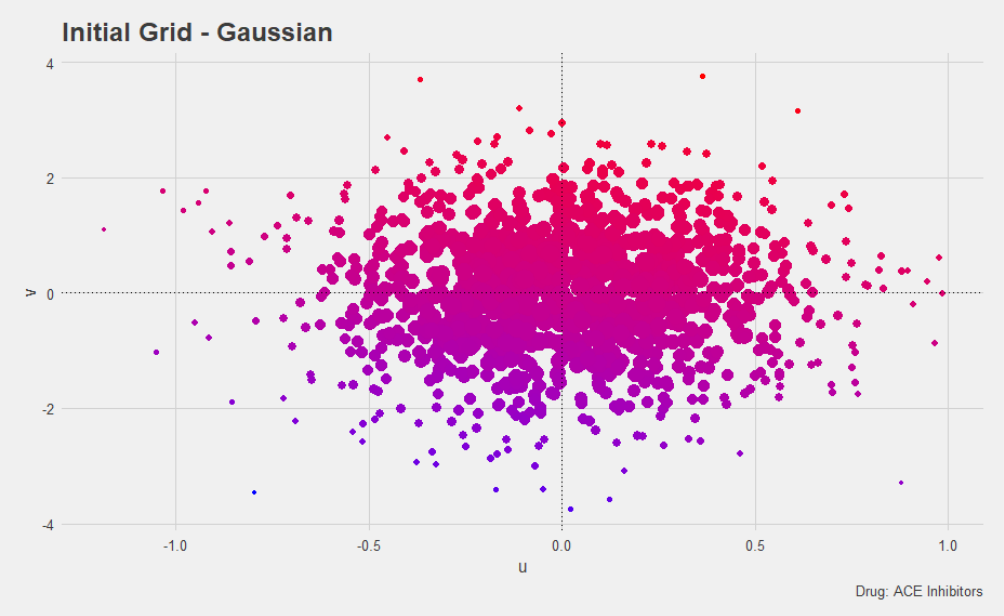

Figure 4.20: Gaussian initial grid for the random effect distribution of the Nonparametric discrete frailty model. Each point is colored according to a gradient scale from blue to red, which distinguishes points associated with decreased risk (blue) from one associated with increased risk (red). The size of each point reflects its weight in the discrete distribution.

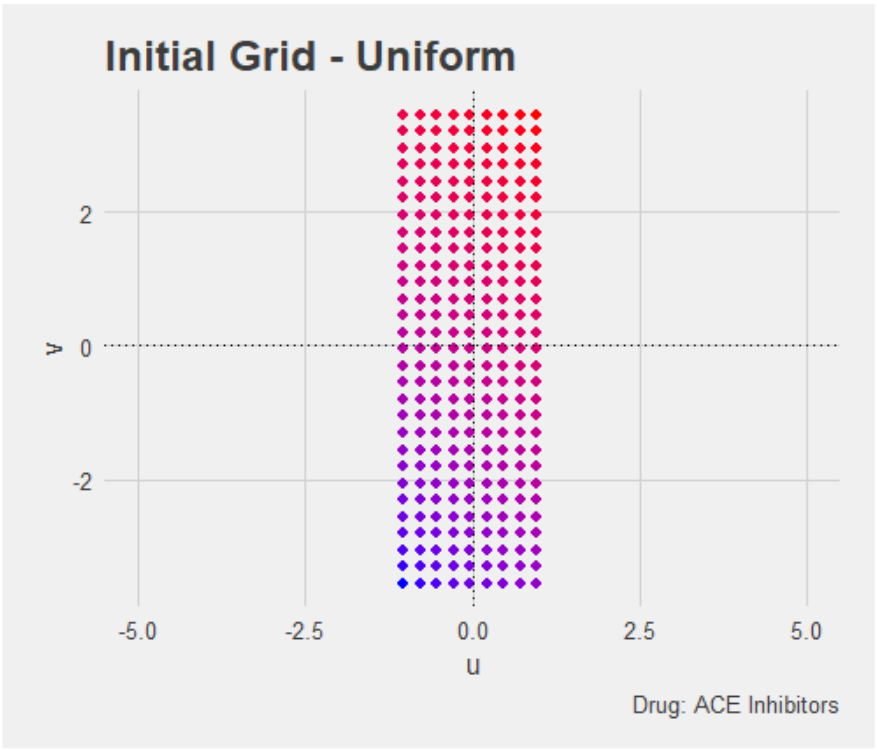

Figure 4.21: Uniform over the rectangle initial grid for the random effect distribution of the Nonparametric discrete frailty model. Each point is colored according to a gradient scale from blue to red, which distinguishes points associated with decreased risk (blue) from one associated with increased risk (red).

in hospitalization, reduced risk in death (purple, right-down corner); increased risk both in hospitalization and death (red, left-high corner). The size of all points is the same, as it is proportional to their weights in the discrete distribution.

The algorithm stops when the support is not reduced in the current iteration and the difference between updated and old weights, computed in maximum norm, is less than *epsw*=1e-03 (see Algorithm 3.3). In the first case it takes 21 iterations to converge, while in the second one 19. In the Uniform initialization case, the algorithm converges in slightly less iterations also in other runs performed during the sensitivity analyses in Subsections 4.4.2 and 4.4.3. However, since our implementation relies on R built-in functions and fast optimization routines (see Section 3.4.4 and Appendix C for details), this does not translate in a significant difference of running time needed for the estimation.

Results of the two models are summarized in Table 4.8 and Table 4.9. The coefficients' estimates are almost identical in the two initialization cases, even if some small differences can be noticed looking at the terminal events process coefficients. However, none of the mentioned differences exceeds the second decimal place: the biggest one corresponds to the estimated coefficient for the Adherent covariate in the terminal processes, that is -0.434 in the Gaussian initialization case against -0.407 in the Uniform one. Along with almost identical standard deviation estimates, this suggests a good stability in coefficients estimation with respect to the initialization method. On the contrary, from the point of view of the discrete identified distribution of random effects, the difference due to the two different initializations is quite remarkable. Although both the Gaussian initialization and the Uniform one identify supports of six points and both the distributions are centered in the origin of $\mathbb{R}^2$, in the first case the distribution is more left skewed, while in the second case is more balanced. The interpretations of both estimated coefficients and random-effect points are given in the following.

Variables	Estimate	StdDev	HR	CI95	pvalue
Recurrent Events					
Sex [M]	0.039	0.019	1.039	[1.002,1.078]	0.037
Adherent [1]	-0.262	0.019	0.770	[0.741,0.799]	<2e-16
AgeEvent	-0.015	0.001	0.985	[0.984,0.987]	<2e-16
Comorbidity	0.122	0.005	1.130	[1.118,1.142]	<2e-16
Terminal Event					
Sex [M]	0.179	0.073	1.196	[1.036,1.381]	0.014
Adherent [1]	-0.434	0.078	0.648	[0.556,0.755]	<2e-16
AgeEvent	0.043	0.004	1.044	[1.035,1.052]	<2e-16
Comorbidity	0.449	0.020	1.566	[1.505,1.629]	<2e-16

Frailty	P1	P2	P3	P4	P5	P6
u	-0.294	0.144	0.264	0.387	0.907	0.577
v	-1.296	0.229	1.223	2.177	2.854	3.487
w	0.496	0.213	0.113	0.118	0.014	0.044

Table 4.8: Summary of the Nonparametric Discrete Frailty model specified in Equations 4.9 with Gaussian initialization. For categorical variables, the considered stratum is indicated between brackets. Pvalues are computed using the Wald statistic value, as specified in Subsection 3.4.6. Points of the identified frailty discrete distribution are characterized through thier abscissa (u), ordinata (v) and weight (w).

Variables	Estimate	StdDev	HR	CI95	pvalue
Recurrent Events					
Sex [M]	0.039	0.019	1.039	[1.003,1.079]	0.034
Adherent [1]	-0.259	0.019	0.771	[0.743,0.800]	<2e-16
AgeEvent	-0.015	0.001	0.985	[0.984,0.987]	<2e-16
Comorbidity	0.123	0.005	1.131	[1.119,1.143]	<2e-16
Terminal Event					
Sex [M]	0.169	0.073	1.184	[1.025,1.366]	0.021
Adherent [1]	-0.407	0.078	0.665	[0.571,0.755]	1.7e-07
AgeEvent	0.039	0.004	1.041	[1.032,1.049]	<2e-16
Comorbidity	0.429	0.020	1.535	[1.476,1.597]	<2e-16

Frailty	P1	P2	P3	P4	P5	P6
u	-0.466	-0.194	0.079	0.231	0.468	0.679
v	-1.872	-0.859	-0.090	1.166	2.277	3.088
w	0.208	0.234	0.206	0.217	0.071	0.063

Table 4.9: Summary of the Nonparametric Discrete Frailty model specified in Equations 4.9 with Uniform initialization. For categorical variables, the considered stratum is indicated between brackets. Pvalues are computed using the Wald statistic value, as specified in Subsection 3.4.6. Points of the identified frailty discrete distribution are characterized through their abscissa (u), ordinata (v) and weight (w).

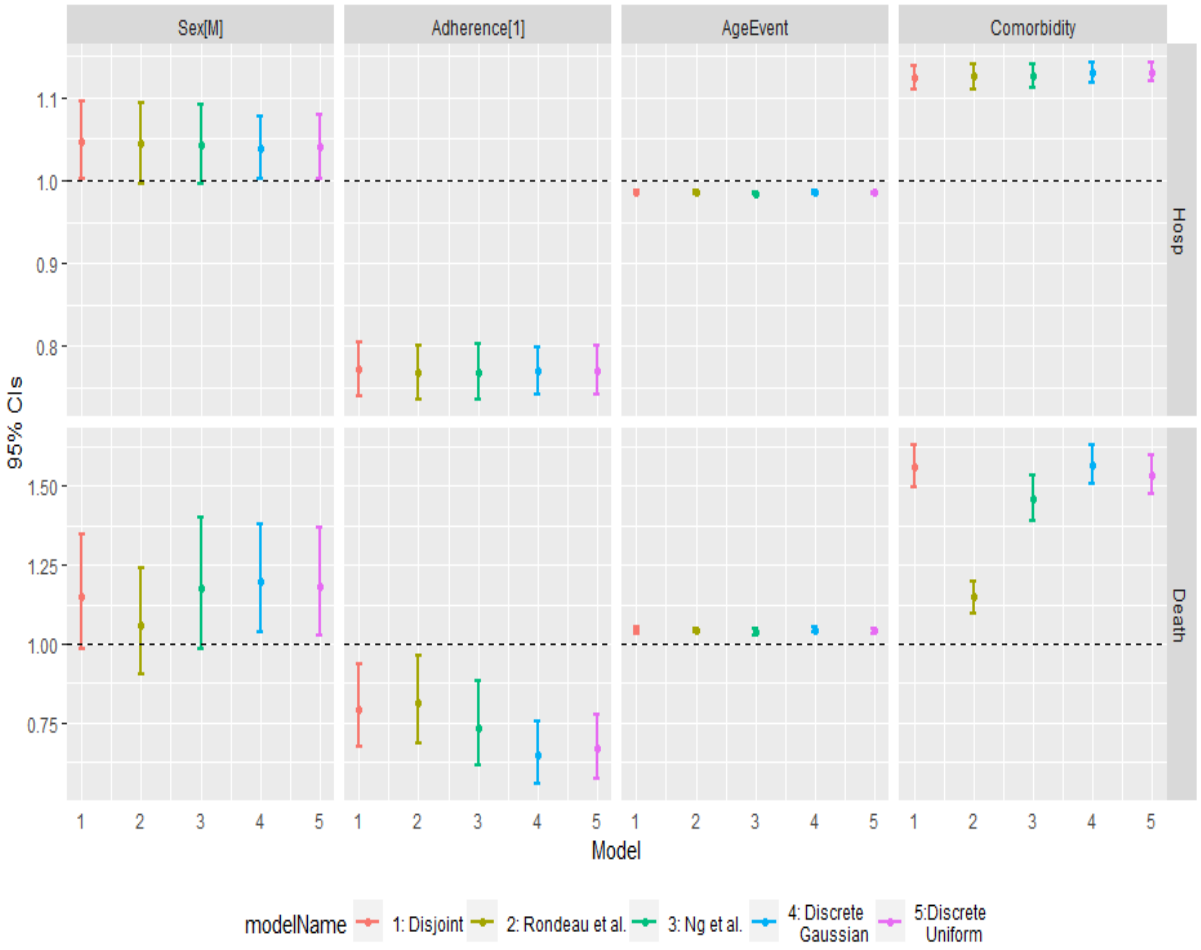

Figure 4.22: Comparison of estimated Hazard Ratios (HRs) and their 95% CI in the trained models. Considered models are: disjoint model in Equations 4.2 and 4.3 (pink); Rondeau et al. model in Equation 4.5 (pistachio green); Ng et al.model in Equation 4.7 (teal); Discrete Nonparametric Frailty model with Gaussian Initialization (light blue) and Discrete Nonparametric Frailty model with Uniform Initialization (purple) (Equation 4.9). Top panels refer to the recurrent events process, whereas bottom panels to terminal event process.

From Figure 4.22 we can note that the estimated hazard ratios (light blue: Discrete Gaussian; purple: Discrete Uniform) are quite concordant with the estimates obtained in the disjoint model (pink) and using the approach by Ng et al. (teal), and are not subject to instability issues present in the model trained through Rondeau et al.'s approach (pistachio green). Some differences with previous models results can be found, although small, and also in this case are concentrated on the terminal events process part (bottom panels).

In both cases the covariate **Sex** results to be statistically significant at 5% for both processes (pvalues, respectively, of 0.037 for hospitalization and 0.014 for death in the Gaussian case, and of 0.034 for hospitalization and 0.021 for death in the Uniform case). As usual, looking at the hazard ratio, male subjects are suggested to be slightly more prone to risk of hospitalization (Gaussian: HR=1.039; Uniform: HR=1.039) and death (HR=1.196, Gaussian case, and HR=1.184, Uniform case). As testified by the panels in the first column of Figure 4.22, the obtained values are compliant with the other models' ones.

The covariate **Adherent** results to be statistically significant at any level for the two processes, in both cases. According to the Gaussian initialization model, being adherent to the treatment yields a 23.0% decrease in the risk of a new hospitalization (HR=0.770) and a 35.2% decrease in the risk of death. According to the Uniform initialization model, instead, being adherent to the treatment yields a 22.9% decrease in the risk of a new hospitalization (HR=0.771) and a 33.5% decrease in the risk of death. From a clinical point of view, such results finally endorse the efficacy of the ACE inhibitors treatment for heart failure, as it lead to a significant reduction of the hospitalizations rate (and thus of critical HF events) of adherent patients during their clinical path, in addition to increasing their survival probability. From the two corresponding panels in the second column of Figure 4.22, we can see that the estimated values for the hospitalization process coefficient are very similar to the previous models ones, while the terminal event process coefficient estimates are slightly smaller with respect to previous estimates. However, the overlapping of 95% confidence intervals suggest that such difference is not statistically significant.

The covariate **AgeEvent** results, for both processes in both the trained models, to be statistically significant at all levels. Its effect in the Gaussian initialization model on the hospitalization hazard is a 1.5% (1.5% also in the Uniform initialization model) reduction of the risk of hospitalization per year (HR=0.985), while on the death hazard it yields an increase of the risk of the 4.4% (4.1% in the Uniform initialization model). Clinically speaking this can be explained by the fact that part of the risk of experiencing a new hospitalization is replaced by the risk to die when patients get older. This is reasonable especially in our case, as we are considering elderly persons (median age at first hospitalization of 74, IQR=[66;80]), but is likely to be different when young subjects are involved. Pointwise estimates are close to previous models ones, as well as their 95% (very narrow) confidence intervals (third column panels in Figure 4.22).

The covariate **Comorbidity** results to be in both models statistically significant for both the processes. It yields, in the Gaussian case, an increase of 13.0% in the risk of hospitalization per comorbidity registered (13.1% in the Uniform case) and a very high increase of 56.6% in the risk of death per comorbidity registered (53.5% in the Uniform case). The role comorbidities have in increasing the mortality and hospitalizations of heart failure patients is well documented in medical literature [61] and confirmed also by our analysis. In particular, patients show very low survival probabilities in presence of multiple comorbidities, which can be possibly explained by the fact that, among the considered ones (see Table 2.3), there are illness of assessed severity. For example, renal insufficieny, which is considered one of the most powerful predictor of mortality in heart failure patients [62], affects the 27.6% of patients in our cohort. The recurrent events process coefficient estimates are very close to the previous models ones (top right panel), while the terminal events process coefficient ones are subject to a little more variability (bottom right panel). However, in both the Gaussian and the Uniform frailty models, the obtained values are coherent with the disjoint and Ng et al. models (avoiding the issues involved using the approach by Rondeau et al.).

Figures 4.23 and 4.24 display the identified discrete distributions of random effects in terms of points in $\mathbb{R}^2$ for the Gaussian and the Uniform initializations, respectively. Each point is colored according to a gradient scale from blue (associated with decreased risk) to red (associated with increased risk) and is sized according to the probability of a patient to belong to the corresponding latent population. In both cases, the support of the identified distributions is composed by six points, whose disposition follows a diagonal shape, as expected considering the results of Ng et al. model (see Figure 4.19). This fact is important, as it actually supports the correctness of our custom implementation of Ng et al. model estimation routine and of the effectiveness of our new nonparametric frailty extension.

In the Gaussian initialization case (see Table 4.8 and Figure 4.23), we notice that the latent population with the highest probability is **P1** (i.e. *Protected* subpopulation) with a probability of about 49.6%, which is linked to a protective random effect ($u = -0.294$, $v = -1.296$), both from the new hospitalization risk and death point of view. Point **P2** (i.e. *Neutral* subpopulation), associated with a probability of 21.3%, identifies a subpopulation almost neutral to added random risk for both processes ($u = 0.144$, $v = 0.229$). Point **P3** (i.e. *Slightly at Risk* subpopulation) identifies a subpopulation associated with

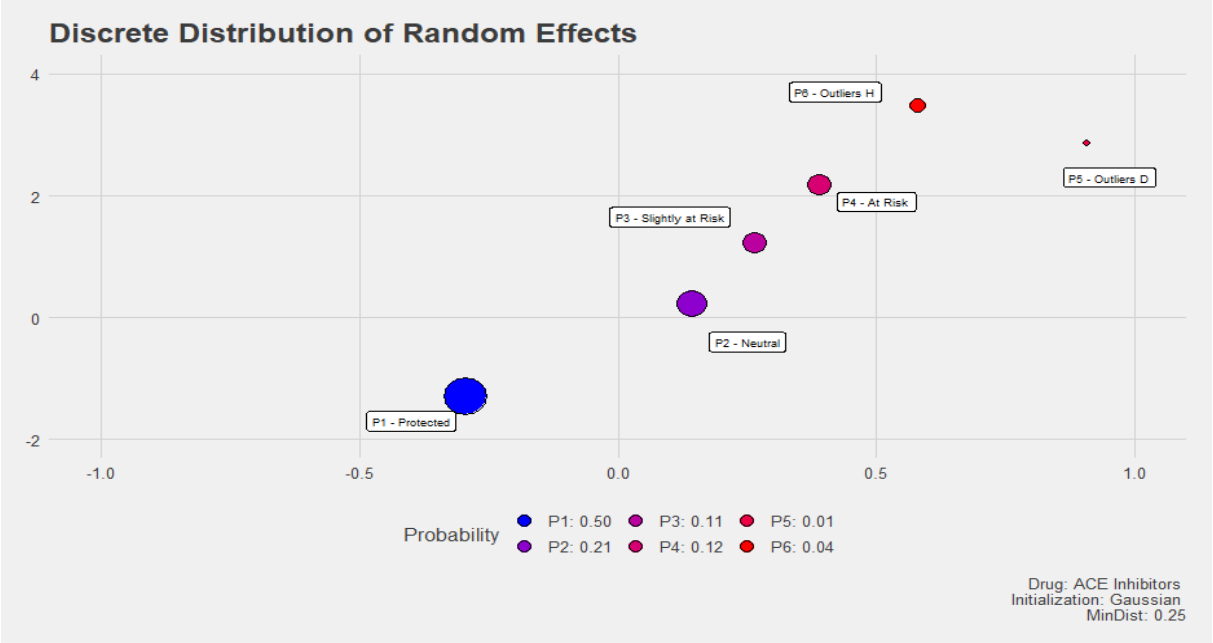

Figure 4.23: Identified Discrete Distribution of Random Effects in the Gaussian Initialization case. The color of points ranges from blue (strong subjects) to red (weak subjects), while their size is representative of the probability of a patient to belong to the corresponding latent population.

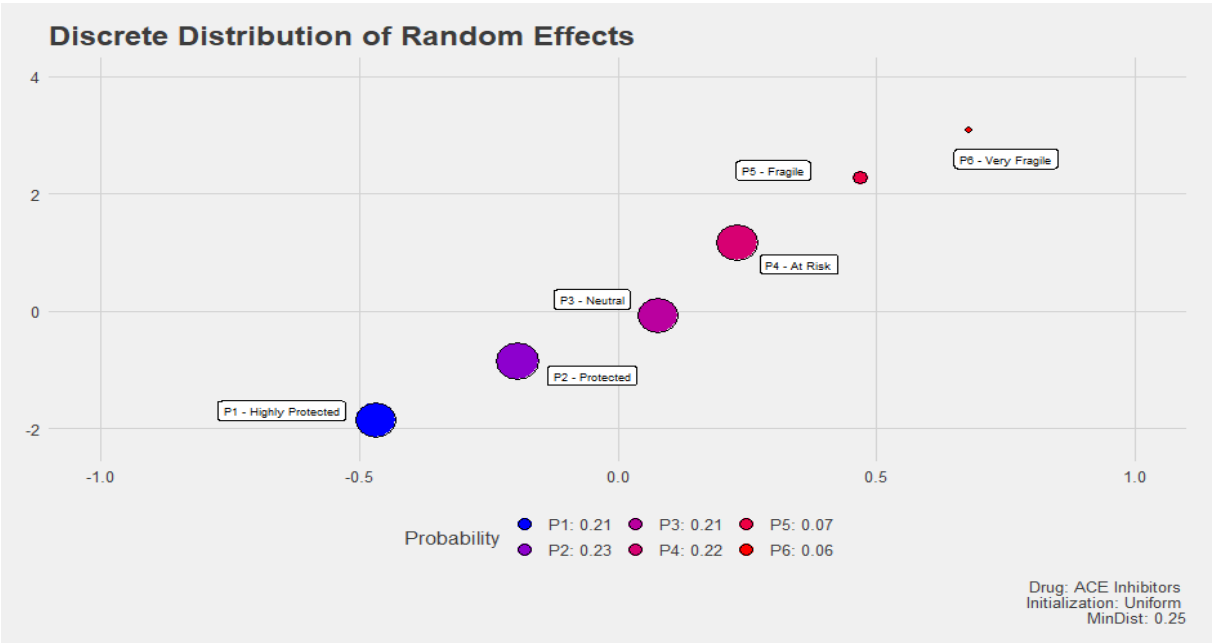

Figure 4.24: Identified Discrete Distribution of Random Effects in the Uniform over a rectangle Initialization case. The color of points ranges from blue (strong subjects) to red (weak subjects), while their size is representative of the probability of a patient to belong to the corresponding latent population.

a 11.3% probability which is slightly more prone to risk for both processes ($u = 0.264$, $v = 1.223$), while point **P4** (i.e. *At Risk* subpopulation) is composed by fragile patients ($u = 0.387$, $v = 2.177$), but still representative of a significant proportion of patients (about 11.8%). Point **P5** (i.e. *Outliers H*) and **P6** (i.e. *Outliers D*) are related to outlying subjects (probability, respectively, of 1.4% and 4.4%) which are extremely fragile. The two distinguish themselves as the former is composed by subjects particularly prone to the risk of a new hospitalization ($u = 0.907$, $v = 2.854$), while the latter by subjects particularly prone to the risk of death ($u = 0.577$, $v = 3.487$).

In the Uniform over a rectangle initialization model, the discovered discrete distribution of random effects support consists in six points, disposed, as in the previous case, in a diagonal pattern (see Table 4.9 and Figure 4.24). The points are however more evenly spaced along the diagonal with respect to the Gaussian initialization case, and are characterized by a more balanced distribution (but still left skewed). Point **P1** ($u = -0.466$, $v = -1.872$) and point **P2** ($u = -0.194$, $v = -0.859$), associated, respectively, with a probability of 20.8% and 23.4%, identify *Highly Protected* and *Protected* subpopulations, respectively. Point **P3** ($u = 0.079$, $v = -0.090$) is related to a subpopulation *Neutral* to random effects, with 20.6% probability for a patient to belong to it. Point **P4** identifies instead a relevant group of patients slightly more prone to the risk of a new hospitalization and death ($u = 0.231$, $v = 1.166$), with a probability of 21.7% (*At Risk* subpopulation). Points **P5** ($u = 0.468$, $v = 2.277$) and **P6** ($u = 0.679$, $v = 3.088$) identify two subpopulations of outliers composed by *Fragile* and *Very Fragile* patients particularly prone to the risk of death, associated with smaller probabilities (respectively, 7.1% and 6.3%).

In general, to quantify the actual effect on survival probabilities (for both the terminal and recurrent events processes) due to the identified discrete distribution of random effects, we look at the induced stratification of baseline survival curves. This comes from a different parametrization of the random effects in Equation 4.9, in which they are encoded as multiplicative factors of the baseline hazards of the two processes. As their estimated distributions are finite, such formulation yields, for each of the two processes, the following mixture of Cox models

$$\begin{aligned} h_i^R(t|\boldsymbol{x}_i R(t)) &= \sum_{l=1}^{L} w_l h_0^R(t) \exp\{u_l\} \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i^R(t)\} \\ &= \sum_{l=1}^{L} w_l h_{0l}^R(t) \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i^R(t)\} \end{aligned} \tag{4.12}$$

$$h_i^D(t|\boldsymbol{x}_i^D(t)) = \sum_{l=1}^{L} w_l h_0^D(t) \exp\{v_l\} \exp\{\boldsymbol{\gamma}^T \boldsymbol{x}_i^D(t)\}$$
$$= \sum_{l=1}^{L} w_l h_{0l}^D(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{x}_i^D(t)\}.$$

Each process model is characterized by composing Cox models which share the same coefficients (i.e. in which the effect of the chosen set of covariates is the same for all components), but have different baseline hazards, obtained multiplying the common baselines of Equation 4.9 by the exponentiated corresponding random effect of the identified discrete distribution. Such formulation is appealing since it allows to visualize, compare and quantify the actual influence of random effects on the survival probability function, composed by a weighted set of functions, obtained from the stratified hazards through Equation 3.20.

Figure 4.25 and Figure 4.26 reports the obtained set of Survival probability curves for the hospitalization and death processes, respectively, in the case of Gaussian Initialization, while Figure 4.27 and Figure 4.28 their Uniform initialization counterparts. Time is expressed in days and each curve is colored according to the gradient scale from blue to red (associated with the risk related to the latent population random effect) also adopted in Figures 4.23 and 4.24. In all the four plots, the white curve represents the corresponding process baseline survival probability function estimated in absence of random effects. Notice that it is reported with the purpose of facilitating the quantification of the random effect influence on the baseline survival curves, but it does not represent an actual latent class of subjects identified by the algorithm. On one hand, the stratified baselines for the hospitalization process shows more fragile subpopulations are linked to curves with gradually steeper slopes (purple-red curves), coherently with the fact that their expected time before a new hospitalization is smaller. The difference between the hospitalization-related survival function of *Protected* subpopulation (point P1 - blue curve) and the others is quite evident in the case of Gaussian Initialization (see Figure 4.25), while in the Uniform Initialization model the presence of a more balanced distribution makes it lighter (see Figure 4.27). However, in both cases the overall difference among subpopulations in terms of survival with respect to a new hospitalization event is significant: in the Gaussian initialization model the median new hospitalization time for *Protected* (point **P1**) subpopulation baseline is 218 days, while for *Outlier H* subpopulation (point **P5**) is only 34 days; in the Uniform initialization model the median new hospitalization time for *Protected* (point **P1**) subpopulation baseline is 290 days, while for *Very Fragile* (point **P6**) subpopulation baseline is only 45 days.

On the other hand, the difference between baseline survival curves for the terminal event stratified by latent population is much more evident, in both the trained models. In the Gaussian Initialization case (see Figure 4.26) we notice that subpopulations *Protected* (point **P1**) and *Slightly at Risk* (**P3**) show different, almost symmetrical curves with respect to the terminal events process baseline survival without random effects, which is mimicked by subpopulation *Neutral* (**P2**). Subpopulation *At Risk* (**P4**) is associated to a steeper curve, as well as subpopulations *Outliers D* and *Outliers H* (point **P6** and **P5**, respectively), which have much lower survival chances with respect to the other subpopulations. In the Uniform Initialization model (Figure 4.28), curves related to to *Highly Protected* (**P1**), *Protected* (**P2**), *Neutral* (**P3**) and *At Risk* (**P4**) identify very different profiles in terms of survival, but are evenly spaced. The two remaining subpopulations (*Fragile* **P5** and *Very Fragile* **P6**), present two very steep curves and are thus likely to depart within a short time (median survival time of 83 and 28 days, respectively), showing a similar pattern as in the Gaussian case.

The described plots are useful to visualize differences among the discovered latent classes of patients, however we must recall that, as the final model can be seen as a mixture of Cox models, each curve is associated to a probability which specifies its frequency in the considered cohort. For this reason, it is useful to complement any analysis suggesting a comprehensive metric, which summarizes all the information provided by the model. In our case, when dealing with a specific patient, a good choice is to compute the expected survival time, with respect to each one of the two considered processes, which can be obtained simply as a weighted average of the expected survival times related to the stratified Cox models mentioned in Equations 4.12.

In this paragraph we have detailed a possible pipeline to analyze the results obtained when fitting our model, distinguishing between the two proposed initializations. We can conclude that the initialization method surely influences the identified discrete distribution of random effects, while its effect on the fixed-effect coefficients is negligible. Both methods well suit the problem (i.e. the identified discrete distributions of random effects are in both cases plausible), and thus we will evaluate both of them also in our further analysis. However, as mentioned in Section 3.4.5, the Uniform initialization should be preferred, as it is less informative.

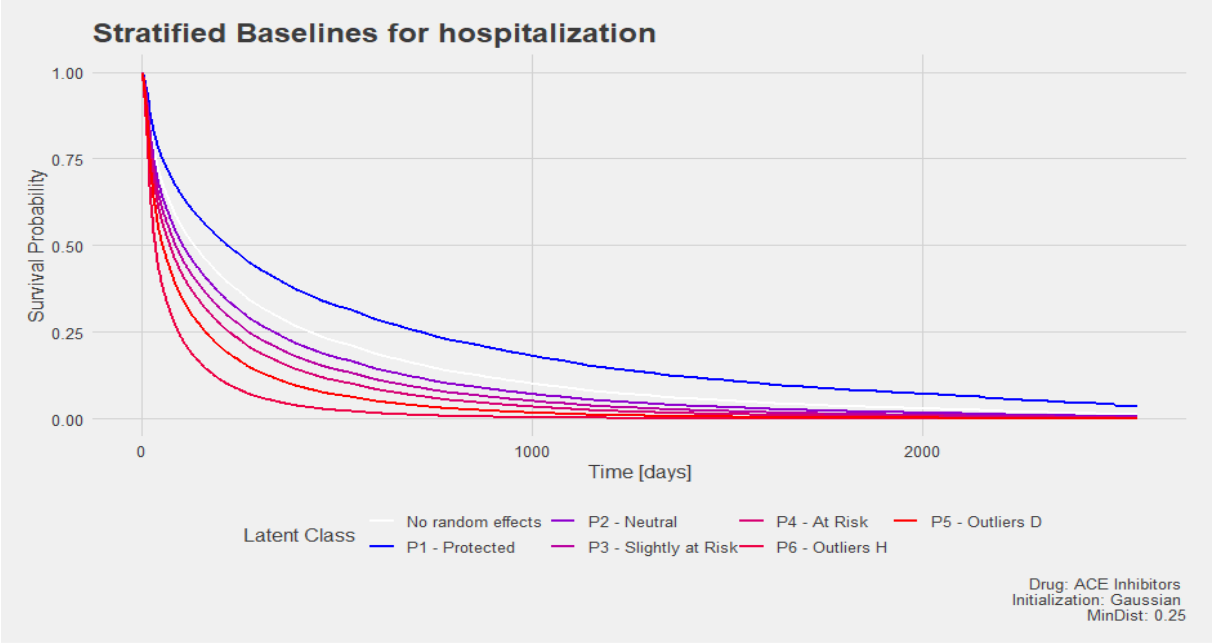

Figure 4.25: Stratified Survival Probability Baseline curves of the hospitalization process, associated to the discrete distribution of random effects identified in the Gaussian Initialization model. The color of each curve is the same of the corresponding random effect point as in Figure 4.23. The white line represents the survival probability baseline curve of the model without random effects.

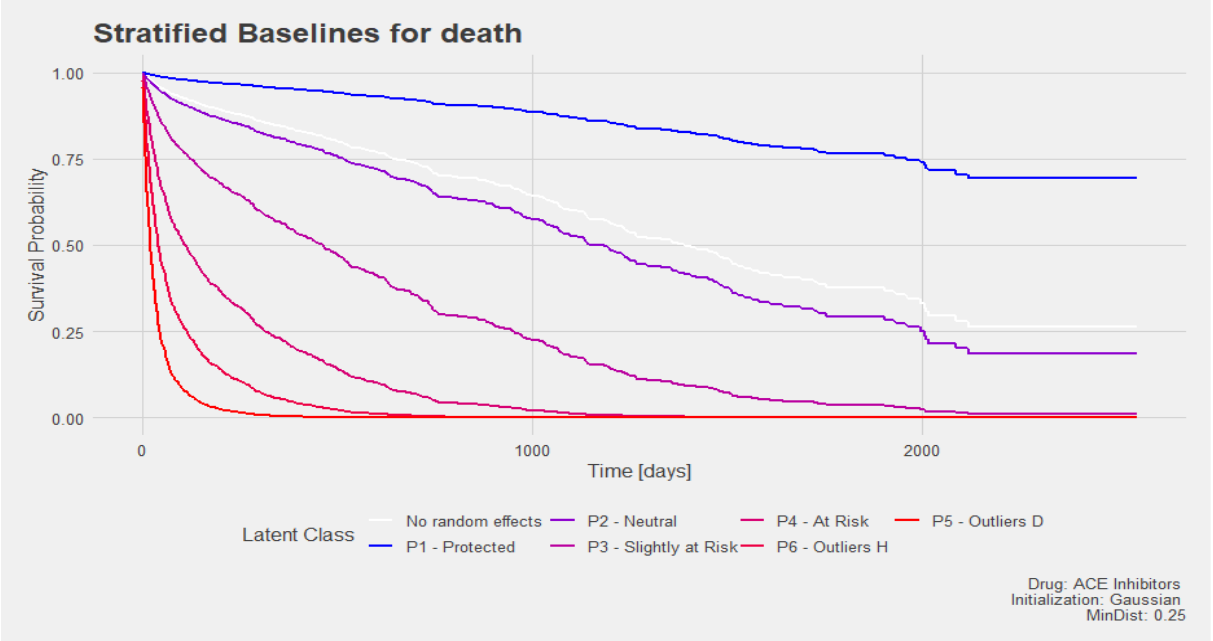

Figure 4.26: Stratified Survival Probability Baseline curves of the terminal event process, associated to the discrete distribution of random effects identified in the Gaussian Initialization model. The color of each curve is the same of the corresponding random effect point as in Figure 4.23. The white line represents the survival probability baseline curve of the model without random effects.

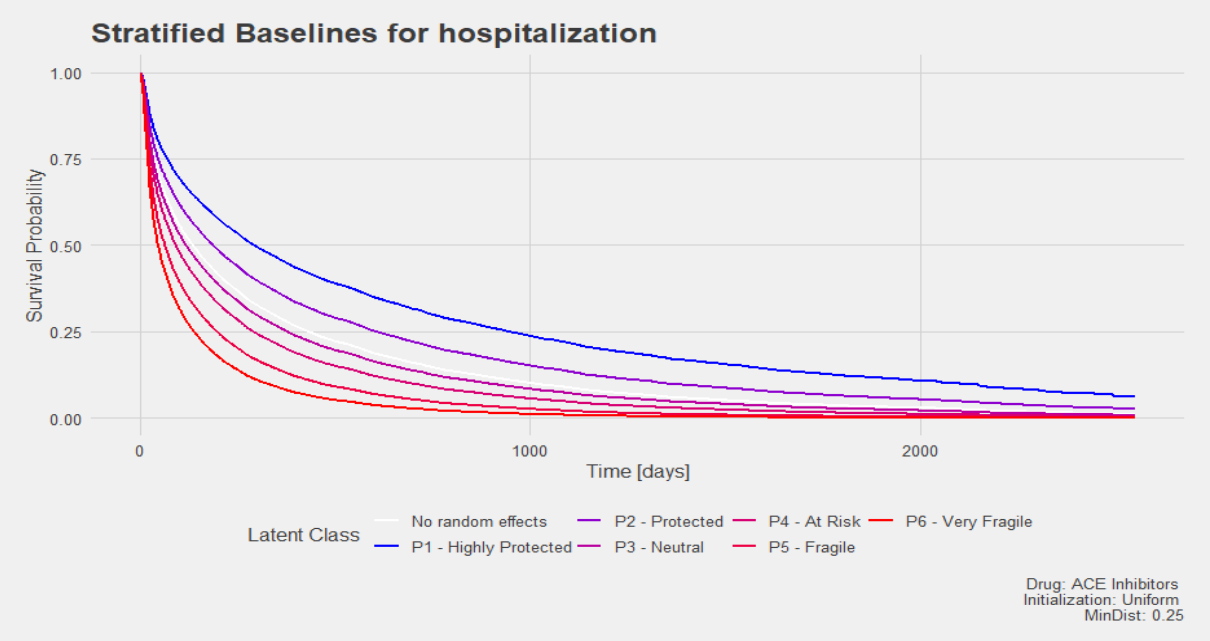

Figure 4.27: Stratified Survival Probability Baseline curves of the hospitalization process, associated to the discrete distribution of random effects identified in the Uniform Initialization model. The color of each curve is the same of the corresponding random effect point as in Figure 4.24. The white line represents the survival probability baseline curve of the model without random effects.

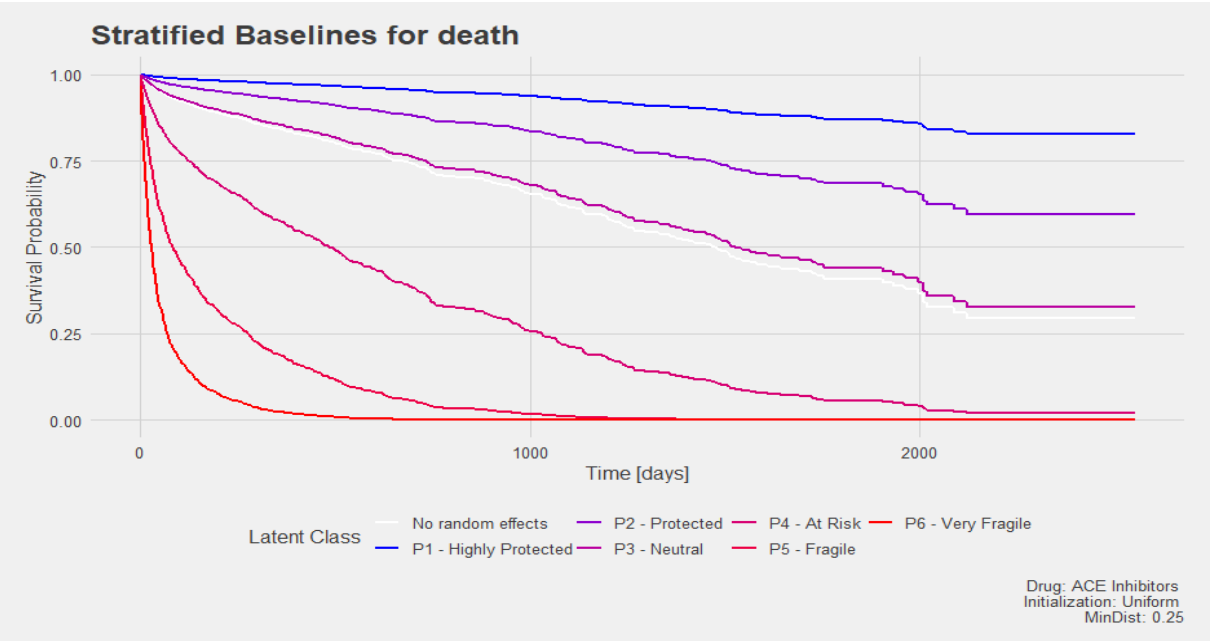

Figure 4.28: Stratified Survival Probability Baseline curves of the terminal event process, associated to the discrete distribution of random effects identified in the Uniform Initialization model. The color of each curve is the same of the corresponding random effect point as in Figure 4.24. The white line represents the survival probability baseline curve of the model without random effects.

4.4.2. Sensitivity Analysis: Choice of *MinDist*

Once we have studied the influence of the different initialization methods on the final discrete distribution identified by the model, we investigate the change in the *MinDist* threshold, which turns out to be the main factor influencing the final distribution discovered. As mentioned, when choosing the *MinDist* value we have to keep in mind that it set the distance between two significantly different points in the random effects distribution, specifying how much we want to be sensitive to diversification in the latent classification of patients (i.e. the bigger the value of *MinDist*, the less we diversify). However, either the problem considered is well known and the available information about the underlying phenomenon can be exploited to define a proper value, or an a priori choice may result very difficult. Moreover, the fact that *MinDist* so strongly influences the final results of the model suggests the necessity to design a rigorous procedure to select it. This represents a complex problem involving several issues, such as the definition of a proper metric to measure the goodness of a found distribution with respect to another, the implementation of an efficient cross validation scheme or the definition of a criterion to identify the candidate values. This would require a separate study and is beyond the scope of this thesis.
Nevertheless, to take a first step in this direction, we perform a sensitivity analysis. In particular, we look at the dependence between the parameter to be tuned and a simple fitting criterion, the Akaike Information Criterion (in the following, AIC) [9], in order to identify promising candidates. We reformulate the AIC as

$$\mathrm{AIC}(d) = 3 * \mathrm{card}(\boldsymbol{w}(d)) - \mathrm{LogL}(\mathcal{M}(d)) \tag{4.13}$$

where d is the current value of *MinDist*, card refers to the number of components of $\boldsymbol{w}$, $w(d)$ is the vector of weights of the model identified using d and $\mathrm{LogL}(\mathcal{M}(d))$ is the same model loglikelihood. The penalization term consists in the cardinality of w multiplied by three as that is the total number of parameters related to the final discrete distribution. First, we define a set of possible candidates for the *MinDist* parameter consisting in 37 values, ranging from 0.1 to 1, which are apart from each other of 0.025. Then, we evaluate the defined AIC for each one of the model identified using these values and plot the resulting curve, trying the spot trends or patterns useful to select the best values. We focus solely on the identified discrete distributions of random effects, since we confirm that the choice of *MinDist* does not affect coefficients' estimation. As in the previous section, the analysis is carried out both in the case of Gaussian initialization and of Uniform over a rectangle initialization.

Following the described procedure, in the Gaussian Initialization case we obtain the AIC curve reported in Figure 4.29. In the plot we highlight different regions of interest: the red rectangle encloses a region where the AIC shows a decreasing trend and high variability, reasons why we classify it as a *transition region*, discarding the related threshold values from the list of plausible candidates. On the contrary, the three shaded boxes identify different regions in which the AIC stays almost constant at different levels, which we classify as *stability regions*. Although the general rule when using the AIC to compare models is to try to minimize it, in our case this approach may result too restrictive. For this reason, we decided to analyze the distributions of random effects identified by models tuned with threshold values from the stability regions. We notice that in such regions stability of the AIC corresponds to stability of the identified final discrete distribution (see first part of Appendix D). Figure 4.30 shows the distributions related to median values of the *MinDist* parameter for each region, i.e 0.55 in region *Stable[1]*, 0.75 in region *Stable[2]* and 0.925 in region *Stable[3]*. For $MinDist = 0.55$ (top panel), the model identifies a three points, well balanced distribution. The suggested latent partition comprehends a *quite protected* subpopulation, a *neutral* to random risk subpopulation and aa *At Risk* subpopulation, with probabilities 46.5%, 36.6% and 16.9% respectively. $MinDist = 0.75$ (central panel) yields a distribution composed by only two points, which partition the cohort in *Stronger* against *Weaker* subjects. The associated random effects yield less extreme effects on the survival probabilities with respect to the previous one and the distribution is still left skewed (probabilities of 76.9% and 23.1% respectively). Finally, $MinDist = 0.925$ (bottom panel) yields probably the most promising result from an interpretative point of view, as it identifies a three point skewed distribution, where two subpopulations mimic the distribution related to $MinDist = 0.75$ (*Stronger* and *Weaker* subpopulations with probabilities of 74.8% and 21.9%), while the third is composed by *Extremely weak* patients. Basically, the most fragile patients belonging to *At Risk* subpopulation of the *Stable[2]* model, who contributed to move the relative random effect point to more extreme values, are identified as outliers (probability of 3.2%) and assigned to a third specific point.

Applying the same procedure to the Uniform Initialization case we obtained the AIC curve reported in Figure 4.31. First thing to notice is that this curve shows a little higher amount of noise and a more regular decreasing trend with respect to the Gaussian initialization case one. Following the same approach we identify some interesting patterns in the curve. The region enclosed in the first red box shows a steep descent of the AIC,

similarly to the transition region in Figure 4.29. The following dashed red box identifies a non-decreasing AIC region which is not stable due to the high amount of variability. The curve reaches an elbow point localized at $MinDist = 0.50$ (purple box) after which the AIC drops again (second red box). Finally, the curve achieves show a sort of stability region, with a minimum point localized at $MinDist = 0.875$ (gray region named as *Stable*), but still presents variability. In Figure 4.32 are reported the two discovered distribution of random effects relative to threshold values identified by the *elbow* (top panel) and the minimum in the *Stable* region (bottom panel)(See the second part of Appendix D for the identified distributions in each classified region in Figure 4.31). In the first panel, representing the distribution corresponding to the elbow value of 0.50, we notice a five points support, with: two outlying subpopulations, one of *Very Strong* (small blue point, probability of 11.1%) and one of *Very Fragile* (small red point, probability of 9.0%) subjects; a relevant subpopulation of significantly *Fragile* patients (magenta point, probability of 24.7%); the highest probability subpopulation, related to *Protected* patients (purple point, probability of 30.1%); a relevant subpopulation related to patients *Neutral* to added random risk (light purple point, probability of 25.1%). The distribution appears to be balanced and spots two interesting outlying classes. The second panel, instead, shows the distribution identified by the model tuned with the threshold value of 0.875, corresponding to the minimum achieved AIC. Its shape is very similar to the one of the distribution found in the Gaussian case for values in the third stability region, identifying a main *Protected* subpopulation (purple point, probability of 54.9%) and another relevant *At Risk* subpopulation (magenta point, probability of 36%), in addition to an outlier point representative of *Very Fragile* patients (small red point, probability of 9%).

The results of this analysis confirm the high influence of the *MinDist* parameter on the discrete distribution of random effects yielded by our model. We can however notice some recurrent features, such as left skewness of distributions and presence of outliers which are characterized by very higher risk rates with respect to the overall population. As this characteristics are well summarized by the distribution relative to $MinDist = 0.875$ in the Uniform initialization model (bottom panel of Figure 4.32), in addition to the fact that the corresponding AIC curve (see Figure 4.31) achieves a minimum for such value, we can suggest it as the most plausible candidate.

We investigate the effect of the *MinDist* value choice on the discovered discrete distributions, suggesting a strategy to identify plausible candidates. In the next Subsection we focus on the sensitivity to randomization of the two different initialization models.

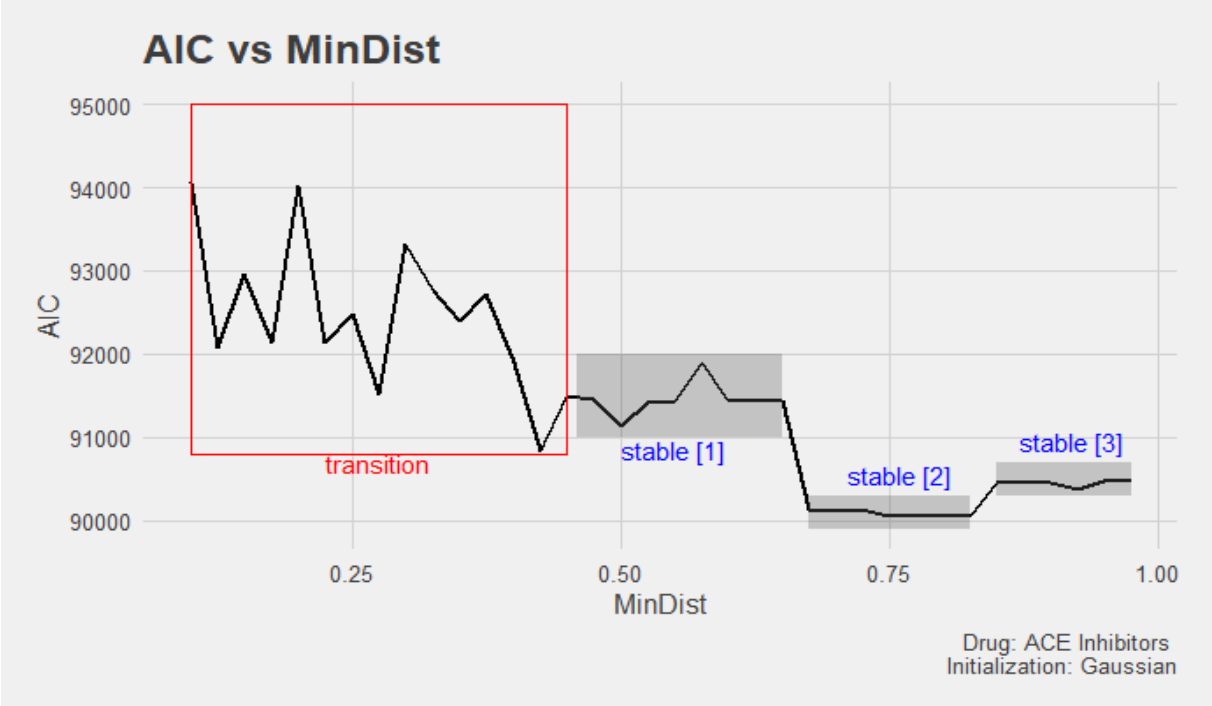

Figure 4.29: AIC curve as function of the *MinDist* parameter, computed through the Gaussian initialization discrete nonparametric frailty model.

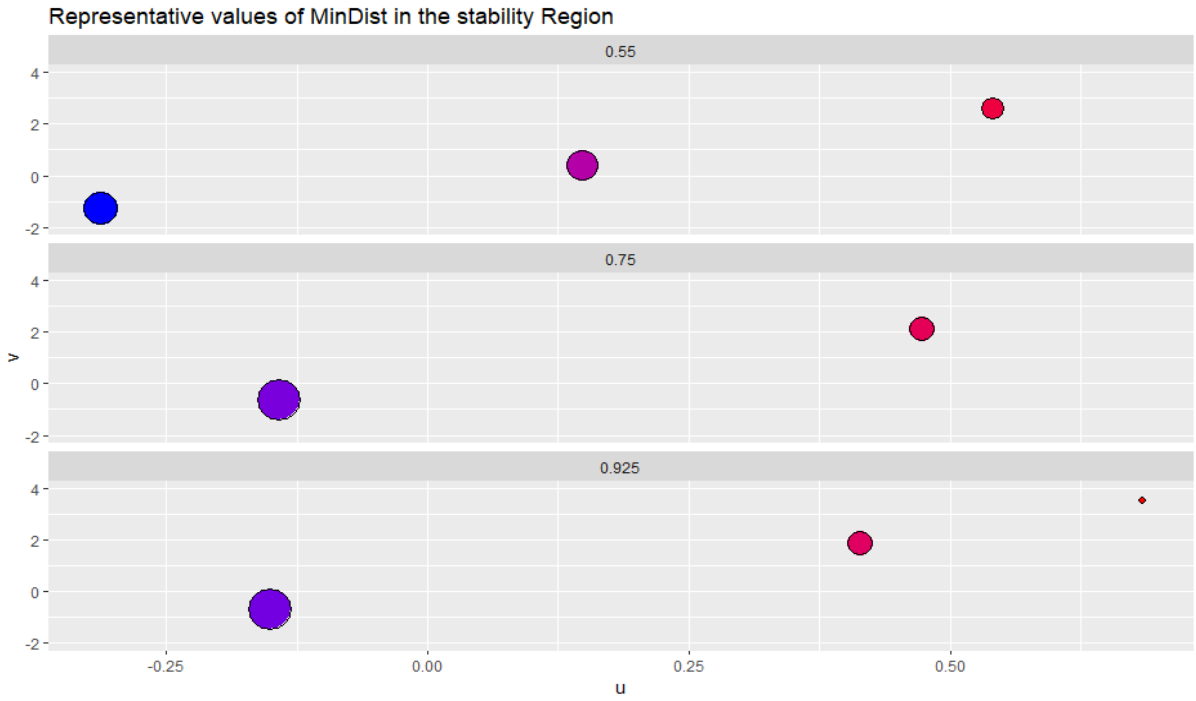

Figure 4.30: Discrete distribution of random effects related to Gaussian Initialization models tuned with values of *MinDist* representative of the stability regions of Figure 4.29. Top panel refers to region *Stable[1]* with a median *MinDist* of 0.55. Central panel refers to region *Stable[2]* with a median *MinDist* of 0.75. Bottom panel refers to region *Stable[3]* with a median *MinDist* of 0.925. The color of points ranges from blue (strong subjects) to red (weak subjects), while their size is representative of the probability of a patient to belong to the corresponding latent population.

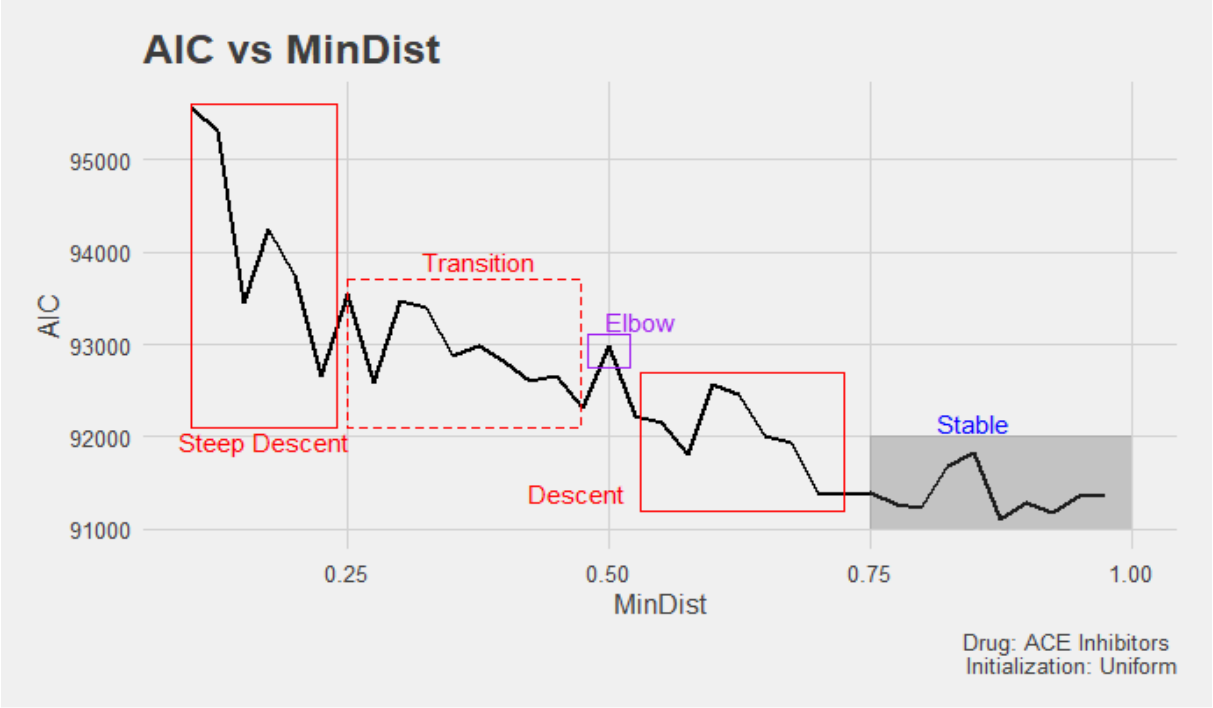

Figure 4.31: AIC curve as function of the *MinDist* parameter, computed through the Uniform initialization discrete nonparametric frailty model.

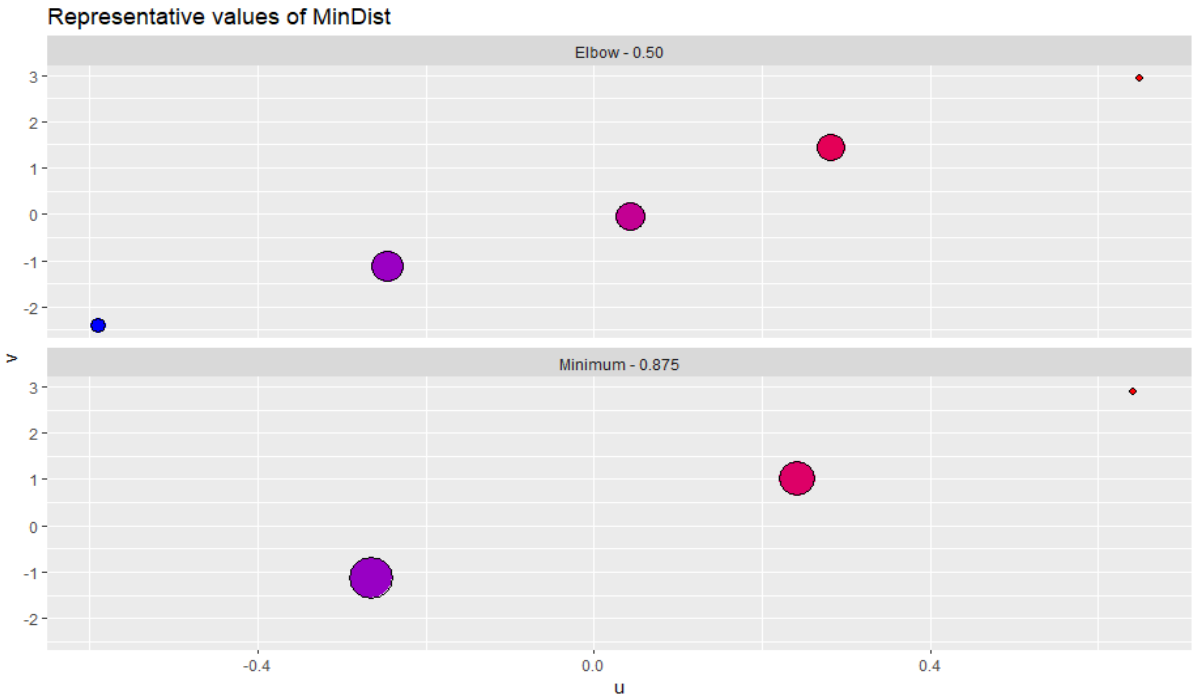

Figure 4.32: Discrete distribution of random effects related to Gaussian Initialization models tuned with relevant values of *MinDist* spotted analyzing Figure 4.31. In particular, the values considered are the elbow localized at $MinDist = 0.50$ (top panel) and the value minimizing the AIC in the Stable region $MinDist = 0.875$ (bottom panel). The color of points ranges from blue (strong subjects) to red (weak subjects), while their size is representative of the probability of a patient to belong to the corresponding latent population.

4.4.3. Sensitivity Analysis: Randomization

In this final paragraph, we present the results of a sensitivity analysis performed on our model and on its estimation algorithm, with respect to randomization. Actually, the following steps in our procedure relies on randomization:

- random allocation of discrete frailties (composing the initial grid) to patients in order to define the initial estimates for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $H_0^R(t)$ and $H_0^D(t)$;
- in the Gaussian Initialization case, the sampling of points to define the initial grid.
- elimination of support points which are not assigned to any subject in each iteration.

Thus, it is reasonable to assess the stability of our estimation algorithm, as well as to quantify the variability of the results due to randomicity. To this end, we repeat the sensitivity analysis performed on the *MinDist* parameter considering 10 different random seeds and we compare the different obtained AIC curves (as function of the parameter *MinDist*). As before, we perform the analysis in the two different initialization cases; in particular, we expect the Gaussian initialization model to be more sensible to randomicity, since the initial points of the support are casually sampled.

Figure 4.33 reports the AIC curves for the Gaussian case, while Figure 4.34 reports the ones for the Uniform Initialization model. In both cases the different curves share a similar shape, suggesting overall stability of the estimation. However, in the first case the variability is quite high, while in the second one the estimated curves are almost overlying, with a much less relevant degree of noise. For this reason, we can argue that the designed algorithm is sensitive to randomicity mostly due to the grid initialization of the Gaussian case. This represents an important drawback of the Gaussian initialization model, since it may invalidate the analysis carried out to identify good candidates for the *MinDist* parameter. Actually, investigating the shape of the AIC curve (i.e. looking for elbows, stability regions, etc) would be a lot more difficult due to the relevant masking effect caused by variability due to randomization. To avoid this problem, a solution may be represented by computing a large sample of the mentioned AIC curves (with different random seeds) to obtain a Monte Carlo estimate. However, this would require a significant increase of the running time and computational power needed for the estimation. For these reasons, adopting the Uniform over a rectangle model seems to be a more robust choice.

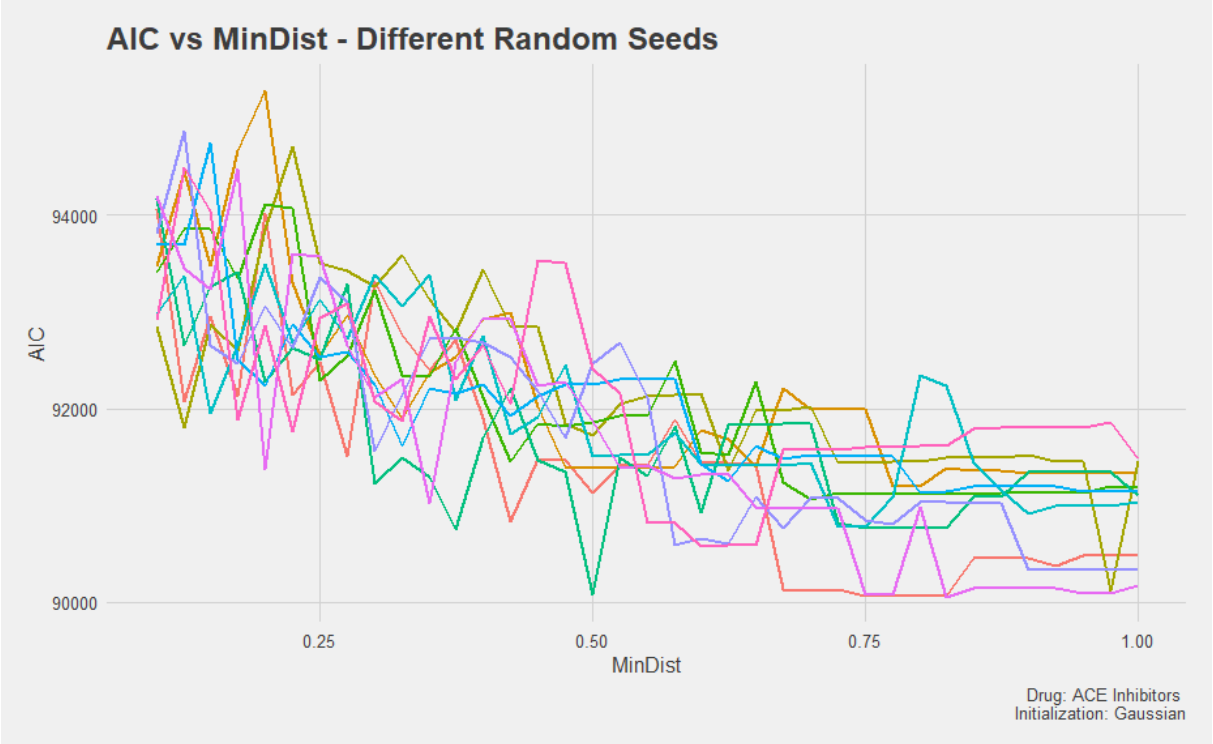

Figure 4.33: AIC curves as function of the *MinDist* parameter, computed through the Gaussian initialization discrete nonparametric frailty model with 10 different random seeds. Different colors refer to different random seeds.

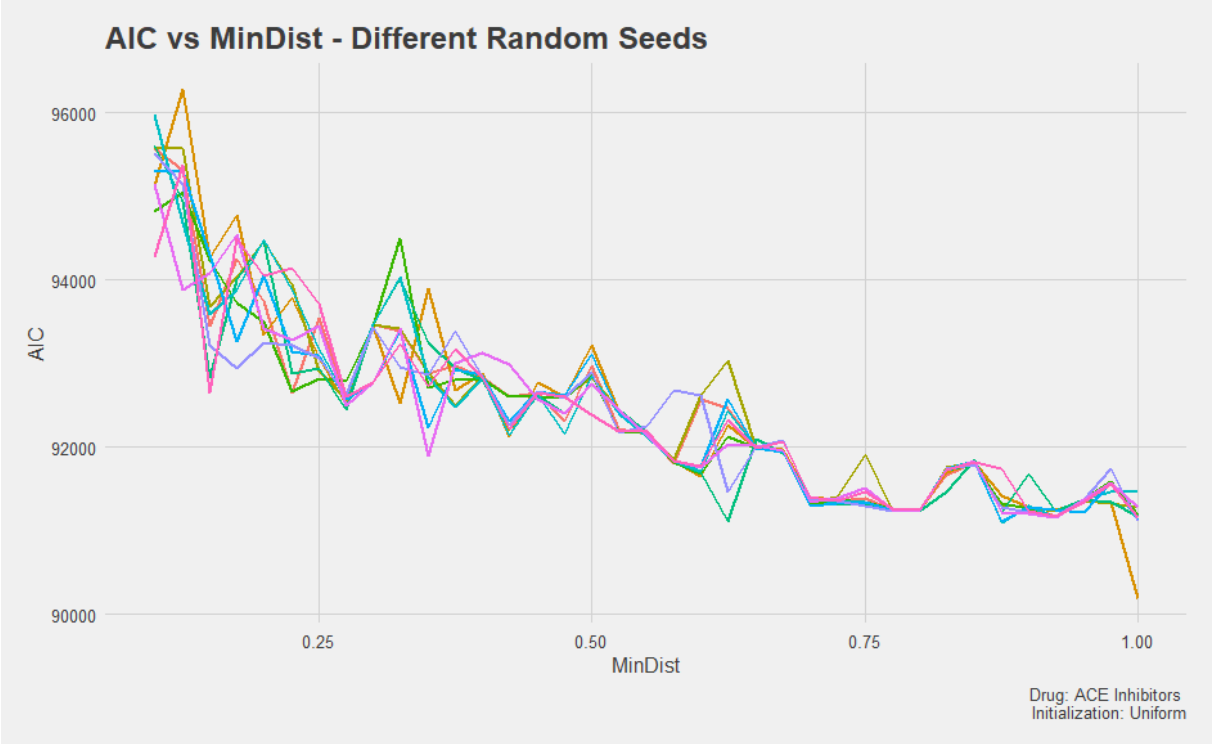

Figure 4.34: AIC curves as function of the *MinDist* parameter, computed through the Uniform initialization discrete nonparametric frailty model with 10 different random seeds. Different colors refer to different random seeds.

Given the results of the analyses performed on our innovative model, we can conclude that the best setting consists in a Uniform initialization model, as it is less informative and more robust to randomization, tuned with a *MinDist* value of 0.875, as it minimizes the corresponding AIC curve and yields a discrete distribution of random effects which is representative of the population features observed during the analysis. With respect to our application, our model assesses the positive effect of the ACE inhibitors in Heart Failure treatment, as it yields a reduced hazard rate for a new hospitalization event and an higher survival probability for adherent patients. Moreover, our model discovers a latent partition of patients in three subpopulations (*Protected*, *At Risk* and *Very Fragile*), characterized by their stratified hospitalization and survival probability baselines, which are significantly different. In particular, this enables further analyses: it allow us to argue, for example, that the assessed positive effect that adherence to the ACE inhibitors treatment have in reducing the hospitalization and death rates is likely to be unappreciable for outlier patients belonging to the *Very Fragile* class, due to their personal fragility. Finally, we conclude that the proposed approach is an effective inferential tool to jointly model hospitalizations and survival of patients, and provides an easy to understand and more informative frailty characterization than the other jointly models studied in this thesis.

5 | Discussion and Conclusions

The main focus of this thesis is to develop a methodology to jointly model two dependent longitudinal processes, the first being a recurrent events process, informative for the second, which instead involves events classified as terminal. In our case, the two processes were represented by hospitalizations and departure of patients affected by heart failure who undergo ACE inhibitors therapy. The work actually stems from the idea of expanding previous analyses developed in the field of pharmacoepidemiology, centered on the effect that adherence to drug prescriptions has on the survival outcome of patients, to include in the modelling the recurrent hospitalizations of a patient, in view of the renewed necessity of a proper managing of hospital beds. A further challenge was represented by the complexity and dimensionality of the Heart Failure Dataset [32], in which the considered data are collected.

To accomplish our goal, we considered classical survival tools, noting their inadequacy as they allow to model only one event at a time. Then, we delved into the recurrent events modelling theory, identifying frailty models as the principal tool around which we construct our analysis. Our first attempt consisted in analyzing two separate frailty models for the two processes, following the approach in [48]. The next step was the joint modelling of the two processes, in order to make hospitalizations informative of death, which was achieved linking frailties. Keeping the disjoint models as reference, we compared different joint models, both from the coefficients estimation and from the frailties characterization point of views. In particular, we considered a model proposed by Rondeau et al. in 2007 [51], which showed possible instability in coefficients estimation and a too naive characterization of the joint modelling of random effects. In addition, we studied a model proposed by Ng et al. in 2020 [37], fixing the stability problem and achieving a higher coherence in coefficients estimation, as well as an elegant joint characterization of random effects. At last, in an attempt to remedy to the difficult practical exploitability of the random effects Normal distribution characterization, we proposed an original model in which we considered discrete non parametric bivariate frailties. This model represents the main contribution of this thesis, as we designed it from scratches, specifying its the-

oretical foundations and its estimation algorithm. The novelty of the model resides in the fact that it identifies a discrete distribution for the bivariate random frailties, which can lead, thanks to its ease of interpretation, to an improved management of patients' hospital courses, in addition to opening the path to insightful latent partition analyses.

The application of the model to the ACE inhibitors dataset yielded a good stability and coherence (with respect to other considered models) in coefficients estimation. From a clinical point of view we can conclude that adherent patients are significantly less at risk of a new hospitalization event and likely to survive longer than non adherent ones. Moreover, the model discovers a discrete distribution of random effects which defines a latent partition of patients in three subpopulations, characterized by fairly different levels of fragility. The first *Protected* subpopulation can be considered as a baseline, since it comprehends the biggest proportion of patients (55%), while the second *At Risk* is characterized by those subjects with a higher degree of fragility (36%), who are thus more likely to experience an adverse event (hospitalization or death). The third subpopulation is instead composed by outliers (9%), i.e. *Very Fragile* patients, characterized by the fact that the protective effect of adherence to ACE inhibitors treatment is likely to be unappreciable in the evolution of their clinical paths due to their personal fragility.

The considerably wide breadth of our work opens the way for several further developments, in each of the many areas that this thesis involves. From a pharmacoepidemiology point of view, the natural continuation of the work is to extend the performed analysis to the other drugs (Angiotensin Receptor Blockers,Anti-Aldosterone Agents, Beta Blocking Agents and Diuretics) presents in the HF dataset, to assess the effect of adherence to prescriptions in the proposed context of recurrent and terminal events. Moreover, an additional step would be the design of a joint model capable of dealing with drug interactions, instead of treating solely adherence to monotherapies. Another possible aspect to develop regards the optimization of the implementation of Ng and others model. Actually, when deciding to adopt a parametric modeling of the correlated frailties, it represents a valid option in the analysis; however, our implementation requires very high computational and time requirements when treating dataset of significant dimensions. An optimized version, capable of compete with R packages from these requirements point of view, would probably make this model a popular alternative. Of course, however, the majority of the research paths that our work opens are related to our non parametric discrete frailty model. Being a new, complex model, even if it relies on strong theoretical foundations, different adjustments may strengthen its performances: the implementation of a com-

prehensive method to compute standard errors, capable of providing estimates also for the parameters of the discrete distribution (e.g. the one suggested in Section 3.4.6); the design of different initialization or different support reduction procedures; the evaluation of different kind of distance with respect to the Euclidean one in the support reduction algorithm; the design of a procedure to tune the threshold under which two points are merged during support reduction, which we confirmed to be the most important model hyperparameter. In particular, a proper investigation of this last topic would allow the model to be considered as a valid, usable alternative, making it worth to implement a dedicated R package.

Bibliography

[1] O. Aelen. Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6:701–726, 1978.

[2] L. Amorim and J. Cai. Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44:324–333, 2014.

[3] P. Andersen and R. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[4] P. Andersen, O. Borgan, R. Gill, and N. Keiding. *Statistical Models Based on Counting Processes.* Springer-Verlag, 1993.

[5] S. Andrade, K. Kahler, F. Frech, and F. Chan. Methods for evaluation of medication adherence and persistence using automated databases. *Pharmacoepidemiology and Drug Safety*, 15:565–574, 2006.

[6] P. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, chapter 9, pages 450–455. Pringer-Verlag, 2006.

[7] N. Breslow. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453, 1974.

[8] T. P. Callanan. Restricted maximum likelihood estimation of variance components: computational aspects. *Retrospective Theses and Dissertations, Iowa State University.*

[9] J. Cavanaugh and A. Neath. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, (3), 2019.

[10] S. Cole, H. Chu, and S. Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology*, 179:252–260, 2014.

[11] D. Collet. *Modelling survival data in medical research*, chapter 3, page 72. Chapman & Hall, 2015.

[12] D. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, 34:187–220, 1972.

[13] D. Cox. Partial likelihood. *Biometrika*, 1995.

[14] D. Eddelbuettel and J. Balamuta. Extending extitR with extitC++: A Brief Introduction to extitRcpp. *The American Statistician*, 72(1):28–36, 2018. doi: 10.1080/00031305.2017.1375990.

[15] B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

[16] European Medicines Agency. ATC code. URL `https://www.ema.europa.eu/en/glossary/atc-code`.

[17] T. R. Fleming and D. Harrington. *Counting Processes and Survival Analysis*, page 164. John Wiley and Sons Inc., 1991.

[18] J. Foulley. A simple argument showing how to derive restricted maximum likelihood. *Journal of Dairy Science*, (8), 1993.

[19] J. Gagne, R. J. Glinn, J. Avorn, R. Levin, and S. Schneeweiss. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *Journal of Clinical Epidemiology*, 64(7):749–759, 2011.

[20] F. Gasperoni, A. M. Paganoni, F. Ieva, C. Jackson, and L. Sharples. Non-parametric frailty cox models for hierarchical time-to-event data. *Biostatistics*, 2020.

[21] D. Hosmer, S. Lemeshow, and S. Mai. *Applied Survival Analysis: Regression Modelling of Time to Event Data*. Wiley-Interscience, 2008.

[22] P. Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1:255–273, 1995.

[23] X. Huang and R. Wolfe. A frailty model for informative censoring. *Biometrics*, 58: 510–520, 2002.

[24] J. Kalbfleish and R. Prentice. *The Statistical Analysis of Failure Times Data*. John Wiley Sons, 2002.

[25] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[26] S. Karve, M. Cleve, M. Helm, T. Hudson, D. West, and B. Martin. Prospective validation of eight different adherence measures for use with administrative claims data among patients with schizophreniahe. *Value in Health*, 12(6), 2009.

[27] D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text.* Springer, 1996.

[28] D. Lin, W. Sun, and Z. Ying. Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika*, 86:59–70, 1999.

[29] D. Lin, L. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society B*, 62: 711–730, 2002.

[30] L. Liu, X. Huang, A. Yaroshinki, and J. Cormier. Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics*, 72:204–214, 2016.

[31] C. Masci, A. M. Paganoni, and F. Ieva. Semiparametric mixed effects models for unsupervised classification of italian schools. *Journal of the Royal Statistical Society*, 2019.

[32] C. Mazzali, A. M. Paganoni, F. Ieva, C. Masella, M. Maistrello, A. Ornella, S. Scalvini, and M. Frigerio. Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in lombardy region, 2000 to 2012. *BMC Health Services Research*, 16(234), 2016.

[33] J. J. McMurray, S. Adamopoulos, and Anker, S. D. ESC Committee for Practice Guidelines (2012). ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. *European heart journal*, 33:1787–1847. doi: https://doi.org/10.1093/eurheartj/ehs104.

[34] Ministero della Salute della Repubblica Italiana. Scompenso Cardiaco, 2022. URL `https://www.salute.gov.it/portale/salute/p1_5.jsp?area=Malattie_cardiovascolari&id=43&lingua=italiano`.

[35] y. . . u. . h. MOX, Laboratory for Modeling and Scientific Computing, Politecnico di Milano, title = HPC@MOX.

[36] W. Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1:27–52, 1969.

[37] S. K. Ng, R. Tawiah, G. McLachlan, and V. Gopalan. Joint frailty modelling of time-to-event data to elicit the evolution pathway of events: A generalised linear mixed model approach. *Biostatistics*, 0(0):1–25, 2020. doi: 10.1093/biostatistics/output.

[38] W. H. Organization. *Introduction to Drug Utilization Research.* WHO Library Cataloguing-in-Publication Data, 2003. URL `https://apps.who.int/iris/bitstream/handle/10665/42627/924156234X.pdf?sequence=1&isAllowed=y`.

[39] A.-K. Ozga, M. Kieser, and G. Rauch. A systematic comparison of recurrent event models for application to composite endpoints. *BMC Medical Research Methodology*, 18, 2018.

[40] G. Paulon, M. De Iorio, A. Guglielmi, and F. Ieva. Joint modeling of recurrent events and survival: a bayesian non-parametric approach. *Biostatistics*, 21:1–14, 2020.

[41] P. Pazos-Lopez et al. The causes, consequences, and treatment of left or right heart failure, Vascular Health and Risk Management.

[42] V. Phillips, B. Hejblum, M. Prague, D. Commenges, and C. Proust-Lima. Robust and efficient optimization using a marquardt-levenberg algorithm with R package marqlevalg.

[43] R. Prentice, B. Williams, and A. Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981.

[44] D. Price and A. Manatunga. Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine*, 20:1515–1527, 2001.

[45] S. Prinja, N. Gupta, and R. Verma. Censoring in clinical trials: review of survival analysis techniques. *Indian J Community Med*, 35:217–221, 2010.

[46] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

[47] Regione Lombardia. HFData project: utilization of regional health source databases for evaluating epidemiology, short- and medium term outcome and process indicators in patients hospitalized for heart failure. Progetto di ricerca finalizzata di Regione Lombardia., 2012. HFData-RF-2009-1483329.

[48] S. Ripatti and J. Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56:1016–1022, 2002.

[49] G. Robinson. That blup is a good thing: the estimation of random effects. *Statistical Science*, (1):15–51, 1991.

[50] V. Rondeau, L. Filleul, and P. Joly. Nested frailty models using maximum penalized likelihood estimation. *Statistics in Medicine*, 25:4036–4052, 2006.

[51] V. Rondeau, S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran. Jointfrailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. *International Journal of Epidemiology*, 8:708–721, 2007.

[52] V. Rondeau, Y. Mazroui, and J. Gonzalez. frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47, 2012.

[53] V. Rondeau, J. R. Gonzalez, Y. Mazroui, A. Mauguen, A. Diakite, A. Laurent, M. Lopez, A. Król, and C. L. Sofeu. *frailtypack: General Frailty Models: Shared, Joint and Nested Frailty Models with Prediction; Evaluation of Failure-Time Surrogate Endpoints*, 2019. URL `https://CRAN.R-project.org/package=frailtypack`. R package version 3.0.3.

[54] R. Scaramuzza. Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete non parametric frailty approach. URL `https://github.com/RiccardoMS/Joint-frailty-modelling-of-hospitalizations-and-death-in-heart-failure-patients`

[55] D. Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69:239–241, 1982.

[56] M. Spreafico. Statistical modelling of adherence to drug prescription and its effects on survival in heart failure patients, 2016.

[57] S. Spruance, J. Reid, and M. Samore. Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, 48(8):2787–, 1982.

[58] B. Strom. *Textbook of Pharmacoepidemiology*. John Wiley and Sons, 2006.

[59] T. M. Therneau. *coxme: Mixed Effects Cox Models*, 2020. URL `https://CRAN.R-project.org/package=coxme`. R package version 2.2-16.

[60] T. M. Therneau. *A Package for Survival Analysis in R*, 2021. URL `https://CRAN.R-project.org/package=survival`. R package version 3.2-13.

[61] H. Van der Wal, V. Van Deursen, P. Van der Meer, and A. Voors. Comorbidities

in heart failure. *Handbook of Experimental Pharmacology*, 243:35–66, 2017. doi: 10.1007/164_2017_27.

[62] F. Widmer. Herzinsuffizienz und komorbiditäten [comorbidity in heart failure]. *Therapeutische Umschau*, 68(2):103–106, 2011. doi: 10.1024/0040-5930/a000127.

[63] World Health Organization Collaborating Center for Drug Statistics Methodology. ATC/DDD index 2022. URL `https://www.whocc.no/atc_ddd_index/`.

[64] W. Yang, C. Jepson, D. Xie, J. Roy, J. Yenchih Hsu, A. Anderson, R. Landis, J. He, and H. Feldman. Statistical methods for recurrent event analysis in cohort studies of ckd. *Clinical Journal of the American Society of Nephrology*, 12:2066–2073, 2017.

A | Appendix A

In this Appendix is reported the arguments followed to choose between a full and a partial likelihood approach in the estimation procedure for the non parametric discrete frailty model presented in Section 3.4.

Let's assume, for simplicity and without loss of generalization, to have to deal with a single process where each subject experiences one (possibly censored) event only. When random effects are included in the linear predictor of a Cox model, the Cox argument ([13]) can be extended. In particular, we can write

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \Big[\frac{\exp\{\boldsymbol{X}_i^T\boldsymbol{\beta} + u_i\}}{\sum_{l\in\mathcal{R}(t)} \exp\{\boldsymbol{X}_l^T\boldsymbol{\beta} + u_l\}} \Big]^{\delta_i} \text{ x}$$
$$\Big[\sum_{l\in\mathcal{R}(t)} \exp\{\boldsymbol{X}_l^T\boldsymbol{\beta} + u_l \Big]^{\delta_i} exp\{-H_0(t)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}+u_i}\}$$

In this case, maximizing the partial likelihood (first term) is suitable assuming that it contains almost all the information regarding $\boldsymbol{\beta}$ and the random effects distribution, while the second term encloses the information regarding the cumulative baseline hazard function. However, we must recall that frailty models have a double nature: reparametrizing the model in order to obtain a multiplicative frailty we obtain a likelihood such as

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{N} [h_0(t_i)e^{u_i}e^{\boldsymbol{X}_{ij}^T\boldsymbol{\beta}}]^{\delta_i} \exp\{-H_0(t_i)e^{u_i}e^{\boldsymbol{X}_{ij}^T\boldsymbol{\beta}}\} \\ &= \prod_{i=1}^{N} [h_{0i}(t_i)e^{\boldsymbol{X}_{ij}^T\boldsymbol{\beta}}]^{\delta_i} \exp\{-H_{0i}(t_i)e^{\boldsymbol{X}_{ij}^T\boldsymbol{\beta}}\} \end{aligned}$$

We can notice that frailties can be interpreted as a way to apply stratification of baseline hazards at patients' level. This twofold interpretation of frailties contrast somehow the Cox assumption, since part of the information useful to characterize frailties is likely to be included in the term discarded by partial likelihood. The partial likelihood reasoning

can still be applied in particular contexts, such as in the model proposed by Ng (Section 3.3.2). Actually, in that case we were considering

- a patient-specific frailty
- a frailty distribution centered in zero

thus, the stratification effect at baselines level is likely to be negligible. However, in our model we consider a discrete and finite (likely to be small) support for the frailties distribution. For this reason, the interpretation of frailties as stratifying factors gains significance with respect to consider them only as nuisance variables in the linear predictor. For this reason, it is conceptually better to adopt a full likelihood approach in the model estimation.

Moreover, there is also a more practical reason to back up this choice. Since the estimation will require the designing of an Expectation-Maximization algorithm (involving the likelihood random variables) we will need to compute probabilities in the expectation step. In this case, the computation will require estimates for the instantaneous and cumulative baseline hazards, thus the adoption of a full likelihood approach is more suitable for our needs.

B | Appendix B

In this Appendix is reported the R script containing the custom implementation of the estimation routine of the model by Ng et al.[37], summarized in Section 3.3.2, as well as two C++ auxiliary functions used for parallelizing the matricial product and computing the REML update. To navigate through the code, refers to the following scheme

- *lines 1 – 46*: data loading and preparation
- *lines 47 – 107*: initial values, data structures and convergence parameters definition
- *line 108*: start routine
  - *lines 120 – 203*: inner loop, Newton-Rapshon update computation
  - *lines 205 – 220*: outer loop, REML parameters update computation
  - *lines 230 – 262*: Variance-covariance matrices of obtained estimates
- *lines 265 – 282*: Summary and Results

```
1  ############################################################
2  #          Ng, Tawiah, MacLachlan, Gopalan                 #
3  ############################################################
4
5  # Load packages
6  print("load packages")
7  library(data.table)
8  library(Rcpp)
9  library(parallel)
10
11 # Load Data
12 print("load data")
13 load("dataRec.RData")
14 load("dataDeath.RData")
15
```

```
# RCpp function to speed up matrix multiplication
print("load cpp functions")
library(Rcpp)
sourceCpp("MatProd.cpp")
sourceCpp("REMLupdate.cpp")

# Order data according to observed gap times
print("order gap times")
order1 <- order(data$GapEvent)
order2 <- order(dataDeath$GapEvent)
data <- data[order1,]
dataDeath <- dataDeath[order2,]

# Variables
print("define variables")
ID_rec          <- factor(data$COD_REG)
ADERENTE_rec    <- factor(data$ADERENTE)
SESSO_rec       <- factor(data$SESSO)
etaEvent_rec    <- as.double(data$etaEvent)
comorbidity_rec <- as.double(data$comorbidity)

ID_term          <- factor(dataDeath$COD_REG)
ADERENTE_term    <- factor(dataDeath$ADERENTE)
SESSO_term       <- factor(dataDeath$SESSO)
etaEvent_term    <- as.double(dataDeath$etaEvent)
comorbidity_term <- as.double(dataDeath$comorbidity)

# Design parameters
n_beta = 4                          # recurrent event covariates
n_gamma = 4                         # terminal event covariates
n_subjects = dim(dataDeath)[1]      # number of units subject to random
    effect

# Initial values: set to zero and
print("set initial values")
# Omega
#beta0  <- rep(0,n_beta)
#gamma0 <- rep(0,n_gamma)
#u0     <- rep(0,n_subjects)
#v0     <- rep(0,n_subjects)
#Omega0 <- c(beta0,gamma0,u0,v0)
#rm(beta0,gamma0,u0,v0)
load("Ninth_run/Omega.RData")
# Phi
```

```
#theta_u<- 0.7
#theta_v<- 0.7
#rho    <- 0.5
#Phi0   <- c(theta_u, theta_v, rho)
#rm(theta_u,theta_v,rho)
load("Ninth_run/Phi.RData")

# General Data Structures: not modified in the loop, only defined once
# Design Matrices [X1,X2,Z1,Z2]
print("design matrices")
X1 = model.matrix(~ SESSO_rec + ADERENTE_rec + etaEvent_rec + comorbidity_
    rec)[,-1]
X2 = model.matrix(~ SESSO_term + ADERENTE_term + etaEvent_term +
    comorbidity_term)[,-1]
Z2 = diag(n_subjects)
Z2 <-Z2[order2,]

loadZ1 <- TRUE
if(loadZ1){
load("RE_design_matrix_FFU.RData")
} else {
print("computing Z1..")
Z1 <- matrix(0,dim(X1)[1],n_subjects)
codici <- levels(ID_rec)
for(i in 1:length(codici)){
     current <- codici[i]
     for(j in 1:dim(X1)[1]){
       if(ID_rec[j]==current)
         Z1[j,i]=1
     }
  }
save(Z2, file="RE_design_matrix_FFU.RData")
}

# Clean memory
rm(SESSO_rec,ADERENTE_rec,etaEvent_rec,comorbidity_rec,SESSO_term,ADERENTE_
    term,etaEvent_term,comorbidity_term)
gc()

# Censoring vectors
DELTA_R <- data$event
DELTA_D <- dataDeath$cens

# Convergence
```

```
conv1 <- FALSE       # External loop
maxit1<- 100
maxit2<- 10
it1   <- 0
treshold <- 1e-3
treshold1<- 1e-6

# Start loop
print("start loop!")
while(!conv1 & it1<maxit1){
  cat(paste("\nCurrent outer it: ", it1))
  print("Current Estimates:")
  print(paste("Beta: ", Omega0[1:n_beta]))
  print(paste("Gamma: ", Omega0[(n_beta+1):(n_beta+n_gamma)]))
  print(paste("u0 NaN: ", table(is.nan(Omega0[(n_beta+n_gamma+1):(n_beta+n_
      gamma+n_subjects)]))))
  print(paste("v0 NaN: ", table(is.nan(Omega0[(n_beta+n_gamma+n_subjects+1)
      :(n_beta+n_gamma+2*n_subjects)]))))
  print(paste("Phi: ", Phi0))

  it2 = 0
  conv2 <- FALSE
  while(!conv2 & it2<maxit2){
    cat(paste("\nCurrent inner it: ", it2))
    print("Current Estimates:")
    print(paste("Beta: ", Omega0[1:n_beta]))
    print(paste("Gamma: ", Omega0[(n_beta+1):(n_beta+n_gamma)]))
    print(paste("u0 NaN: ", table(is.nan(Omega0[(n_beta+n_gamma+1):(n_beta+
        n_gamma+n_subjects)]))))
    print(paste("v0 NaN: ", table(is.nan(Omega0[(n_beta+n_gamma+n_subjects
        +1):(n_beta+n_gamma+2*n_subjects)]))))
    print(paste("Phi: ", Phi0))

    # G first   (depend on beta and u --> in the loop)
    Q1       <- matrix(0,dim(X1)[1],dim(X1)[1])
    diag(Q1)<- exp(MatProd(X1,Omega0[1:n_beta]) + MatProd(Z1,Omega0[(n_beta
        +n_gamma+1):(n_beta+n_gamma+n_subjects)]))
    E1       <- matrix(0,dim(X1)[1],dim(X1)[1])
    diag(E1)<- DELTA_R/rev(cumsum(rev(diag(Q1))))
    F1       <- matrix(0,dim(X1)[1],dim(X1)[1])
    F1[lower.tri(F1)] <- 1
    diag(F1)<-1
    S1       <- matrix(0,dim(X1)[1],dim(X1)[1])
    diag(S1)<- cumsum(diag(E1))
```

```
139     D1<- MatProd(Q1,S1) - MatProd(Q1,MatProd(F1,MatProd(E1^2,MatProd(t(F1),
            Q1))))
140     dL1_dEta <-DELTA_R - MatProd(Q1,MatProd(F1,MatProd(E1,rep(1,dim(X1)[1])
            )))
141     rm(Q1,E1,F1,S1)
142     gc()
143
144     Q2      <- matrix(0,dim(X2)[1],dim(X2)[1])
145     diag(Q2)<- exp(MatProd(X2,Omega0[(n_beta+1):(n_beta+n_gamma)])+ MatProd
            (Z2,Omega0[(n_beta+n_gamma+n_subjects+1):(n_beta+n_gamma+2*n_
            subjects)]))
146     E2      <- matrix(0,dim(X2)[1],dim(X2)[1])
147     diag(E2)<- DELTA_D/rev(cumsum(rev(diag(Q2))))
148     F2      <- matrix(0,dim(X2)[1],dim(X2)[1])
149     F2[lower.tri(F2)] <- 1
150     diag(F2)<-1
151     S2      <- matrix(0,dim(X2)[1],dim(X2)[1])
152     diag(S2)<- cumsum(diag(E2))
153     D2<- MatProd(Q2,S2) - MatProd(Q2,MatProd(F2,MatProd(E2^2,MatProd(t(F2),
            Q2))))
154     dL1_dZeta<-DELTA_D - MatProd(Q2,MatProd(F2,MatProd(E2,rep(1,dim(X2)[1])
            )))
155     rm(Q2,E2,F2,S2)
156     gc()
157
158     # G2
159     temp <- matrix(0,2,2)
160     temp[1,1]<-1/(Phi0[1]*(1-Phi0[3]^2))
161     temp[1,2]<- -Phi0[3]/(sqrt(Phi0[1]*Phi0[2])*(1-Phi0[3]^2))
162     temp[2,1]<- -Phi0[3]/(sqrt(Phi0[1]*Phi0[2])*(1-Phi0[3]^2))
163     temp[2,2]<-1/(Phi0[2]*(1-Phi0[3]^2))
164     G2 <- kronecker(temp,diag(nrow=n_subjects))
165
166     # G
167     G = matrix(0, n_beta+n_gamma+2*n_subjects,n_beta+n_gamma+2*n_subjects)
168     G[1:n_beta,1:n_beta]=MatProd(t(X1),MatProd(D1,X1))
169     G[(n_beta+1):(n_beta+n_gamma),(n_beta+1):(n_beta+n_gamma)]=MatProd(t(X2
            ),MatProd(D2,X2))
170     G[1:n_beta,(n_beta+n_gamma+1):(n_beta+n_gamma+n_subjects)]=MatProd(t(X1
            ),MatProd(D1,Z1))
171     G[(n_beta+1):(n_beta+n_gamma),(n_beta+n_gamma+n_subjects+1):(n_beta+n_
            gamma+2*n_subjects)]=MatProd(t(X2),MatProd(D2,Z2))
172     G[(n_beta+n_gamma+1):(n_beta+n_gamma+n_subjects),1:n_beta]=MatProd(t(Z1
            ),MatProd(D1,X1))
```

```
173     G[(n_beta+n_gamma+n_subjects+1):(n_beta+n_gamma+2*n_subjects),(n_beta
            +1):(n_beta+n_gamma)]=MatProd(t(Z2),MatProd(D2,X2))
174     G[(n_beta+n_gamma+1):(n_beta+n_gamma+n_subjects),(n_beta+n_gamma+1):(n_
            beta+n_gamma+n_subjects)]=MatProd(t(Z1),MatProd(D1,Z1))
175     G[(n_beta+n_gamma+n_subjects+1):(n_beta+n_gamma+2*n_subjects),(n_beta+n
            _gamma+n_subjects+1):(n_beta+n_gamma+2*n_subjects)]=MatProd(t(Z2),
            MatProd(D2,Z2))
176     rm(D1,D2)
177     gc()
178
179     G[(n_beta+n_gamma+1):(n_beta+n_gamma+2*n_subjects),(n_beta+n_gamma+1):(
            n_beta+n_gamma+2*n_subjects)] = G[(n_beta+n_gamma+1):(n_beta+n_gamma
            +2*n_subjects),(n_beta+n_gamma+1):(n_beta+n_gamma+2*n_subjects)] +
            G2
180     rm(G2)
181     gc()
182
183     # Gradient dL/dOmega [upd_step]
184     dBeta <-MatProd(t(X1),dL1_dEta)
185     dGamma<-MatProd(t(X2),dL1_dZeta)
186     dU      <-MatProd(t(Z1),dL1_dEta)   - Omega0[(n_beta+n_gamma+1):(n_beta+n_
            gamma+n_subjects)]/(Phi0[1]*(1-Phi0[3]^2)) + (Phi0[3]/(sqrt(Phi0[1]*
            Phi0[2])*(1-Phi0[3]^2)))*Omega0[(n_beta+n_gamma+n_subjects+1):(n_
            beta+n_gamma+2*n_subjects)]
187     dV      <-MatProd(t(Z2),dL1_dZeta) - Omega0[(n_beta+n_gamma+n_subjects+1)
            :(n_beta+n_gamma+2*n_subjects)]/(Phi0[2]*(1-Phi0[3]^2)) + (Phi0[3]/(
            sqrt(Phi0[1]*Phi0[2])*(1-Phi0[3]^2)))*Omega0[(n_beta+n_gamma+1):(n_
            beta+n_gamma+n_subjects)]
188     upd_step <- c(dBeta,dGamma,dU,dV)
189
190     # clean memory
191     rm(dBeta,dGamma,dU,dV)
192     gc()
193
194     # update Omega
195     x       <- qr.solve(G,upd_step, tol=1e-10)
196     Omega0 <- Omega0 + x
197     # update iteration
198     it2 <- it2+1
199
200     # check convergence
201     if(norm(x,type="2")<treshold1)
202       conv2=TRUE
203   }
```

```
204
205   ## update Phi
206   # invert G
207   invG    <- chol2inv(chol(G))
208   invG_bb<- invG[(n_beta+n_gamma+1):nrow(invG),(n_beta+n_gamma+1):ncol(invG
          )]
209
210   # update
211   Phi1 <- REMLupdate(Omega0[(n_beta+n_gamma+1):length(Omega0)],invG_bb,n_
          subjects)
212   converged <- norm(Phi0 - Phi1,type="2")
213   Phi0 <- Phi1
214   gc()
215
216   # update iteration
217   it1 <- it1+1
218
219   # check convergence
220   if(converged<treshold)
221     conv1=TRUE
222 }
223
224 save(Omega0, file="Omega.RData")
225 save(Phi0, file="Phi.RData")
226
227 # Clean Memory
228 rm(G)
229
230 ## Variance-covariance Matrices for obtained estimates
231 # Beta & Gamma
232 varBeta <- invG[1:n_beta,1:n_beta]
233 varGamma<- invG[(n_beta+1):(n_beta+n_gamma),(n_beta+1):(n_beta+n_gamma)]
234 # Theta_u2,Theta_v2,rho
235 Epsilon <- kronecker(matrix(c(Phi0[1],Phi0[3]*sqrt(Phi0[2]*Phi0[1]),Phi0[3]
          *sqrt(Phi0[2]*Phi0[1]),
236                                       Phi0[2]),2,2),diag(nrow=n_subjects))
237 dEpsInv_dThetaU2<-kronecker(matrix(c((-1)/(Phi0[1]^2*(1-Phi0[3]^2)),Phi0[3]
          /(2*sqrt(Phi0[2])*(1-Phi0[3]^2)*Phi0[1]^3),Phi0[3]/(2*sqrt(Phi0[2])*(1-
          Phi0[3]^2)*Phi0^3),
238                                                 0),2,2),diag(nrow=n_subjects))
239 dEpsInv_dThetaV2<-kronecker(matrix(c(0,Phi0[3]/(2*sqrt(Phi0[1])*(1-Phi0
          [3]^2)*Phi0[2]^3),Phi0[3]/(2*sqrt(Phi0[1])*(1-Phi0[3]^2)*Phi0[2]^3),
240                                                 (-1)/(Phi0[2]^2*(1-Phi0[3]^2))),2,2),
                                                      diag(nrow=n_subjects))
```

```
dEpsInv_dRho<-kronecker(matrix(c((2*Phi0[3])/(Phi0[1]*(1-Phi0[3]^2)^2),-(Phi0[3]^2+1)/(sqrt(Phi0[1]*Phi0[2])*(1-Phi0[3]^2)^2),-(Phi0[3]^2+1)/(sqrt(Phi0[1]*Phi0[2])*(1-Phi0[3]^2)^2),
                                        (2*Phi0[3])/(Phi0[2]*(1-Phi0[3]^2)^2))
                                            ,2,2),diag(nrow=n_subjects))

J1     <- MatProd(invG[(n_beta+n_gamma+1):(n_beta+n_gamma+2*n_subjects),(n_beta+n_gamma+1):(n_beta+n_gamma+2*n_subjects)],dEpsInv_dThetaU2)
J2     <- MatProd(Epsilon,dEpsInv_dThetaU2)
J3     <- MatProd(invG[(n_beta+n_gamma+1):(n_beta+n_gamma+2*n_subjects),(n_beta+n_gamma+1):(n_beta+n_gamma+2*n_subjects)],dEpsInv_dThetaV2)
J4     <- MatProd(Epsilon,dEpsInv_dThetaV2)
J5     <- MatProd(invG[(n_beta+n_gamma+1):(n_beta+n_gamma+2*n_subjects),(n_beta+n_gamma+1):(n_beta+n_gamma+2*n_subjects)],dEpsInv_dRho)
J6     <- MatProd(Epsilon,dEpsInv_dRho)

A      <- matrix(0,3,3)
A[1,1]<- sum(diag((J1-J2)))^2
A[1,2]<- sum(diag((MatProd(J1,J3)+MatProd(J2,J4)-2*MatProd(J1,J4))))
A[1,3]<- sum(diag((MatProd(J1,J5)+MatProd(J2,J6)-2*MatProd(J1,J6))))
A[2,2]<- sum(diag((J3-J4)))^4
A[2,3]<- sum(diag((MatProd(J3,J5)+MatProd(J4,J6)-2*MatProd(J3,J6))))
A[3,3]<- sum(diag((J5-J6)))^2
A[3,1]<- A[1,3]
A[2,1]<- A[1,2]
A[3,2]<- A[2,3]
VarPhi<- 2*chol2inv(chol(A))
rm(A,J1,J2,J3,J4,J5,J6,Epsilon,dEpsInv_dRho,dEpsInv_dThetaU2,dEpsInv_dThetaV2)


# CI for HR, Survival probabilities
# HR and 95pc CI
summary <- data.frame(
   Estimate=c(Omega0[1:(n_beta)],Omega0[(n_beta+1):(n_beta+n_gamma)],Phi0),
   StDev    =c(sqrt(diag(varBeta)),sqrt(diag(varGamma)),sqrt(diag(VarPhi))),
   HR       =exp(c(Omega0[1:(n_beta)],Omega0[(n_beta+1):(n_beta+n_gamma)],
       Phi0)),
   L95      = exp(c(Omega0[1:(n_beta)],Omega0[(n_beta+1):(n_beta+n_gamma)],
       Phi0) - 1.96*c(sqrt(diag(varBeta)),sqrt(diag(varGamma)),sqrt(diag(
       VarPhi)))),
   U95      = exp(c(Omega0[1:(n_beta)],Omega0[(n_beta+1):(n_beta+n_gamma)],
       Phi0) + 1.96*c(sqrt(diag(varBeta)),sqrt(diag(varGamma)),sqrt(diag(
       VarPhi)))))
```

```
row.names(summary) <- c("Beta1","Beta2","Beta3","Beta4","Gamma1","Gamma2","
    Gamma3","Gamma4","ThetaU2","ThetaV2","rho")
summary

# Save Results
FFU_try <- list(beta=Omega0[1:n_beta],gamma=Omega0[(n_beta+1):(n_beta+n_
    gamma)],
                    u=Omega0[(n_beta+n_gamma+1):(n_beta+n_gamma+n_subjects)],
                    v=Omega0[(n_beta+n_gamma+n_subjects+1):(n_beta+n_gamma+2*n
                        _subjects)],
                    thetaU2=Phi0[1],thetaV2=Phi0[2],rho=Phi0[3],
                    HR=summary$HR,summary=summary,VarBeta=varBeta,varGamma=
                        varGamma,VarPhi=VarPhi)
save(FFU_try,file='FFU_ACE.RData')
```

```
// Parallel Matricial Product

// [[Rcpp::depends(RcppArmadillo, RcppEigen)]]
// [[Rcpp::plugins(openmp)]]

#include <omp.h>
#include <RcppArmadillo.h>
#include <RcppEigen.h>

// [[Rcpp::export]]
SEXP MatProd(const Eigen::Map<Eigen::MatrixXd> A,
                    Eigen::Map<Eigen::MatrixXd> B,
                    int n_cores=40){

  Eigen::setNbThreads(n_cores);
  Eigen::MatrixXd C = A * B;
  return Rcpp::wrap(C);
}
```

```
// REML update

#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericVector REMLupdate(NumericVector R, NumericMatrix Bqq, int M) {
  NumericVector T(3);
  NumericVector out(3);

  for(int i=0;i<M;i++){
```

```
12      T[0]+=Bqq(i,i)+pow(R[i],2);
13      T[2]+=Bqq(M+i,M+i)+pow(R[M+i],2);
14      T[1]+=(Bqq(i,M+i)+Bqq(M+i,i)+2*R[i]*R[M+i])/2;
15    }
16
17    out[1]=T[2]/M;
18    out[0]=T[0]/M;
19    out[2]=T[1]/sqrt(T[0]*T[2]);
20
21    return out;
22
23  }
```

C | Appendix C

In this Appendix is reported the R script implementing the EM algorithm designed to estimate the nonparametric discrete frailty model (Section 3.4). To navigate through the code, refers to the following scheme

- *lines 1 – 98*: data loading, preprocessing, initialization of data structures
- *lines 100 – 106*: convergence parameters definition
- *lines 110 – 120*: Hyperparameters choice (Gaussian/Uniform Initialization & MinDist)
- *lines 121 – 128*: Loglikelihood Function Definition
- *lines 130 – 180*: Grid Initialization
  - *lines 130 – 140*: Retrieve variance parameters to define Region of interest
  - *lines 141 – 148*: Gaussian Initialization
  - *lines 149 – 157*: Uniform over a rectangle Initialization
  - *lines 159 – 180*: Support Reduction step (in case of Gaussian Initialization)
- *lines 181 – 210*: Other parameters Initialization
  - *lines 181 – 185*: Random frailty allocation to patients
  - *lines 187 – 197*: Hospitalization process parameters initialization
  - *lines 199 – 210*: Death process parameters initialization
- *lines 211 – 219*: Definition of auxiliary structures for computations and saving
- *lines 220 – 396*: Start loop
  - *lines 222 – 259*: Support Reduction
  - *lines 260 – 306*: Expectation Step
  - *lines 308 – 371*: Maximization Step
    - *lines 308 – 324*: Support Reduction - elimination of unassigned points

    * *lines 326 –332*: Update probabilities
    * *lines 334 – 339*: Update support of random effects discrete distributions
    * *341 – 354*: Covariate coefficients' update
    * *355 – 371*: Baseline Hazards and Cumulative Hazards update
  - *lines 372 – 396*: Check convergence and save estimated values

The resulting models described in Section 4.4 are obtained using this version of the code. Different versions, comprehending various features like an alternative to the support reduction procedure or a constrained optimization step in the support update, can be found at [54], as well as the scripts used to perform the analysis of the obtained models.

```
##########################################
#  Nonparametric Discrete Frailty #
##########################################
rm(list=ls())
set.seed(210197)

# load packages
library(data.table)
library(survival)
library(mvtnorm)
library(survminer)

## data loading
load("dataRec.RData")
load("dataDeath.RData")

## data preprocessing
# ordering (not necessary?)
data <- data[order(data$GapEvent),]
dataDeath <- dataDeath[order(dataDeath$GapEvent),]

# arrange data
ID_rec          <- factor(data$COD_REG)
ADERENTE_rec    <- factor(data$ADERENTE)
SESSO_rec       <- factor(data$SESSO)
etaEvent_rec    <- as.double(data$etaEvent)
comorbidity_rec <- as.double(data$comorbidity)
```

```
28
29 ID_term           <- factor(dataDeath$COD_REG)
30 ADERENTE_term     <- factor(dataDeath$ADERENTE)
31 SESSO_term        <- factor(dataDeath$SESSO)
32 etaEvent_term     <- as.double(dataDeath$etaEvent)
33 comorbidity_term <- as.double(dataDeath$comorbidity)
34
35 # Model Matrices
36 X1 <- model.matrix(~ SESSO_rec + ADERENTE_rec + etaEvent_rec + comorbidity_
       rec )[,-1]
37 X2 <- model.matrix(~ SESSO_term + ADERENTE_term + etaEvent_term +
       comorbidity_term )[,-1]
38
39 # Response vectors
40 time1 <- data$GapEvent
41 time2 <- dataDeath$GapEvent
42
43 # Number of patients
44 N = length(unique(ID_rec))
45
46 # Number of unique times
47 nt1 = length(unique(data$GapEvent))
48 nt2 = length(unique(dataDeath$GapEvent))
49
50 # Number of observations
51 nobs1 = nrow(data)
52 nobs2 = nrow(dataDeath)
53
54 # Number of covariates
55 ncov1 = ncol(X1)
56 ncov2 = ncol(X2)
57
58 # Cumulative Hazards: data frame encoding
59 cumhaz1 = as.data.frame( cbind( hazard=rep( 0, nt1 ), time = sort( unique(
       time1 ))))
60 cumhaz2 = as.data.frame( cbind( hazard=rep( 0, nt2 ), time = sort( unique(
       time2 ))))
61
62 # Instantaneous hazards
63 haz1 = as.data.frame( cbind( hazard=rep( 0, nt1 ), time = sort( unique(
       time1 ))))
64 haz2 = as.data.frame( cbind( hazard=rep( 0, nt2 ), time = sort( unique(
       time2 ))))
65
```

```
# Number of events per patient
D1 <- table( ID_rec[ data$event == 1 ] )
orderD1<-match(unique(ID_rec),unique(data$COD_REG)[order(unique(data$COD_
    REG))])
D1<- D1[orderD1]
D2 <- table( ID_term[dataDeath$cens == 1])
orderD2<-match(unique(ID_rec),unique(dataDeath$COD_REG)[order(unique(
    dataDeath$COD_REG))])
D2<-D2[orderD2]

# Risk sets
risk_index1 <- matrix( 0, nrow = nobs1, ncol = nt1 )
risk_index2 <- matrix(0, nrow = nobs2, ncol = nt2)

# Time lists
time_list1 <- sapply( 1:nt1, function(x) !is.na(match(time1, cumhaz1$time[x
    ])))
time_list2 <- sapply( 1:nt2, function(x) !is.na(match(time2, cumhaz2$time[x
    ])))

# Number of ties
m1 <- sapply( 1:dim(time_list1)[2], function(x) sum(data$event[time_list1[,
    x]]))
m2 <- sapply( 1:dim(time_list2)[2], function(x) sum(dataDeath$cens[time_
    list2[,x]]))

# fill risk index
for( l in 1:nt1 )
{
  risk_index1[ which( time1 >= cumhaz1$time[ l ]), l ] <- 1
}
for( l in 1:nt2 )
{
  risk_index2[ which( time2 >= cumhaz2$time[ l ]), l ] <- 1
}

# Encode ID as numeric
groups1 <- match(ID_rec, unique(ID_rec))
groups2 <- match(ID_term, unique(ID_rec))

## Set Convergence Indices
# Count
count <- 0
count_conv <- 100
```

```
104
105 # epsilon
106 eps_conv = 1e-3
107 eps      = 1e5
108
109
110 ## Set Environmental variables
111 # Uniform  : if True, Initializes the grid as a discrete uniform over
112 #                     the rectangle of dimension +/- 3*sqrt(diag(Sigma))
113 #            if False, Initializes the grid sampling from the bivariate
114 #                     Normal distribution characterized by Mu and Sigma
115 # MinDist  : continuous value, set the threshold under which two points
116 #            are merged if support reduction is the chosen method
117
118 Uniform = TRUE
119 MinDist = 0.25
120
121 ## Support Functions
122 # LogLikelihood computation
123 LOGL<-function(Z,w,E_formula1,E_formula2,E_haz1,E_haz2){
124   lw<-sum(Z%*%matrix(log(w)))
125   lr<- sum(Z*(E_haz1-E_formula1))
126   ld<- sum(Z*(E_haz2-E_formula2))
127   return (lw+lr+ld)
128 }
129
130 ## Grid Initialization
131 # Print Ng estimates for initialization of Mu and Sigma
132 load("FFU_ACE_runs/Tenth_run_FFU_ACE/FFU_try.RData")
133
134 # Initialize Mu and Sigma
135 library(MASS)
136 K= 1000
137 Sigma <- matrix(c(0.12,0.0,0.0,1.4),nrow = 2, ncol = 2)
138 mu=c(0,0)
139
140 # Grid Initialization
141 if (!Uniform){
142   P <- mvrnorm(K,mu,Sigma)
143   w<-dmvnorm(P,mu,Sigma)
144   w<-w/sum(w)
145   P_show<-P
146   P_show[,1]<-P[,1]-rep(w%*%P[,1],length(P[,1]))
147   P_show[,2]<-P[,2]-rep(w%*%P[,2],length(P[,2]))
```

```
} else {
  temp_u<-seq(-3*sqrt(Sigma[1,1]),3*sqrt(Sigma[1,1]), by=MinDist)
  temp_v<-seq(-3*sqrt(Sigma[2,2]),3*sqrt(Sigma[2,2]), by=MinDist)
  P      <- unname(data.matrix(expand.grid(temp_u,temp_v)))
  w      <- rep(1/dim(P)[1],dim(P)[1])
  P_show<-P
  P_show[,1]<-P[,1]-rep(w%*%P[,1],length(P[,1]))
  P_show[,2]<-P[,2]-rep(w%*%P[,2],length(P[,2]))
  K<-dim(P)[1]
}

# Support Reduction
is_near<-TRUE
while(is_near){
  D<-dist(P)
  D<-as.matrix(D)
  D[upper.tri(D)]<-10
  diag(D)<-10
  out<-which(D == min(D), arr.ind = TRUE)
  if(D[out][1]<MinDist){
    #merge
    P[out[1,2],]=(P[out[1,2],]+P[out[1,1],])/2
    P<-P[-out[1,1],]
    #update weights
    w[out[1,2]]<-w[out[1,2]]+w[out[1,1]]
    w<-w[-out[1,1]]
    w<-w/sum(w)
    K<-K-1
  } else {
    is_near=FALSE
  }
}

## Other Parameters initialization
# Assign patient to random frailty, built frailties vectors
P_index <- sample(1:K, size=N, replace = T, prob = w)
P_off1  <- P[,1][P_index[groups1]]
P_off2  <- P[,2][P_index[groups2]]

# Estimate Initial Recurrent model, Cumulative Hazard and Hazard
cox1 <- coxph(Surv(time1,event)~SESSO_rec + ADERENTE_rec + etaEvent_rec +
                comorbidity_rec + offset(P_off1),data=data)
beta<-cox1$coefficients
s1   <- survfit(cox1,data=data)
```

```
cumhaz1$hazard = s1$cumhaz
haz1$hazard = diff(c(0,cumhaz1$hazard))
for(j in 1:length(haz1$hazard)){
  if(haz1$hazard[j]==0)
    haz1$hazard[j]<-haz1$hazard[j-1]
}

# Estimate Initial Recurrent model, Cumulative Hazard and Hazard
cox2 <- coxph(Surv(time2,cens)~SESSO_term + ADERENTE_term + etaEvent_term +
                comorbidity_term + offset(P_off2),data=dataDeath)
gamma <-cox2$coefficients
s2   <- survfit(cox2,data=dataDeath)
cumhaz2$hazard = s2$cumhaz
haz2$hazard = diff(c(0,cumhaz2$hazard))
for(j in 1:length(haz2$hazard)){
  if(haz2$hazard[j]==0)
    haz2$hazard[j]<-haz2$hazard[j-1]
}

## Support Structures Definition
# Structures for computations
numerator <- rep( 0, K )
Z <- E_formula1 <- E_formula2 <- E_haz1<-E_haz2<- matrix( 0, nrow = N, ncol
    = K)
E_part1 <- E_part2<-rep( 0, N)

# Saving structure
Saved <- list()

## Estimation Routine
# Start loop
while (eps_conv < eps & count < 100 ){

  # Save current w estimates
  w_old <- w

  # Save current Z estimates
  Z_old <- Z

  # Save current P estimates
  P_old <-P

  # Support Reduction
  is_near=TRUE
```

```
235   while(is_near){
236     D<-dist(P)
237     D<-as.matrix(D)
238     D[upper.tri(D)]<-10
239     diag(D)<-10
240     out<-which(D == min(D), arr.ind = TRUE)
241     if(D[out][1] < MinDist){
242       #merge
243       P[out[1,2],]=(P[out[1,2],]+P[out[1,1],])/2
244       P<-P[-out[1,1],]
245       #update weights
246       w[out[1,2]]<-w[out[1,2]]+w[out[1,1]]
247       w<-w[-out[1,1]]
248       w<-w/sum(w)
249       K<-K-1
250     }
251     else{
252       is_near<-FALSE
253     }
254   }
255
256   # Clean Structures
257   Z <- E_formula1 <- E_formula2 <- E_haz1 <- E_haz2<- matrix( 0, nrow = N,
        ncol = K)
258   numerator <- rep(0,K)
259
260   # Expectation Step
261   for(i in 1:N){
262
263     current_patient1 <- groups1==i
264     current_patient2 <- groups2==i
265
266     ebz1 <- exp( X1[current_patient1,] %*% beta )
267     ebz2 <- exp( X2[current_patient2,] %*% gamma)
268
269     tRij <- match(time1[current_patient1], cumhaz1$time)
270     H01t <- cumhaz1$hazard[tRij]
271     lh01t<-  log(haz1$hazard[tRij])
272
273     tDi <- match(time2[current_patient2], cumhaz2$time)
274     H02t <- cumhaz2$hazard[tDi]
275     lh02t<-  log(haz2$hazard[tDi])
276
277     E_part1[i] <- ifelse( ncov1 > 0,
```

```
                              sum( H01t*ebz1 ),
                              sum( H01t ) )
    E_part2[i] <- ifelse( ncov2 > 0,
                              H02t*ebz2 ,
                              H02t)

    for(l in 1:K){
      E_formula1[i,l] <- ifelse( ncov1 > 0,
                                 sum( H01t*ebz1*exp(P[l,1])),
                                 sum( H01t*exp(P[l,1]) ) )
      E_formula2[i,l] <- ifelse( ncov2 > 0,
                                 H02t*ebz2*exp(P[l,2]),
                                 H02t*exp(P[l,2]))
      E_haz1[i,l]      <- sum(data$event[current_patient1]*(lh01t+log(ebz1)+
          P[l,1]))

      E_haz2[i,l]      <- dataDeath$cens[current_patient2]*(lh02t+log(ebz2)+
          P[l,2])

      pivot <- min(as.numeric(D1)[i]*(P[l,1]) - E_formula1[i,l] +
                     + as.numeric(D2)[i]*(P[l,2]) - E_formula2[i,l])
      numerator[l] <- w[l]*exp(as.numeric(D1)[i]*(P[l,1]) - E_formula1[i,l]
          +
                                 + as.numeric(D2)[i]*(P[l,2]) - E_formula2[
                                     i,l])

      if(max(numerator)==0)
        numerator <- 1e-16/((as.numeric(D1)[i]*P[,1]-E_formula1[i,]+as.
            numeric(D2)[i]*P[,2]-E_formula2[i,])/pivot)
    }

    Z[i,] <- numerator/sum(numerator)

  }

  # Maximization Step
  # Latent partition
  belonging <- as.numeric( apply(Z, 1, which.max) )

  # Support Reduction - Unassigned points
  t<-table(factor(belonging, levels = 1:K))
  to_elim<-which(as.numeric(t)==0)
  if(length(to_elim)>0){
    Z<-Z[,-to_elim]
```

```
317     E_formula1<-E_formula1[,-to_elim]
318     E_formula2<-E_formula2[,-to_elim]
319     E_haz1<-E_haz1[,-to_elim]
320     E_haz2<-E_haz2[,-to_elim]
321     numerator <-numerator[-to_elim]
322     P<-P[-to_elim,]
323     K<-K-length(to_elim)
324   }
325
326   # Vector of proportions
327   if(K>1)
328     w <- (colSums(Z))/ N
329   else{
330     w <- 1
331     P <- matrix(c(0,0),nrow = 1,ncol = 2)
332   }
333
334   # Unconstrained Optimization
335   P[,1]    <- log(( as.numeric(D1) %*% Z )/( E_part1 %*% Z))
336   P[,2]    <- log(( as.numeric(D2) %*% Z )/( E_part2 %*% Z))
337   P_show     <-P
338   P_show[,1]<-P[,1]-w%*%P[,1]
339   P_show[,2]<-P[,2]-w%*%P[,2]
340
341   # Frailties update
342   P_off1 <-log((Z%*%exp(matrix(P[,1])))[groups1])
343   P_off2 <-log((Z%*%exp(matrix(P[,2])))[groups2])
344
345   # Estimate Betas
346   temp_model1 <- coxph(Surv(time1,event)~ SESSO_rec + ADERENTE_rec +
        etaEvent_rec +
347                          comorbidity_rec + offset(P_off1),data=data,method
                             = "breslow")
348   beta        <- temp_model1$coef
349
350   # Estimate Gammas
351   temp_model2 <- coxph(Surv(time2,cens)~ SESSO_term + ADERENTE_term +
        etaEvent_term +
352                          comorbidity_term + offset(P_off2),data=dataDeath,
                             method = "breslow")
353   gamma        <- temp_model2$coef
354
355   # Estimate Cumulative Hazard and Hazard - Recurrent
356   s1    <- survfit(temp_model1,data=data)
```

```
357   cumhaz1$hazard = s1$cumhaz
358   haz1$hazard = diff(c(0,cumhaz1$hazard))
359   for(j in 1:length(haz1$hazard)){
360     if(haz1$hazard[j]==0)
361       haz1$hazard[j]<-haz1$hazard[j-1]
362   }
363
364   # Estimate Cumulative Hazard and Hazard - Terminal
365   s2    <- survfit(temp_model2,data=dataDeath)
366   cumhaz2$hazard = s2$cumhaz
367   haz2$hazard = diff(c(0,cumhaz2$hazard))
368   for(j in 1:length(haz2$hazard)){
369     if(haz2$hazard[j]==0)
370       haz2$hazard[j]<-haz2$hazard[j-1]
371   }
372
373   # Convergence
374   if(length(w)==length(w_old))
375     eps <- max(abs(w - w_old),na.rm=T)
376   else if (length(w)==1)
377     eps <- 0
378   else
379     eps<-1
380
381   # Count
382   count <- count + 1
383
384   # Print
385   #print(P)
386   #print(w)
387   #print(table(belonging))
388
389   # AIC computation
390   LogL= 3*length(w)-LOGL(Z=Z,w=w,E_formula1 = E_formula1,E_formula2 = E_
          formula2,
391                          E_haz1 = E_haz1, E_haz2=E_haz2)
392   # Save
393   temp_list<-list("modelR"=temp_model1,"modelT"=temp_model2,"w"=w,"P"=P,"P_
          show"=P_show,"cumhaz1"=cumhaz1,"cumhaz2"=cumhaz2,"LogL"=LogL,"Table"=
          table(belonging))
394   Saved[[count]]<-temp_list
395 }
```

D | Appendix D

In this Appendix are reported the identified discrete distributions of random effects obtained during the sensitivity analysis about the *MinDist* parameter, applied to the Gaussian Initialization and Uniform initialization models. We recall that the candidates for the threshold parameter are 37, ranging from 0.1 to 1 with step 0.025. Obtained distributions are partitioned in different sets, according to the regions of interest defined in the analysis of Figure 4.29 and Figure 4.31. In particular, Figures D.1 to D.4 refers to the the Gaussian initialization model, while figure D.5 to D.8 to the Uniform initialization one.

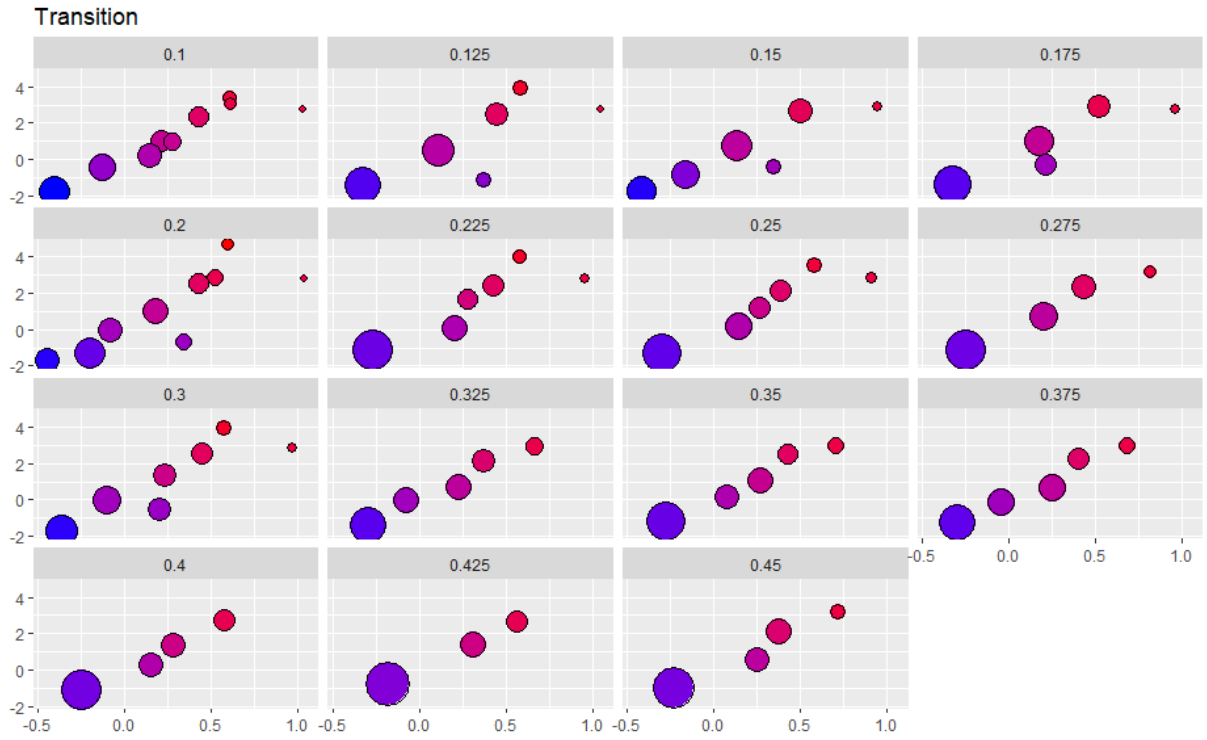

Figure D.1: Identified discrete distributions of random effects using the Gaussian Initialization model, tuned with values of the threshold parameter belonging to the *Transition* region of Figure 4.29.

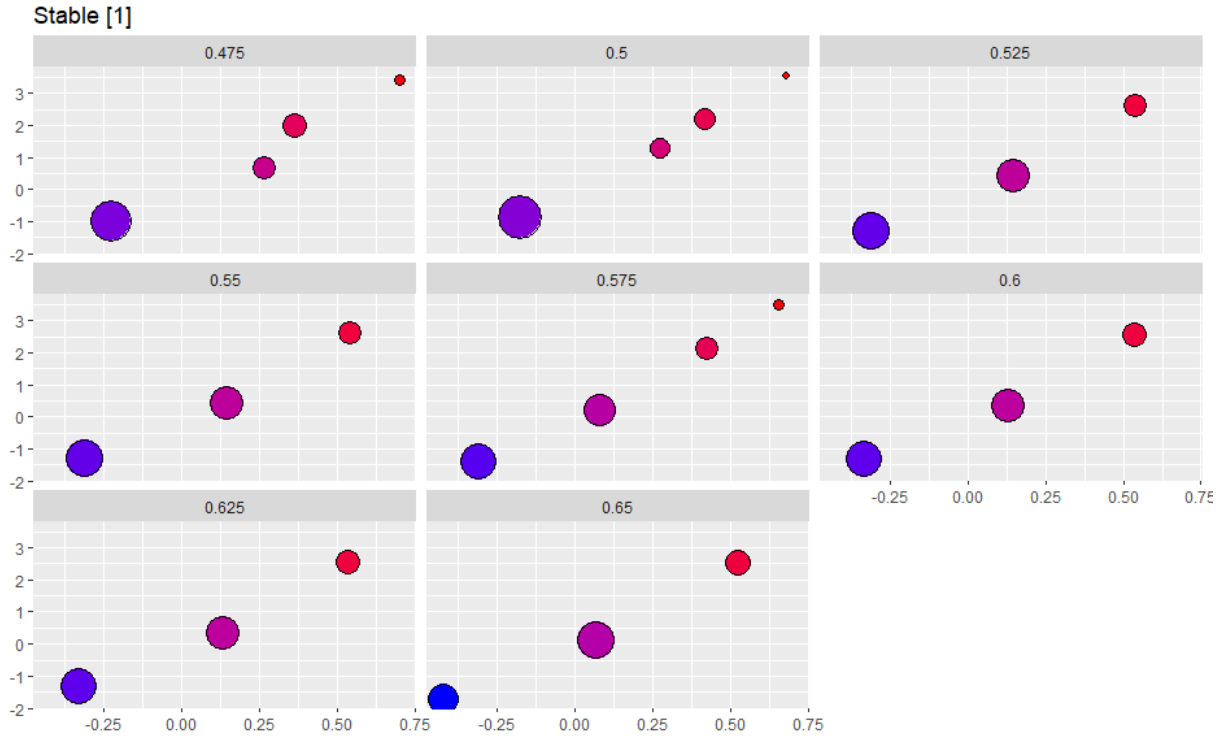

Figure D.2: Identified discrete distributions of random effects using the Gaussian Initialization model, tuned with values of the threshold parameter belonging to the *Stable[1]* region of Figure 4.29.

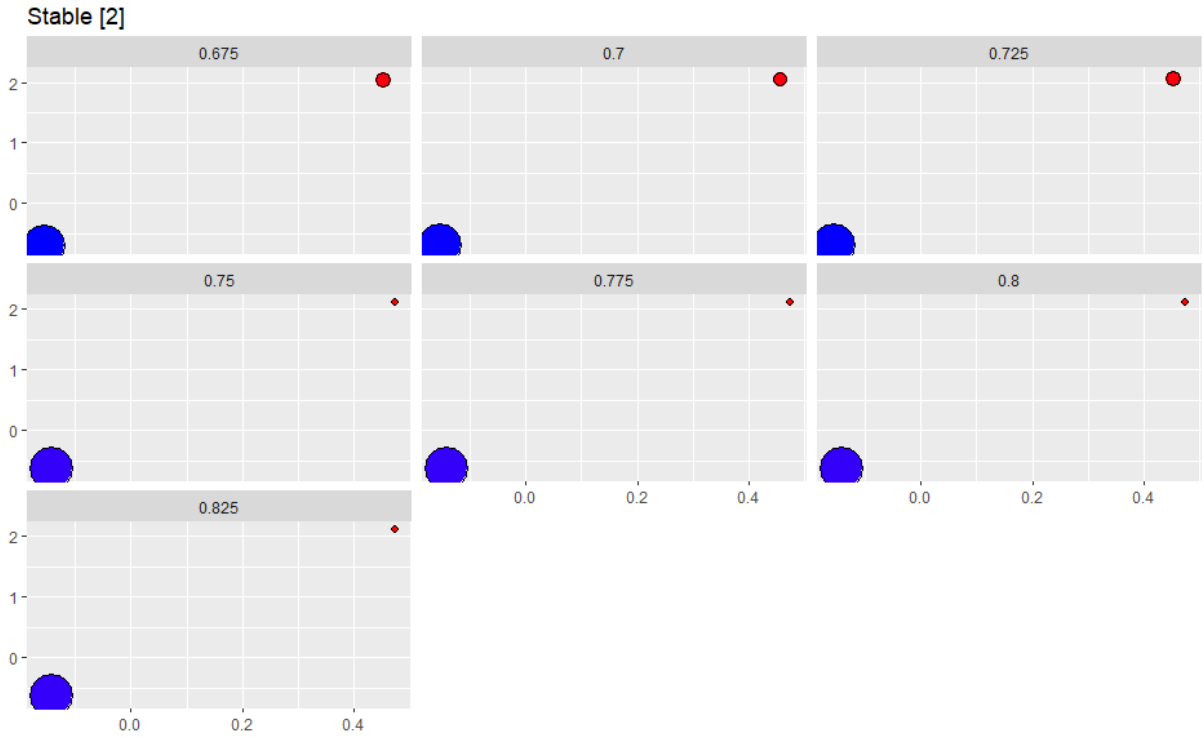

Figure D.3: Identified discrete distributions of random effects using the Gaussian Initialization model, tuned with values of the threshold parameter belonging to the *Stable[2]* region of Figure 4.29.

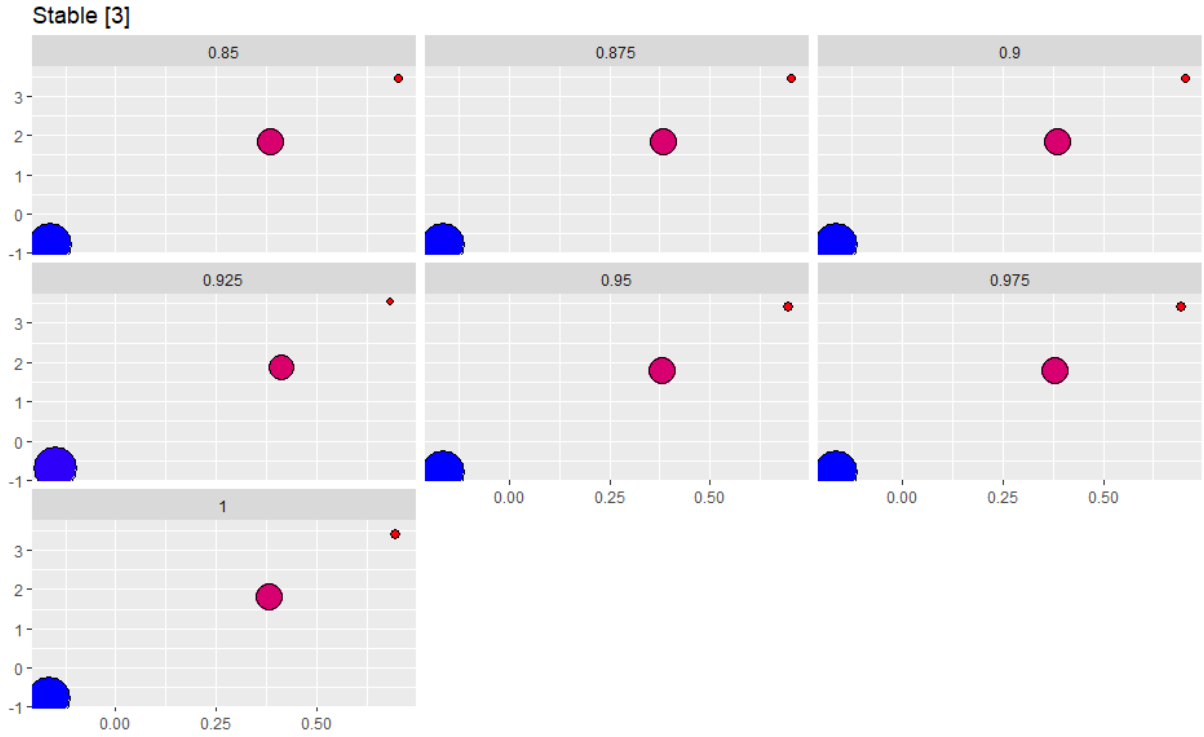

Figure D.4: Identified discrete distributions of random effects using the Gaussian Initialization model, tuned with values of the threshold parameter belonging to the *Stable[3]* region of Figure 4.29.

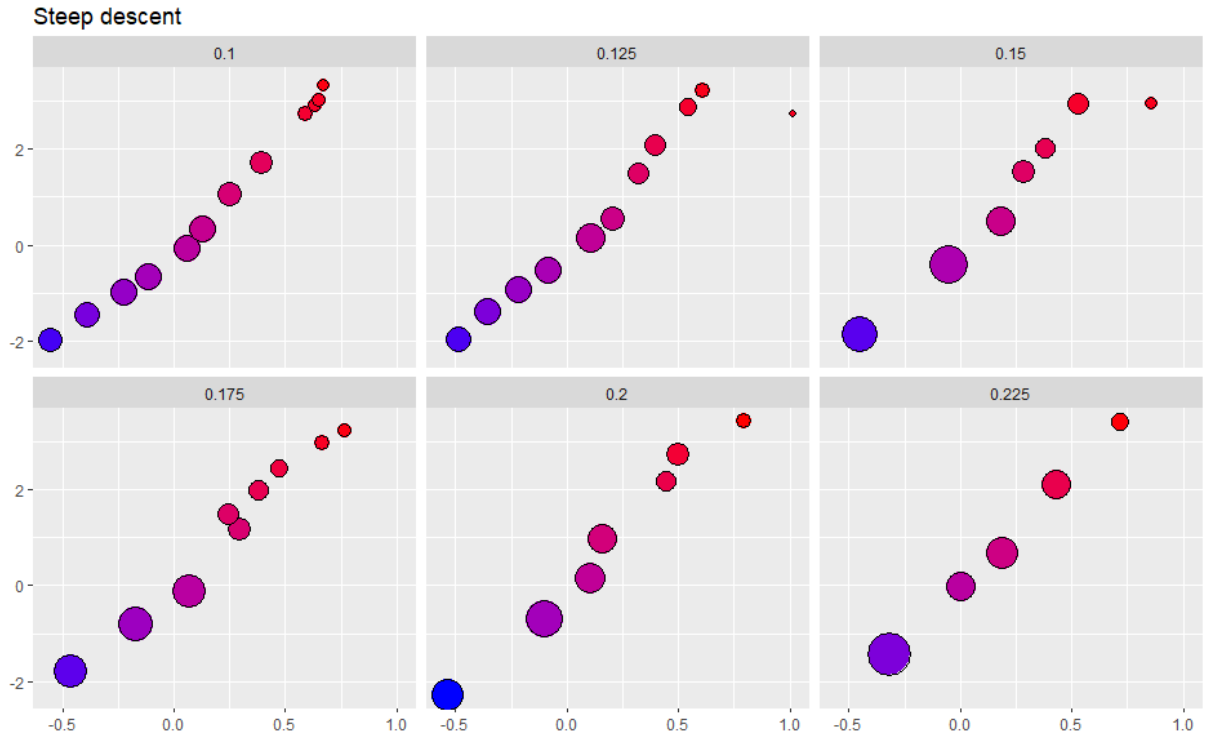

Figure D.5: Identified discrete distributions of random effects using the Uniform Initialization model, tuned with values of the threshold parameter belonging to the *Steep Descent* region of Figure 4.31.

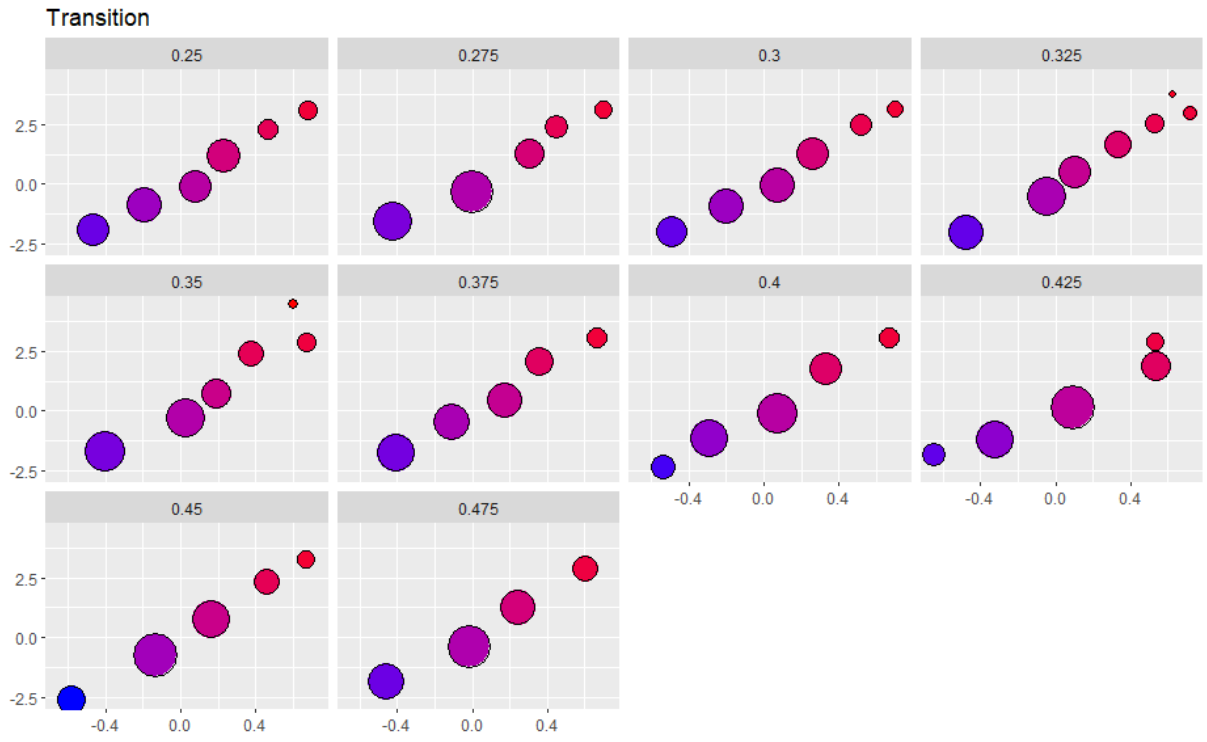

Figure D.6: Identified discrete distributions of random effects using the Uniform Initialization model, tuned with values of the threshold parameter belonging to the *Unstable* region of Figure 4.31.

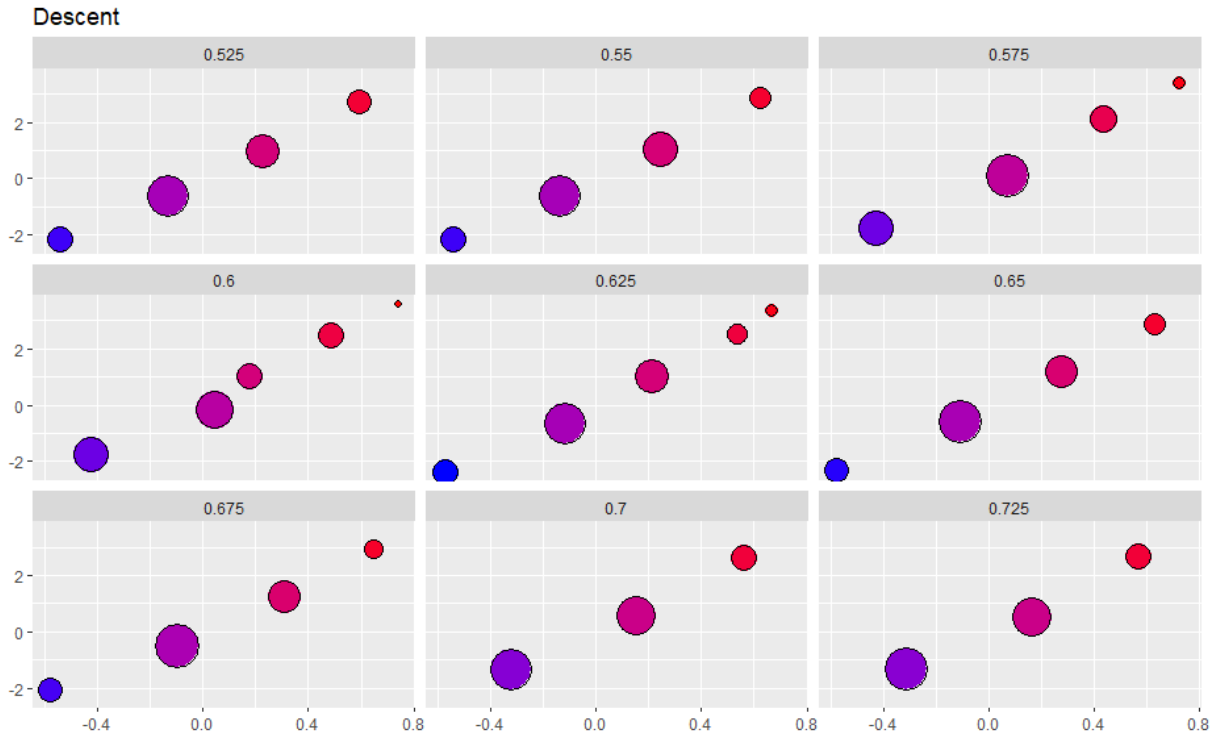

Figure D.7: Identified discrete distributions of random effects using the Uniform Initialization model, tuned with values of the threshold parameter belonging to the *Second Descent* region of Figure 4.31.

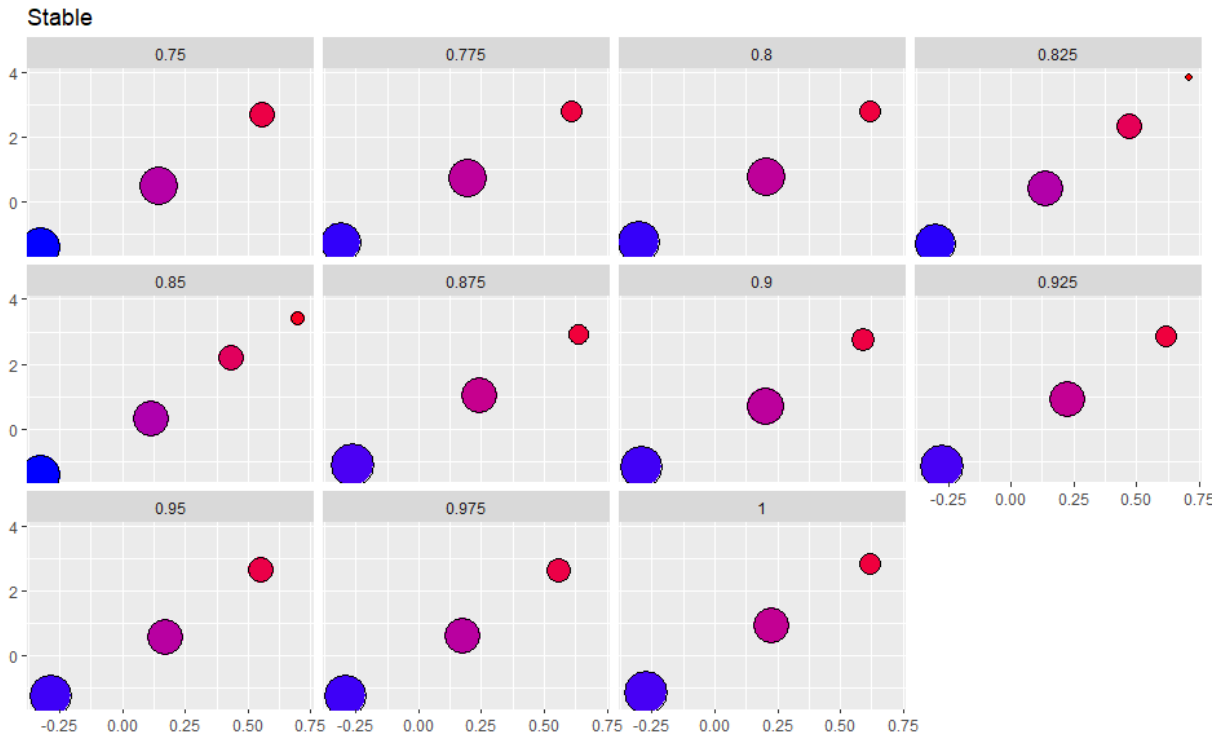

Figure D.8: Identified discrete distributions of random effects using the Uniform Initialization model, tuned with values of the threshold parameter belonging to the *Stable* region of Figure 4.31 .

List of Figures

List of Tables

Ringraziamenti

Arrivato alla fine di questo lungo percorso vorrei dedicare qualche parola a tutte le persone che lo hanno reso possibile.

Per prima cosa, vorrei ringraziare la professoressa Francesca Ieva, che ha accettato di seguirmi in questo percorso nonostante i numerosi impegni, ed insieme a lei le dottoresse Marta Spreafico e Chiara Masci. Tutte hanno mostrato una disponibilità ed un entusiasmo che raramente ho incontrato durante il mio percorso accademico, assecondando le mie idee e non facendomi mai mancare il loro supporto. In particolare, vorrei ringraziarle per avermi guidato nei momenti di difficoltà, per l'enorme aiuto che mi hanno dato nella stesura di questo scritto e il tempo che ci hanno dedicato. Lavorare con voi è stato un piacere.

Grazie a Gianluca e Milena, per avermi dato la possibilità di scegliere il mio percorso e portarlo a termine come meglio credevo. Il vostro supporto non mi è mai mancato non solo durante questi ultimi sei mesi, ma da quando non volevo scrivere "ambulanza" per una pagina intera. Vorrei ringraziarvi per aver avuto fiducia in quel ragazzino che, qualche volta, ancora non ha voglia di fare i compiti. Vorrei anche chiedere scusa sia a voi che a Marta: so bene che, in certi momenti, vivere con me in questi ultimi cinque anni non deve essere stato facile. Una dedica speciale poi va proprio a lei, perchè ora è il suo turno. Insieme a voi voglio ringraziare anche il resto della famiglia, perchè in qualsisasi occasione non avete mai fatto mancare il vostro aiuto e il vostro sorriso. Quello che sono oggi lo devo a tutti voi.

Ringrazio tutti gli amici del Poli. Con voi ho passato giornate interminabili, studiato un numero di ore che non so contare e superato tante situazioni difficili, ma quello che mi torna alla mente ora sono le risate, le serate e i bei momenti degli ultimi anni. La verità è che l'uni non l'ho ancora finita, ma sento già la vostra nostalgia. Quello di cui sono sicuro è che ci incontreremo ancora spesso al baretto, anche se forse non alle tre del pomeriggio.

Ringrazio gli amici di sempre, perchè senza di voi la mia vita sarebbe senza dubbio più triste. In questi anni ne abbiamo fatte di ogni, ed è bello sapere che c'è qualcuno sempre pronto a ridere delle tue disgrazie, prima di (forse) darti una mano.

Infine, vorrei ringraziare Martina. Vorrei ringraziarla perchè mi è sempre vicina, anche quando è difficile. Vorrei ringraziarla perchè è sempre pronta ad ascoltarmi, anche quando ha altri problemi per la testa. Vorrei ringraziarla perchè tante volte ha più fiducia in me di quanta ne abbia io stesso.

Grazie a tutti,
Riccardo