POLITECNICO DI MILANO

SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING
Department of Mathematics

Master of Science in Mathematical Engineering

Statistical modelling of adherence to drug prescription and its effects on survival in Heart Failure patients

Advisor: Prof. Francesca Ieva
Co-advisor: Dr. Francesca Gasperoni

Candidate:
Marta Spreafico
Matr. 853354

Academic Year 2016 - 2017

Contents

List of Figures

List of Tables

Sommario

Lo Scompenso Cardiaco (SC) è una malattia cronica molto diffusa che colpisce in particolare le persone al di sopra dei 65 anni di età. In Italia, è considerato la causa principale di ricovero e ogni anno vengono registrati circa 80,000 nuovi casi. L'ampia diffusione ed il relativo impatto socio-economico-sanitario di questa patologia la rendono un problema molto stimolante sia da un punto di vista clinico che amministrativo.

In questo lavoro analizziamo i dati relativi a pazienti affetti da SC provenienti dalle banche dati amministrative di Regione Lombardia e Regione Friuli Venezia Giulia, concentrando il nostro studio su un approccio innovativo volto ad investigare l'effetto che il consumo di farmaci ha sulla probabilità di sopravvivenza dei pazienti.

Questa tesi ha due principali obiettivi. Il primo consiste nell'elaborazione e nel calcolo di una variabile tempo-dipendente che rappresenti l'assunzione di farmaci nel tempo e che risulti più informativa di una semplice variabile binaria di aderenza. Il secondo riguarda lo studio di questa variabile tramite l'implementazione di un adeguato modello congiunto di dati longitudinali e tempo all'evento.

Per raggiungere questi obiettivi, in primo luogo eseguiamo un processo di selezione rispetto ai databases originali, fissando opportuni criteri di inclusione. In seguito, approfittando delle molteplici informazioni contenute nei due datasets sulla storia farmacologica dei pazienti, ricostruiamo i dati tempo-dipendenti riguardanti il consumo di farmaci. Infine, utilizziamo i dati sopra citati per studiare l'influenza del processo longitudinale che descrive l'aderenza alla terapia farmacologica sull'outcome di sopravvivenza dei pazienti, mediante innovativi modelli congiunti e di sopravvivenza

Parole chiave: Scompenso Cardiaco, consumo di farmaci, aderenza, dati amministrativi, covariate tempo-dipendenti, modelli congiunti.

Abstract

Heart Failure (HF) condition is a widespread chronic heart disease that affects people aged over 65. In Italy, it is considered the principal cause of hospitalization and about 80,000 new cases per year are recorded. The relevant presence and the corresponding socioeconomic and health impact of this disease make it a challenging issue both from a clinical and an administrative point of view.

In this work we analyse HF data collected from the administrative databases of Lombardy Region and Friuli Venezia Giulia Region, concentrating our study on an innovative approach for investigating the effect of drugs consumption on survival outcomes of patients.

This thesis has two different purposes. The first one concerns the elaboration and computation of a time-varying variable of drug assumption which results more informative than a simple binary variable for adherence. The second one concerns the investigation of this time-dependent variable, through the implementation of adequate Joint Models of longitudinal and time-to-event data.

In order to achieve these aims, firstly we perform a selection process on the original databases, defining specific inclusion criteria. Then, taking advantage of several information about patients' pharmacological history collected in our databases, we recover time-dependent data concerning drug assumption over time. Finally, we use the aforementioned data to study the influence of the longitudinal process given by adherence to pharmacological treatment on patients' survival outcomes, by means of innovative joint and survival models.

Keywords: Heart Failure, drug consumption, adherence, administrative data, time-varying covariate, Joint Models.

Introduction

According to the majority of dissertations and scientific papers from the field of pharmacoepidemiology, adherence to therapies is a critical and important issue, especially in chronic diseases. Indeed, medication nonadherence is usually associated with adverse health conditions and increased economic burden to the healthcare system [18].

This thesis is a methodological work on datasets gathering information about patients hospitalized for Heart Failure (HF) in Lombardy Region (LR) and Friuli Venezia Giulia Region (FVGR). The richness of these datasets, especially the presence of patient's pharmacological prescriptions, allow us to investigate and test adherence to treatments, applying several statistical techniques.
At the beginning of our work, we established two purposes. Firstly, since adherence is usually considered as a binary variable, we want to introduce an innovative technique for adherence measure consisting in a time-varying covariate. Pharmacoepidemiological studies help us to better understand how to take advantage of administrative databases in order to achieve the first aim of our work: compute a time-dependent variable of drug assumption that is more informative than a simple binary one.
The second purpose of this thesis concerns the investigation of this time-dependent variable, through the implementation of adequate Joint Models of longitudinal and time-to-event data. In particular, we want to find a methodology to deal with time-dependent covariates related to drug consumption, which results richer than a simple Cox's model in terms of adherence and that can also be a starting point for the development of future analyses.

The dissertation is structured as follows.
In this introduction we begin our work presenting the characteristics of HF pathology. In particular, we give a brief description of symptoms that characterize patients and the effects of this disease.
We proceed presenting some concepts and definitions of pharmacoepidemiology in Chapter 1, such as dosages, coverage days and adherence. We thought these

information and examples are necessary to have a clear overall perspective and to better understand some characteristics of data and procedures we will use in the analyses.
In Chapter 2 we show in details the features of the two administrative databases (LR and FVGR), the information furnished and the patients that are included in the study at the beginning. In particular, we highlight the way these data have been collected and rearranged in order to obtain our study cohorts of patients that meet specific criteria. Moreover, we introduce the computation of two innovative time-dependent covariates for drug assumption processes: curves of cumulative days covered by drug assumption and curves of assumed dose.
The background theory for this thesis is presented in Chapter 3. First of all, we discuss the fundamental characteristics of survival analysis and we introduce the Proportional-Hazard Cox model. Then, we describe Joint Modelling of longitudinal and time-to-event data, the methodology we use to deal with time-dependent covariates.
Once described the necessary theory, in Chapter 4 we present the results achieved in both cases, LR and FVGR. Firstly, we give the presentation and the descriptive analyses of the datasets, which are key points in the field of survival analysis. Next, we verify if our time-varying curves highlight some differences in terms of adherence, using a Functional K-mean approach. Finally, we present the results of Cox's and Joint models, comparing the use of a dichotomized variable for adherence in the first case and of a time-varying variable for drug assumption in the second one.
We conclude our work with a discussion of results that have been obtained and few proposals for future developments.

All the analyses are carried out using the `R` software [24]. In particular, we use `survival` [33] and `JM` [27] packages.

Heart Failure (HF)

Heart failure (HF) is a complex clinical syndrome caused by structural or functional cardiac disorders that impair the ability of one or both ventricles to fill with or eject blood [13]. HF can develop gradually over time, called Chronic Heart Failure, or very quickly, named Acute Heart Failure.

In general, a number of different problems usually leads to HF condition so, usually, it has not a single cause. Main causes of HF are myocardial ischemia, high blood pressure (HBP), cardiomyopathies, valvular heart disease, pulmonary hypertension (PHT) and congenital heart disease [23]. The diagnosis of HF is not an easy task due to the huge variety of symptoms of this pathology. For this purpose there are different tests that facilitate doctors in the disease diagnosis, such as blood test, electrocardiogram (ECG) and echocardiogram. Generally for HF identification, symptoms (such as shortness of breath at rest or during exertion or fatigue), signs of fluid retention (such as pulmonary congestion or ankle swelling), and objective evidence of a decrease in myocardial performance at rest are required [10].

HF is widespread all over the world, especially among people over 65 years: the mean age of HF patients in industrialized societies is approximately 75 years. The prevalence of HF can be estimated at 1%-2% in Western countries and the incidence approaches 5 to 10 per 1000 persons per year [23]. In particular, in Italy about 80,000 new cases per year are recorded [21] and it is the second cause of hospitalization, after vaginal delivery.

Moreover, mortality rate is relatively high in the first few weeks after the occurrence of HF but it presents a much more gradual slope in the following period. According to data from different studies conducted in America and Europe, 30-day, 1-year, and 5-year mortality are around 10% to 20%, 30%, and 65% respectively [23]. Continuous advances in therapy are changing the prognosis and improving survival. In fact, in the Framingham heart study, the 1-year and 5-year mortality rates from HF in men declined from 30% and 70%, respectively, in the period 1950 to 1969 to 28% and 59% in the period 1990 to 1999. In women, 1-year mortality rates decreased from 28% to 24% and the 5-year mortality rates decreased from 57% to 45% during the same period [20]. These results have been confirmed in other population-based studies [29].

The goals of treatment are reduction in symptoms, a decrease in the rate of hospitalizations and the prevention of premature death. The cornerstone of treatment is pharmacological therapies: the most used are ACE Inhibitors, Beta Blocking agents, Angiotensin Receptor Blockers, Anti Aldosterone agents and Diuretics. A fair combination of lifestyle changes and adherence to pharmacological treatments can lead to a good disease monitoring.

For all these data, studying HF condition can lead to healthcare improvements, social benefits and economic utilities.

Chapter 1

Pharmacoepidemiology

In this Chapter we introduce some pharmacological concepts and definitions that are useful to better understand some characteristics of data and procedures used in the following analysis. First of all, in Section 1.1 we explain what Drug Utilization Research is and which are its goals. In Section 1.2 we present the concept of dosage. Then, in Sections 1.3 and 1.4 we show how to determine some fundamental information about drug and we illustrate a method to calculate the coverage days. Finally, in Section 1.5 we explain the concepts of adherence and Proportion of Days Covered.

1.1 Why Drug Utilization Research?

Pharmacoepidemiology applies epidemiological methods to the study of the clinical use of drugs in populations. It is defined in [30] as:

Definition 1.1.1 (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.*

The branch of pharmacoepidemiology that deals with the use of drugs is known as Drug Utilization Research (DUR). As mentioned in [34], it was defined in 1977 by the *World Health Organization* (WHO):

Definition 1.1.2 (Drug Utilization Research). *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 goal of DUR is to facilitate the rational use of drugs in patients populations; therefore the prescription of drugs must be in the optimal dose for the therapeutic indication, with the correct information and at an affordable price.

Studies of DUR provide information about:

1. *Profile of drugs use* (how much the drugs are used - trends in the use and cost of drugs over time)

2. *Quality of drugs use* (drug choice - costs - dosages - interactions)

3. *Determinants of drugs use*(users characteristics - prescriber's characteristics)

For these purposes two different types of databases, administrative and clinical ones, may serve. Administrative databases usually give information about personal data and clinical characteristics, dynamics of prescriptions and survival outcomes for admission in hospital. Clinical databases usually contains details about habits and lifestyles, diagnostic route and intermediate clinical outcomes.
The use of both administrative (dynamics of prescriptions) and clinical (diagnostic route) databases allow to measure the effective drug utilization with a big limitation: we are not able to assert if the patient is currently consuming the dispensed drug.

Among all the aspects mentioned above, in this thesis we focus on:

- users characteristics

- drug choice

- how much the drugs are used

- dosages

- trends in the drugs use over time

1.2 Dosage

In order to evaluate the use of a drug we need a statistical measure of drug consumption. The *World Health Organization* in [34] defines different types of drug dosages and the most used are the Defined Daily Dose (DDD) and the Prescribed Daily Dose (PDD).

Definition 1.2.1 (Defined Daily Dose)**.** *The defined daily dose (DDD) is the assumed average maintenance dose per day for a drug used for its main indication in adults.*

Definition 1.2.2 (Prescribed Daily Dose)**.** *The prescribed daily dose (PDD) is defined as the average dose prescribed according to a representative sample of prescriptions.*

DDD is used to standardize the comparison of drug usage between different drugs or between different healthcare environments. Furthermore DDD is a unit of measurement and it does not necessarily correspond to the recommended or prescribed daily dose: while DDD is fixed over time, PDD can change according to different prescriptions.

As we said before, it should be noted that these dosages do not necessarily reflect actual drug utilization. Indeed a limit, which is common to DUR, consists in the impossibility of knowing the real drug consumption: the patient does not always take all the medications that are dispensed and we cannot know if it happens or not.

1.3 Available information about drugs

Drugs are classified in different ways, based on the information they can give. First of all, there exist three different bands of drugs that tell us if a drug is free (band A, for chronic diseases) or not (band C) or only for hospital use (band H). However, the most important characteristics are connected to other factors, like therapeutic and pharmacological properties or information related to the specific medicine box, that are given by the so called ATC and AIC codes.

1.3.1 ATC codes

The Anatomical Therapeutic Chemical (ATC) classification system is used for the classification of active substances of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is controlled by the *World Health Organization Collaborating Centre for Drug Statistics Methodology* (WHOCC) and it was first published in 1976.

According to the ATC classification system, drugs are classified in groups at five different levels:

1. The first level indicates the anatomical main group and consists of one letter
2. The second level indicates the therapeutic subgroup and consists of two digits
3. The third level indicates the therapeutic/pharmacological subgroup and consists of one letter

4. The fourth level indicates the chemical/therapeutic/pharmacological subgroup and consists of one letter

5. The fifth level indicates the chemical substance and consists of two digits

Table 1.1 shows the first level of ATC code, consisting of one letter corresponding to the anatomical main group, and an example of ATC code is reported in Table 1.2.

Code	Contents
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito-urinary system and sex hormones
H	Systemic hormonal preparations, excluding sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

Table 1.1: First Level of ATC code

Pharmaceutical prescriptions regarding therapies of interest are identified extracting levels of ATC code of each record. In particular we focus our work on five different pharmacological classes: ACE-Inhibitors (ACE), Angiotensin Receptor Blockers (ARB), Beta-Blocking agents (BB), Anti-Aldosterone agents (AA) and Diuretics (DIU), whose corresponding ATC levels are shown in Table 1.3.

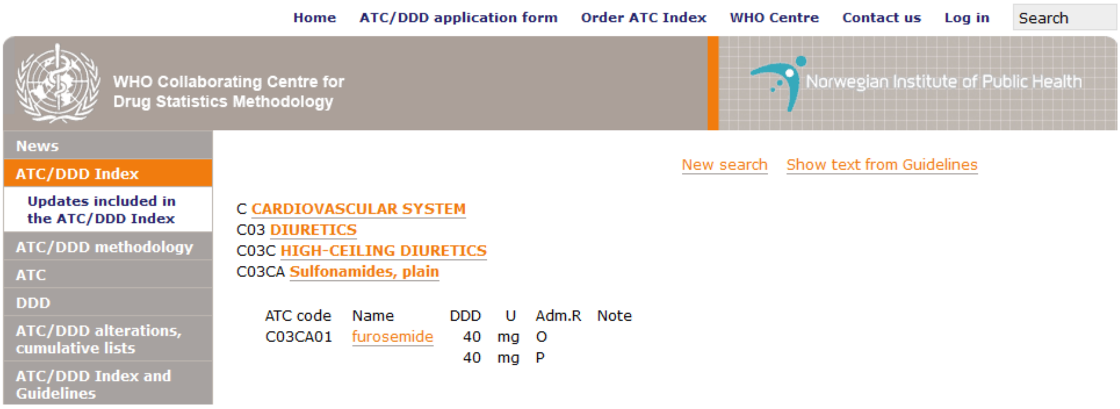
Level	Code	Contents
First	C	Cardiovascular system
Second	C03	Diuretics
Thirs	C03C	High-ceiling diuretics
Fourth	C03CA	Sulfonamides
Fifth	C03CA01	Furosemide

Table 1.2: Example of ATC code and corresponding levels of *Furosemide*, a diuretic used for the treatment of edema and water retention.

Type of drug	ATC levels
ACE-Inhibitors	level 3 $\in$ {C09A, C09B, C09X}
Angiotensin Receptor Blockers	level 3 $\in$ {C09C, C09D}
Beta-Blocking agents	level 2 = C07
Anti-Aldosterone agents	level 3 $\in$ {C03D, C03E}
Diuretics	level 3 = C03C or ATC = C03BA08

Table 1.3: Levels of ATC code for different types of drugs used in our analysis.

The ATC system gives us not only the type of drug but it also includes defined daily doses (DDDs) and routes of administration for drugs containing molecules of only one pharmacological class. Indeed, using the informations available on the WHO's website (`https://www.whocc.no/atc_ddd_index/`), for each ATC code we can find the corresponding DDD. An example is shown in Figure 1.1: for the ATC code C03CA01 the corresponding DDD is 40mg with two possible routes of administration, oral (O) or parenteral (P).


Home
ATC/DDD application form
Order ATC Index
WHO Centre
Contact us
Log in
Search
WHO Collaborating Centre for Drug Statistics Methodology
Norwegian Institute of Public Health
News
ATC/DDD Index
Updates included in the ATC/DDD Index
ATC/DDD methodology
ATC
DDD
ATC/DDD alterations, cumulative lists
ATC/DDD Index and Guidelines
New search
Show text from Guidelines

C CARDIOVASCULAR SYSTEM
C03 DIURETICS
C03C HIGH-CEILING DIURETICS
C03CA Sulfonamides, plain

ATC code	Name	DDD	U	Adm.R	Note
C03CA01	furosemide	40	mg	O	
		40	mg	P	

Figure 1.1: Example of *WHO*'s website use for ATC code C03CA01.

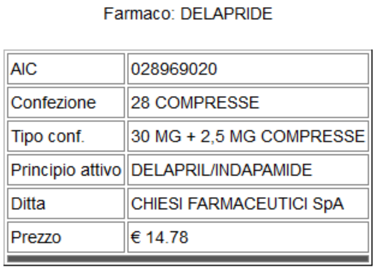
1.3.2 AIC codes

The AIC code (Autorizzazione all'Immissione in Commercio) is a marketing authorization code issued by the Italian Medicines Agency (AIFA - Agenzia Italiana del Farmaco) that identifies every drug box on the pharmacological market in Italy in a unique way. This means that different formats of the same drug have different AIC codes. It consists in a numeric code of nine digits and it is important because it allows us to go back to some fundamental characteristics about drug, such as:

- number of tablets in one box

- *mg* in one tablet
- percentage of the first active principle in one tablet
- route of administration
- cost of a box

For example, the box of drug *Delapride*, with AIC code = 028969020, is composed[1] of 28 tablets of $32.5mg$ where the first active principle has a weight of $30mg$, as shown in Figure 1.2.



Farmaco: DELAPRIDE

AIC	028969020
Confezione	28 COMPRESSE
Tipo conf.	30 MG + 2,5 MG COMPRESSE
Principio attivo	DELAPRIL/INDAPAMIDE
Ditta	CHIESI FARMACEUTICI SpA
Prezzo	€ 14.78

Figure 1.2: Website results for AIC code 028969020, which corresponds to *Delapride*, a drug used for the hypertension treatment.

1.4 Duration of a prescription

The duration of a prescription is a really important information because it indicates the coverage days, that correspond to the period in which the patient consumes the prescribed drug. For example, if we have a prescription in date "2018-01-01" with a duration of 15 days we know that the patient assumes the drug from "2018-01-01" to "2018-01-15".

It could happen that data about the duration of a prescription are not available in the administrative dataset but in some cases we can recover it. In order to do that we need:

- daily dose
- number of tablets in one box

[1] `http://www.blia.it/utili/farmacia/index.php?aic=028969020` (in this case the reference AIC code is 028969020)

- *mg* of the first active principle in one tablet
  (if there are more than one active principle we consider only the first that, i.e., the most important one)

For the daily dose we use PDD if available, otherwise DDD. The other two characteristics can be obtained from the AIC code so its presence in the dataset is essential. Using all this information we have:

Definition 1.4.1 (Coverage days)**.** *The coverage days is the number of days covered by a single box, which is given by*

$$\textit{coverage days} = \frac{\textit{number of tablets} \cdot \textit{mg of the first active principle}}{\textit{daily dose}} \tag{1.1}$$

For example, suppose that a patient has a prescription for *Delapride* with AIC code equal to 028969020 and a PDD of $20mg$ but the duration is not available. Using Equation (1.1) and data available in Figure 1.2, we can calculate that the coverage days of a box are 42:

$$\text{coverage days} = \frac{28 \cdot 30mg}{20mg} = 42$$

1.5 Adherence

In addition to evaluate the assumed drug quantity, we want also to establish if the drug was taken continuously during all the follow up period. In order to do that, we consider adherence to prescribed medications which is a key factor in effective disease management of many chronic conditions. The term adherence generally means if a patient follows or not the prescribed treatment [4]:

Definition 1.5.1 (Adherence)**.** *Adherence (or compliance) generally refers to whether a patient takes a prescribed medication according to schedule.*

There exist lots of different adherence measures. According to [18], the two best ones are the Proportion of Days Covered (PDC) and the Medical Possession Ratio (MPR):

Definition 1.5.2 (Proportion of Days Covered)**.** *The Proportion of Days Covered is defined as*

$$PDC = \frac{\textit{number of distinct coverage days}}{\textit{number of days in the observation period}} \in [0,1] \tag{1.2}$$

Definition 1.5.3 (Medical Possession Ratio)**.** *The Medical Possession Ratio is defined as*

$$MPR = \frac{\textit{number of days supply during observation period}}{\textit{number of days in the observation period}} \in \mathbb{R}^+ \tag{1.3}$$

The term "distinct" in (1.2) underlines the fact that, in case of overlapping periods, PDC considers the first event entirely and only the days of the second one not covered by the first. Conversely, MPR shifts the second event at the day after the end of the first one, preserving all its duration.
For our work we decide to use PDC and an observation period of 365 days, as done in [18].

Finally we use PDC to determine:

- *Adherent patients*
  Adherence measure can be categorized into two levels for which a patient is considered adherent if he reaches a certain level of the measure, that we set at 0.80:

$$\begin{aligned} \text{adherent} = 1 \quad &\longleftrightarrow \quad 0.80 \leq \text{PDC} \leq 1 \\ \text{adherent} = 0 \quad &\longleftrightarrow \quad 0 \leq \text{PDC} < 0.80 \end{aligned}$$

- *Adherence levels*
  Adherence measure can be categorized into four levels based on PDC value:

$$\begin{aligned} \text{level} = 1 \quad &\longleftrightarrow \quad 0 \leq \text{PDC} < 0.25 \\ \text{level} = 2 \quad &\longleftrightarrow \quad 0.25 \leq \text{PDC} < 0.50 \\ \text{level} = 3 \quad &\longleftrightarrow \quad 0.50 \leq \text{PDC} < 0.75 \\ \text{level} = 4 \quad &\longleftrightarrow \quad 0.75 \leq \text{PDC} \leq 1 \end{aligned}$$

1.5.1 Example of PDC calculation

Consider two different patients, A and B, with corresponding coverage days shown in Figures 1.3 and 1.4. We can observe that patient A has four different prescriptions without any overlaps that correspond to three coverage periods. On the other hand, patient B presents six different prescriptions with overlaps between first and second and between fourth and fifth prescription. Considering only distinct days,

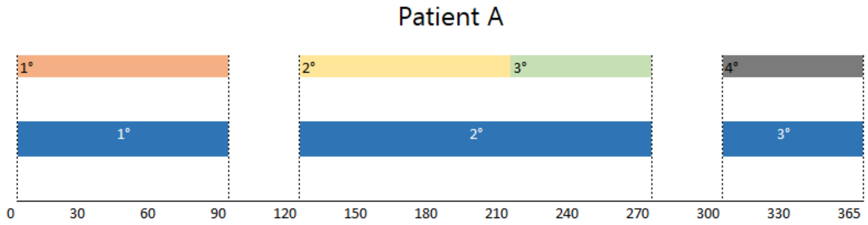


Figure 1.3: Patient A: example without overlaps

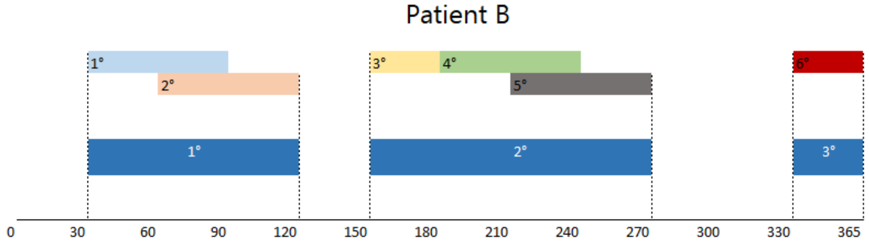


Figure 1.4: Patient B: example with overlaps

we obtain the three periods of coverage days given by the blue bands.
Using Equation (1.2) we have that

$$\text{PDC patient A} = \frac{90+150+65}{365} = \frac{305}{365} = 0.8356$$

$$\text{PDC patient B} = \frac{90+120+35}{365} = \frac{245}{365} = 0.6712$$

Consequently patient A is adherent with level of adherence equal to 4 and patient B is not adherent with a level of adherence equal to 3.

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.

Chapter 2

Datasets

In this Chapter we describe the two administrative databases analysed within this thesis work. In particular, in Section 2.1 we describe the main steps performed in order to collect and rearrange the data from Lombardy Region (LR), whereas in Section 2.2 we describe the same details about data arising from Friuli Venezia Giulia Region (FVGR).

2.1 Lombardy Region dataset

In the Lombardy Region (LR) dataset information about patients hospitalized for HF from 2000 to 2012, as described in [22], are collected. For our work, we used a representative sample composed by 1,333,954 events related to 4,872 patients with their first HF hospitalization during the period 2006-2012.

2.1.1 Variables

For each patient, identified by its unique anonymous ID code, we have some characteristics like age at each event, gender, date of enrolment and the state at the end of the study. Administrative censoring date is December 31st, 2012 ("2012-12-31"). The final state may be *dead* if the patient death occurs before the end of the study, *truncated* if censored, *lost* if lost to follow up. All these variables are reported in Table 2.1.

Each record in the dataset is related to an event, which can be an hospitalization or a pharmacological event. In the first case the date of discharge from hospital and the length of stay in hospital are given. In the second one we know the date of prescription for drugs and the number of days of treatment covered by the prescriptions. All these variables are reported in Table 2.2.

Moreover, further information is available about patients' medical history when events consist of HF hospitalizations, such as particular procedures and a list of comorbidities (see [22] and [11]), as shown in Tables 2.3 and 2.4, respectively.

Variable	Description
COD_REG	Anonymous ID code of each patient
age	Patient's age
gender	Patient's gender
data_rif_ev	Date of first discharge for HF event
data_studio_out	Date of death/censoring
desc_studio_out	Label at the end of the study

Table 2.1: Patient's information in LR dataset.

Variable	Description	
	Hospitalization event	Pharmacological event
tipo_prest	Hospitalization	Pharmaceutical prescription
class_prest	CCS-principal diagnosis (Clinical Classification Software by CMS)	ATC code
data_prest	Date of discharge	Date of prescription
qta_prest_Sum	Length Of Stay in hospital	Duration of the prescription

Table 2.2: Event information in LR dataset.

Variable	Description
ICD	Binary flag which marks if patient has received an Implantable Cardioverter Defribillator
SHOCK	Binary flag which marks if patient has had a circulatory shock
CABG	Binary flag which marks if patient went through a Coronary Artery Bypass Surgery
PTCA	Binary flag which marks if patient has received a Percutaneous Transluminal Coronorary Angioplasty

Table 2.3: Variables for hospitalization procedures in LR dataset.

It is important to observe that AIC codes are not available so the presence of the duration of each prescription, given by qt_prest_Sum in Table 2.2, is fundamental for our analysis, as it will be used instead of the coverage days computation through AIC (Section 1.4).

Variable	Description
`metastatic`	Binary flag which marks the presence of metastasis as a comorbidity
`chf`	Binary flag which marks the presence of CHF as a comorbidity
`dementia`	Binary flag which marks the presence of dementia as a comorbidity
`renal`	Binary flag which marks the presence of renal related issues as a comorbidity
`wtloss`	Binary flag which marks the presence of weight loss as a comorbidity
`hemiplegia`	Binary flag which marks the presence of hemiplegia as a comorbidity
`alcohol`	Binary flag which marks the presence of alcohol use disorders
`tumor`	Binary flag which marks the presence of tumours as a comorbidity
`arrhythmia`	Binary flag which marks the presence of arrhythmia as a comorbidity
`pulmonarydz`	Binary flag which marks the presence of one or more pulmonary diseases as a comorbidity
`coagulopathy`	Binary flag which marks the presence of coagulopathy as a comorbidity
`compdiabetes`	Binary flag which marks the presence of diabetes as a comorbidity
`anemia`	Binary flag which marks the presence of anemia as a comorbidity
`electrolytes`	Binary flag which marks the presence of electrolytes related issues as a comorbidity
`liver`	Binary flag which marks he presence of liver issues as a comorbidity
`pvd`	Binary flag which marks the presence of peripheral vascular disease as a comorbidity
`psychosis`	Binary flag which marks the presence of psychosis as a comorbidity
`pulmcirc`	Binary flag which marks the presence of pulmonary circulation issues as a comorbidity
`hivaids`	Binary flag which marks the presence of HIV/AIDS as a comorbidity
`hypertension`	Binary flag which marks the presence of hypertension as a comorbidity

Table 2.4: Variables for hospitalization comorbidities in LR dataset.

2.1.2 Patients and events selection

The first step of our analysis consists in selecting the proper cohort of patients. According to selection criteria reported in [22], we decide to consider as reference time the date of the first HF discharge and not of admission in order to exclude those patients who died during the first hospitalization. Therefore we select patients with the first discharge for HF before the censoring date and who survived at least one year because, as mentioned in Section 1.5, we are interested in investigating one year of adherence.

Regarding events, we keep only pharmacological events related to ACE, ARB, BB, AA and DIU and hospitalizations. To select only these specific pharmacological classes we consider the ATC codes, as we have explained in Section 1.3.1 (Table 1.3). In order to do that we change the global variable `class_prest` (Table 2.2) into a variable valid only for pharmacological events, that we called `ATC`, and we introduce another categorical feature, named `classe_pharma`, which indicate the type of drug (*ACE-Inhibitors, Angiotensin Receptor Blockers, Beta-Blocking agents, Anti-Aldosterone agents* and *Diuretics*). Then, since we want a follow up period of one year, we select only those events within 365 days. At the end we consider only those patients with at least one hospitalization and one pharmacological event.

We end up with a final dataset of 94,151 events corresponding to 4,406 patients. All the steps of this procedures are outlined in Figure 2.1.

In particular, in Figure 2.1 we observe that selecting event within one year of follow up we delete 518,901 events. This means that only the 27.3% (174,882 events) of the 639,783 events collected in our dataset are within the first year. Moreover, the last step of the procedure is needed to exclude those patients without events within one year of follow up or without pharmacological events of ACE, ARB, BB, AA and DIU.

2.1.3 Adding auxiliary variables

At this point we need to introduce some variables which are necessary to develop our analysis. They are reported in Table 2.5.

First of all, since our final aim is to perform a survival analysis (see Chapter 3), we need the follow up time, named `timeOUT`, given by the difference between the date of death/censoring and the date of the first discharge for HF, and the binary flags which marks if, at the end of the study, a patients is dead or not (for this

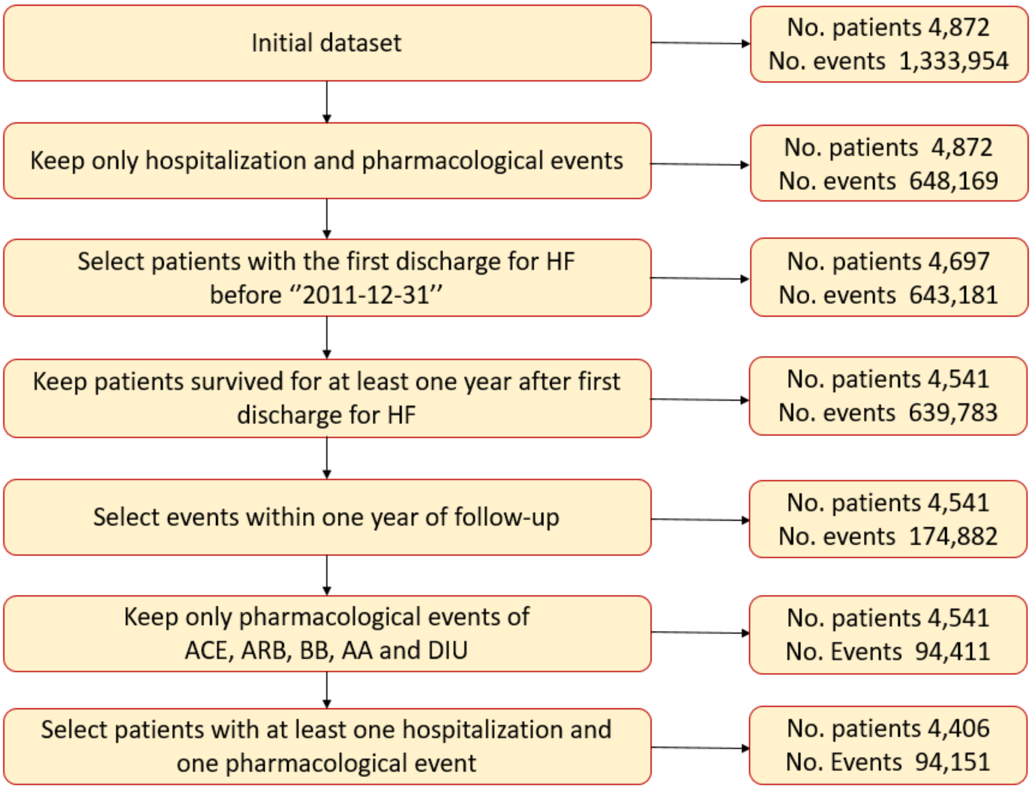

Initial dataset
No. patients 4,872
No. events 1,333,954
Keep only hospitalization and pharmacological events
No. patients 4,872
No. events 648,169
Select patients with the first discharge for HF before "2011-12-31"
No. patients 4,697
No. events 643,181
Keep patients survived for at least one year after first discharge for HF
No. patients 4,541
No. events 639,783
Select events within one year of follow-up
No. patients 4,541
No. events 174,882
Keep only pharmacological events of ACE, ARB, BB, AA and DIU
No. patients 4,541
No. Events 94,411
Select patients with at least one hospitalization and one pharmacological event
No. patients 4,406
No. Events 94,151

Figure 2.1: Patients and events selection procedure for LR dataset.

last purpose we renamed `desc_studio_out` in Table 2.1 into `labelOUT`).

$$\texttt{timeOUT} := \texttt{data\_studio\_out} - \texttt{data\_rif\_ev}$$

$$\texttt{death} := \begin{cases} 1 & \text{if } \texttt{labelOUT} = \text{dead} \\ 0 & \text{otherwise} \end{cases}$$

Then we devide `qt_prest_Sum` into two new variables, `LOS` and `qt_pharma`, which respectively indicate the length of stay in hospital for hospital admissions and the number of days of treatment covered by the prescriptions for pharmacological events. Furthermore, since for hospitalizations `data_prest` is the date of discharge from hospital, we introduce `dataADM` which is the data of admission in hospital and it is given by:

$$\texttt{dataADM} := \texttt{data\_prest} - \texttt{LOS}$$

We also insert two other variables, `hosp` and `pharm`, for the index of hospitalization and pharmacological event. Finally, using ATC codes and WHO's website as explained in Section 1.3.1, we add the Defined Daily Dose (`DDD`) which is not available for combinations (i.e. drugs containing molecules of several pharmacological classes). Therefore we also introduce a binary flag, named `COMBO`, which marks if the drug is a combination of other drugs or not.

Variable	Description
`ATC`	ATC code
`classe_pharma`	Type of drug (ACE, ARB, BB, AA, DIU)
`timeOUT`	Follow up time
`hosp`	Index of hospitalization
`pharm`	Index of pharmacological event
`LOS`	Lenght of stay in hospital
`qt_pharma`	Number of days of treatment covered by the prescriptions
`dataADM`	Date of admission in hospital
`death`	Binary flag which marks if the patient is dead
`DDD`	Defined Daily Dose
`COMBO`	Binary flag which marks if the drug is a combination

Table 2.5: Added variables in LR dataset.

2.1.4 Adherence variables

Since we want to establish if the drug was taken continuously during all the follow up period, we insert some adherence variables, summarized in Table 2.6. We remind that the reference date of each patient is the one of the first discharge from hospital so, in adherence computation, we do not consider the first hospitalization.

First of all, for each patient we compute the number of distinct coverage days during an observation period of 365 days and we call it `ADERENZA`. Dividing this last data by 365 (number of days in the observation periods) like in Equation (1.2), we obtain a new variable, `PDC`, which represent the Proportion of Days Covered and it is a number between 0 and 1. Then, as explained in Section 1.5, we determine adherent patients and adherence levels, respectively named `ADERENTE` and `PDC_CLA`, in this way:

$$\mathtt{ADERENTE} = \begin{cases} 1 & \text{if} \quad 0.8 \leq \mathtt{PDC} \leq 1 \\ 0 & \text{if} \quad 0 \leq \mathtt{PDC} < 0.8 \end{cases}$$

$$\mathtt{PDC\_CLA} = \begin{cases} 1 & \text{if} \quad 0 \leq \mathtt{PDC} < 0.25 \\ 2 & \text{if} \quad 0.25 \leq \mathtt{PDC} < 0.5 \\ 3 & \text{if} \quad 0.5 \leq \mathtt{PDC} < 0.75 \\ 4 & \text{if} \quad 0.75 \leq \mathtt{PDC} \leq 1 \end{cases}$$

Variable	Description
ADERENZA	Coverage days during observation period
PDC	Proportion of Days Covered
ADERENTE	Binary flags which marks if a patient is adherent
PDC_CLA	Adherence class (or level)

Table 2.6: Adherence variables for LR dataset.

2.1.5 Curve of cumulative days covered by drug assumption

At this point, we decide to include a time-dependent variable which, at time t, indicates the total days covered by the type of drug up to that time. Potentially there are five different curves for each patient, one for each type of drug (ACE, ARB, BB, AA and DIU) depending on which drugs he/she assumes. As we have mentioned in Sections 1.5.1 and 2.1.4, we set an observation period of 365 days and, in case of overlapping periods, we consider only distinct days. Furthermore, we hypothesize that all the prescribed types of drug are assumed by patients during the whole period of hospitalization and we do not consider the first hospitalization because the reference date of each patient is the one of the first discharge from hospital.

In order to better explain this concept, we report a real example in Tables 2.7, 2.8 and 2.9. This is concerned with a male patient whose curves of drugs assumption are reported in Figure 2.2. Considering this case, corresponding to COD_REG = 10006065 with data reported in Table 2.10, we note that his observation period ranges from data_rif_ev = "2006-08-29" to 365 days later, that is "2007-08-28". Then, from classe_pharma, we can observe that this patient assumes three different types of drug: AA (four prescription), ACE (nine prescriptions) and DIU (six prescriptions). Therefore we have to calculate three different curves.

About Diuretics (DIU), there are six prescriptions and one hospitalization which cover seven different periods, as shown in Table 2.7. We observe that, since there is an overlap between the third and the fourth, we have to shift the staring date of event no.4 in order to consider only distinct days. The resulting curve is the blue one in Figure 2.2.

No.	Event	Start	End	Days	Overlaps
1	hosp=2	"2006-10-24"	"2006-10-26"	3	
2	pharm=8	"2006-10-31"	"2006-12-07"	38	
3	pharm=10	"2006-12-18"	"2007-01-24"	38	
4	pharm=12	~~"2007-01-19"~~	"2007-02-25"	~~38~~	6
		"2007-01-25"	"2007-02-25"	31	
5	pharm=14	"2007-03-20"	"2007-04-11"	23	
6	pharm=16	"2007-05-22"	"2007-06-13"	23	
7	pharm=18	"2007-06-26"	"2007-08-02"	38	

Table 2.7: Medical history of Diuretics of patient 10006065.

On the other hand, about ACE inhibitors (ACE), there are nine prescriptions and one hospitalization which cover ten different periods, as shown in Table 2.8. We observe that, also in this case, there are several overlaps so we have to shift dates in order to consider only distinct days. Furthermore, events no.8-9-10 are not reported because at event no.7 we get the end of the observation period ("2007-08-28"). The resulting curve is the red one in Figure 2.2.

No.	Event	Start	End	Days	Overlaps
1	pharm=2	"2006-08-29"	"2006-10-23"	56	
2	pharm=4	~~"2006-09-15"~~	"2006-11-09"	~~56~~	39
		"2006-10-24"	"2006-11-09"	17	
3	pharm=7	~~"2006-10-17"~~	"2006-12-11"	~~56~~	23
		"2006-11-10"	"2006-12-11"	33	
4	hosp=2	~~"2006-10-24"~~	~~"2006-10-26"~~	~~3~~	3
		Total	overlap	0	
5	pharm=9	~~"2006-10-31"~~	"2007-06-11"	~~224~~	42
		"2006-12-12"	"2007-06-11"	182	
6	pharm=11	~~"2006-12-18"~~	"2007-07-29"	~~224~~	176
		"2007-06-12"	"2007-07-29"	48	
7	pharm=13	~~"2007-01-19"~~	~~"2007-08-30"~~	~~224~~	192
		"2007-07-30"	"2007-08-28"	29	
8	pharm=15			0	
9	pharm=17			0	
10	pharm=19			0	

Table 2.8: Medical history of ACE Inhibitors of patient 10006065.

Finally, about Anti Aldosterone agents (AA), there are four prescriptions and one hospitalization which cover five different periods, as shown in Table 2.8. We observe that, also in this case, there are several overlaps so we have to shift dates in order to consider only distinct days. The resulting curve is the orange one in Figure 2.2.

No.	Event	Start	End	Days	Overlaps
1	pharm=1	"2006-08-29"	"2006-09-17"	20	
2	pharm=3	~~"2006-09-15"~~	"2006-10-04"	~~20~~	3
		"2006-09-18"	"2006-10-04"	17	
3	pharm=5	"2006-10-06"	"2006-10-25"	20	
4	pharm=6	~~"2006-10-17"~~ "2006-10-26"	"2006-11-05" "2006-11-05"	~~20~~ 11	9
5	hosp=2	~~"2006-10-24"~~	~~"2006-10-26"~~	~~3~~ 0	3

Table 2.9: Medical history of Anti Aldosterone agents of patient 10006065.

We notice that all curves are monotone and non-decreasing, as they are expected being the cumulative sum function of the drug assumption on 365 consecutive days: each day can be covered (in this case the added value is 1) or not (in this case the added value is 0) so the number of accumulated days cannot decrease during time. A more detailed description of the curves of cumulative days covered by drug assumption for LR patients is given in Appendix B.1.

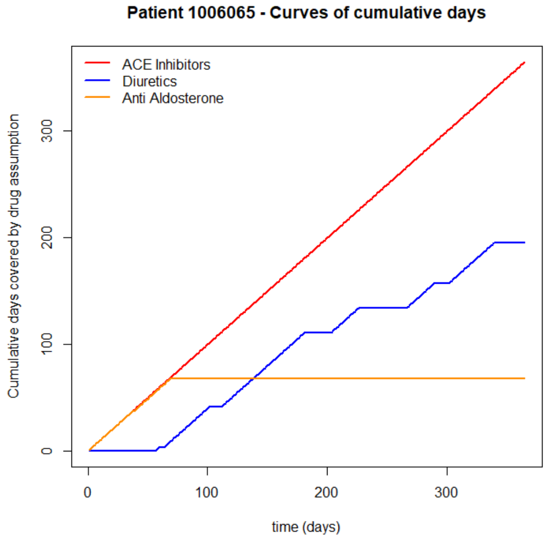


Figure 2.2: Curves of cumulative days covered by drug assumption of patient 10006065 for ACE (red), DIU (blue) and AA (orange).

	COD_REG	data_rif_ev	data_studio_out	labelOUT	data_prest	hosp	pharm	dataADM	LOS	classe_pharma	ATC	qt_pharma	DDD	COMBO
1	10006065	2006-08-29	2012-12-31	CENSORED	2006-08-29	1		2006-08-14	15					
2	10006065	2006-08-29	2012-12-31	CENSORED	2006-08-29		1			AA	C03EB01	20		1
3	10006065	2006-08-29	2012-12-31	CENSORED	2006-08-29		2			ACE	C09AA05	56	2.50	0
4	10006065	2006-08-29	2012-12-31	CENSORED	2006-09-15		3			AA	C03EB01	20		1
5	10006065	2006-08-29	2012-12-31	CENSORED	2006-09-15		4			ACE	C09AA05	56	2.50	0
6	10006065	2006-08-29	2012-12-31	CENSORED	2006-10-06		5			AA	C03EB01	20		1
7	10006065	2006-08-29	2012-12-31	CENSORED	2006-10-17		6			AA	C03EB01	20		1
8	10006065	2006-08-29	2012-12-31	CENSORED	2006-10-17		7			ACE	C09AA05	56	2.50	0
9	10006065	2006-08-29	2012-12-31	CENSORED	2006-10-27	2		2006-10-24	3					
10	10006065	2006-08-29	2012-12-31	CENSORED	2006-10-31		8			DIU	C03CA01	38	40.00	0
11	10006065	2006-08-29	2012-12-31	CENSORED	2006-10-31		9			ACE	C09AA05	224	2.50	0
12	10006065	2006-08-29	2012-12-31	CENSORED	2006-12-18		10			DIU	C03CA01	38	40.00	0
13	10006065	2006-08-29	2012-12-31	CENSORED	2006-12-18		11			ACE	C09AA05	224	2.50	0
14	10006065	2006-08-29	2012-12-31	CENSORED	2007-01-19		12			DIU	C03CA01	38	40.00	0
15	10006065	2006-08-29	2012-12-31	CENSORED	2007-01-19		13			ACE	C09AA05	224	2.50	0
16	10006065	2006-08-29	2012-12-31	CENSORED	2007-03-20		14			DIU	C03CA01	23	40.00	0
17	10006065	2006-08-29	2012-12-31	CENSORED	2007-03-20		15			ACE	C09AA05	224	2.50	0
18	10006065	2006-08-29	2012-12-31	CENSORED	2007-05-22		16			DIU	C03CA01	23	40.00	0
19	10006065	2006-08-29	2012-12-31	CENSORED	2007-05-22		17			ACE	C09AA05	224	2.50	0
20	10006065	2006-08-29	2012-12-31	CENSORED	2007-06-26		18			DIU	C03CA01	38	40.00	0
21	10006065	2006-08-29	2012-12-31	CENSORED	2007-06-26		19			ACE	C09AA05	224	2.50	0

Table 2.10: Events data of LR patient 10006065.

	COD_REG	data_rif_ev	data_studio_out	labelOUT	data_prest	hosp	pharm	dataADM	LOS	classe_pharma	ATC	qt_pharma	DDD	COMBO
1	10009476	2008-12-22	2011-08-22	DEAD	2008-12-22	1		2008-12-07	15					
2	10009476	2008-12-22	2011-08-22	DEAD	2009-01-10	2		2008-12-22	19					
3	10009476	2008-12-22	2011-08-22	DEAD	2009-01-12		1			DIU	C03CA01	56	40.00	0
4	10009476	2008-12-22	2011-08-22	DEAD	2009-01-12		2			ACE	C09AA05	56	2.50	0
5	10009476	2008-12-22	2011-08-22	DEAD	2009-02-13	3		2009-02-01	12					
6	10009476	2008-12-22	2011-08-22	DEAD	2009-02-24	4		2009-02-13	11					
7	10009476	2008-12-22	2011-08-22	DEAD	2009-02-25		3			AA	C03DA03	60		1
8	10009476	2008-12-22	2011-08-22	DEAD	2009-02-25		4			ACE	C09AA01	75	50.00	0
9	10009476	2008-12-22	2011-08-22	DEAD	2009-05-08		5			AA	C03DA03	60		1
10	10009476	2008-12-22	2011-08-22	DEAD	2009-05-08		6			ACE	C09AA01	75	50.00	0
11	10009476	2008-12-22	2011-08-22	DEAD	2009-05-16		7			DIU	C03CA01	56	40.00	0
12	10009476	2008-12-22	2011-08-22	DEAD	2009-06-11		8			DIU	C03CA01	56	40.00	0
13	10009476	2008-12-22	2011-08-22	DEAD	2009-07-23		9			DIU	C03CA01	94	40.00	0
14	10009476	2008-12-22	2011-08-22	DEAD	2009-07-23		10			ACE	C09AA01	75	50.00	0
15	10009476	2008-12-22	2011-08-22	DEAD	2009-09-02		11			DIU	C03CA01	112	40.00	0
16	10009476	2008-12-22	2011-08-22	DEAD	2009-11-28	5		2009-11-17	11					
17	10009476	2008-12-22	2011-08-22	DEAD	2009-11-30		12			BB	C07AG02	14	37.50	0
18	10009476	2008-12-22	2011-08-22	DEAD	2009-12-09		13			DIU	C03CA01	112	40.00	0
19	10009476	2008-12-22	2011-08-22	DEAD	2009-12-09		14			AA	C03DA03	60		1

Table 2.11: Events data of LR patient 10009476.

2.1.6 Curve of assumed dose

Starting from curves of accumulated days, we generate a second set of functions representing the total assumed dose at time t. Also in this case each patient might have five different curves, one for each type of drug (ACE, ARB, BB, AA and DIU) depending on which drugs he/she assumes.

As we did before, we set the observation period at 365 days and, in case of overlapping periods, we consider only distinct days. We assume again that all the prescribed types of drug are assumed by patients during the whole period of hospitalization and we do not consider the first hospitalization because the reference date of each patient is the one of the first discharge from hospital.

In this case we need two additional hypothesis for cases in which the DDD is not available (hospitalizations or pharmacological events with `COMBO=1`). We consider separately each pharmacological class (ACE, ARB, AA, BB and DIU) and:

1. to hospitalizations we assign the DDD of the previous prescription of the pharmacological class considered (if present), otherwise the median of DDDs of all the prescriptions of the pharmacological class considered with `COMBO=0`
2. to pharmacological events with `COMBO=1` we assign the median value of DDDs of all the prescriptions of the pharmacological class considered with `COMBO=0`

This simplifying hypotheses are needed to come up with curves of assumed dose for all the patients, avoiding the exclusion of `COMBO=1` cases (10% of the pharmacological events).
Whenever we meet a case of type 1 or 2 in the dataset we report it with * in Tables 2.12, 2.13 and 2.14.

It may happen that, for a type of drug, a patient presents only hospitalizations and pharmacological events with `COMBO=1` since we have no dose to assign to events (all DDDs are missing, as explained in Section 2.1.3). In this case, it is not possible to calculate the curve. Due to this fact, for each patient and for each type of drug, we add a binary variable, named `curvaMG`, which marks if the curve of assumed dose is available.

We report again a real example of a female patient, corresponding to `COD_REG` = 10009476, whose events data are reported in Table 2.11. We note that her observation period ranges from `data_rif_ev`="2008-12-22" to "2009-12-21". Then, from `classe_pharma`, we can observe that she presents four hospitalizations and assumes four different types of drug: BB (one prescription), DIU (six prescriptions), ACE (four prescriptions) and AA (three prescriptions).

About Beta Blocking agents (BB), there are five different assumption periods without overlaps. Since there is only one prescription, we assign the corresponding DDD to all the hospitalizations, as reported in Table 2.12. The resulting curve is the green one in Figure 2.3.

No.	Event	Start	End	Days	Overlaps	Dose (mg)
1	`hosp=2`	"2008-12-22"	"2009-01-09"	19		37.5*
2	`hosp=3`	"2009-02-01"	"2009-02-12"	12		37.5*
3	`hosp=4`	"2009-02-13"	"2009-02-23"	11		37.5*
4	`hosp=5`	"2009-11-17"	"2009-11-27"	11		37.5*
5	`pharm=12`	"2009-11-30"	"2009-12-13"	14		37.5

Table 2.12: Medical history of Beta Blocking agents of patient 10009476

About Diuretics (DIU), there are ten different assumption periods with various overlaps. Since all the prescriptions have the same dose, we assign it to all the hospitalizations, as reported in Table 2.13. The resulting curve is the blue one in Figure 2.3.

No.	Event	Start	End	Days	Overlaps	Dose (mg)
1	`hosp=2`	"2008-12-22"	"2009-01-09"	19		40*
2	`pharm=1`	"2009-01-12"	"2009-03-08"	56		40
3	`hosp=3`	~~"2009-02-01"~~	~~"2009-02-12"~~	~~12~~	12	40*
		Total	overlaps	0		
4	`hosp=4`	~~"2009-02-13"~~	~~"2009-02-23"~~	~~11~~	11	40*
		Total	overlaps	0		
5	`pharm=7`	"2009-05-16"	"2009-07-10"	56		40
6	`pharm=8`	~~"2009-06-11"~~	"2009-08-05"	~~56~~	30	40
		"2009-07-11"	"2009-08-05"	26		
7	`pharm=9`	~~2009-07-23"~~	"2009-10-24"	~~94~~	13	40
		"2009-08-06"	"2009-10-24"	81		
8	`pharm=11`	~~"2009-09-02"~~	~~"2009-12-22"~~	~~112-1~~	52	40
		"2009-10-25"	"2009-12-21"	59		
9	`hosp=5`			0		
10	`pharm=13`			0		

Table 2.13: Medical history of Diuretics of patient 10009476

No.	Event	Start	End	Days	Overlaps	Dose (*mg*)
1	`hosp=2`	"2008-12-22"	"2009-01-09"	19		50.0*
2	`pharm=2`	"2009-01-12"	"2009-03-08"	56		2.5
3	`hosp=3`	~~"2009-02-01"~~	~~"2009-02-12"~~	~~12~~	12	2.5*
		Total	overlaps	0		
4	`hosp=4`	~~"2009-02-13"~~	~~"2009-02-23"~~	~~11~~	11	2.5*
		Total	overlaps	0		
5	`pharm=4`	~~"2009-02-25"~~	"2009-05-10"	~~75~~	12	50.0
		"2009-03-09"	"2009-05-10"	63		
6	`pharm=6`	~~"2009-05-08"~~	"2009-07-21"	~~75~~	3	50.0
		"2009-05-11"	"2009-07-21"	72		
7	`pharm=10`	"2009-07-23"	"2009-10-05"	75		50.0
8	`hosp=5`	"2009-11-17"	"2009-11-27"	11		50.0*

Table 2.14: Medical history of ACE Inhibitors of patient 10009476

About ACE Inhibitors (ACE), there are eight different assumption periods with some overlaps, as shown in Table 2.14. Since event no.1 is an hospitalization, we assign to it the median value of all DDDs (no previous prescription). The resulting curve is the red one in Figure 2.3.

Finally, about Anti Aldosterone agents (AA), we observe that all the pharmacological events present `COMBO=1` so there does not exist any available dose. Therefore, we cannot calculate the corresponding curve and `curvaMG=0`.

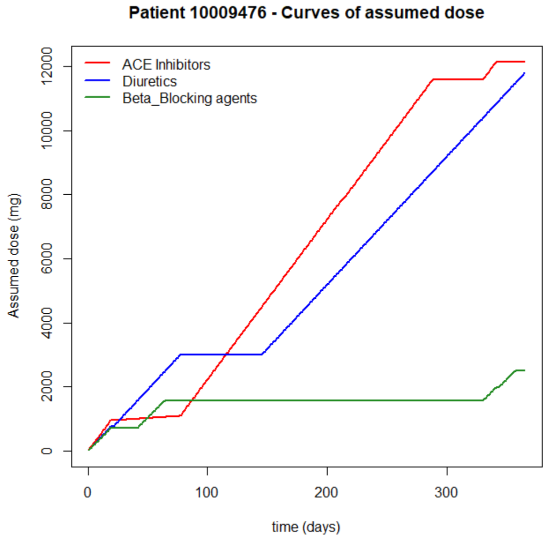


Figure 2.3: Curves of assumed doses of patient 10009476 for ACE (red), DIU (blue) and BB (green).

Note that also in this case all curves are monotone and non-decreasing, since they represent the assumed doses of drugs over time. A more detailed description of the curves of assumed dose for LR patients is given in Appendix B.2.

2.1.7 Final datasets

At the end of the procedure described in the previous Sections, for each type of drug we assemble a final dataset selecting a list of patient's features. Let's keep in mind that our aim is to set a dataset that is handleable for survival analysis with time-dependent covariates. For each type of drug and for each patient, we decide to maintain some characteristics and to modify other the ones creating four new variables:

- patient's age at the first discharge for HF
- total number of patient's hospitalizations in the reference period
- total number of patient's comorbidities at the first HF hospitalization
- total number of patient's procedures at the first HF hospitalization

We also keep the curves of cumulative coverage days and of assumed dose over time for each patient. The resulting variables are summarized in Table 2.15.

We underline that our initial dataset contains several rows for each patient, one for each event, so it is in a long format. After all data rearrangements, we end up with five final datasets, one for each pharmacological class (ACE, AA, ARB, BB and DIU), with only one row per patient. Therefore, each final dataset contains as many rows as there are patients who follow the treatment.

As example we report the female patient, corresponding to `COD_REG` = 10009476, that we have considered in Section 2.1.6 for the computation of the curves of assumed dose. For this patient, the initial dataset contains 19 events, whose principal characteristics are reported in Table 2.11. Since she follows four different treatments (ACE, AA, BB and DIU), she has been inserted in the final datasets of ACE, AA, BB and DIU. In particular:

- Table 2.16 shows her retained variables for ACE treatment
- Table 2.17 shows her retained variables for AA treatment
- Table 2.18 shows her retained variables for BB treatment
- Table 2.19 shows her retained variables for DIU treatment

Variable	Description
COD_REG	Anonymous ID code of each patient
classe_pharma	Type of drug (ACE, ARB, BB, AA, DIU) patient assumes
death	Binary flag which marks if the patient is dead by the end of the study
labelOUT	Status at the end of the study
timeOUT	Follow up time [days]
age_in	Patient's age at the first discharge for HF
gender	Patient's gender
tot_hosp	Total number of patient's hospitalizations
comorbidity	Total number of patient's comorbidities
tot_procedures	Total number of procedures the patient underwent
ADERENTE	Binary flag which marks if a patient is adherent or not
ADERENZA	Coverage days during observation period for the assuming drug
PDC	Proportion of Days Covered
PDC_CLA	Adherence class (or level)
curvaMG	Binary flag which marks if the curve of dose is available for the patient
day_t	Value of the curve of accumulated days at time t
dose_t	Value of the curve of assumed dose at time t

Table 2.15: Variables retained for each patients of the LR final datasets.

	COD_REG	classe_pharma	death	labelOUT	timeOUT	age_in	gender
1	10009476	ACE	1	DEAD	973	82	F
	tot_hosp	comorbidity	tot_procedures	ADERENTE	ADERENZA	PDC	PDC_CLA
1	5	1	1	1	296	0.8109589	4
	curvaMG	day_1		day_365	dose_1		dose_365
1	1	1	...	296	50	...	12140

Table 2.16: Row of LR patient 10009476 in the final dataset of ACE Inhibitors.

	COD_REG	classe_pharma	death	labelOUT	timeOUT	age_in	gender
1	10009476	AA	1	DEAD	973	82	F
	tot_hosp	comorbidity	tot_procedures	ADERENTE	ADERENZA	PDC	PDC_CLA
1	5	1	1	0	186	0.509589	3
	curvaMG	day_1		day_365	dose_1		dose_365
1	0	1	...	186	NA	...	NA

Table 2.17: Row of LR patient 10009476 in the final dataset of Anti Aldosterone agents.

	COD_REG	classe_pharma	death	labelOUT	timeOUT	age_in	gender
1	10009476	DIU	1	DEAD	973	82	F
	tot_hosp	comorbidity	tot_procedures	ADERENTE	ADERENZA	PDC	PDC_CLA
1	5	1	1	0	67	0.1835616	1
	curvaMG	day_1		day_365	dose_1		dose_365
1	1	1	...	67	37.5	...	2512.5

Table 2.18: Row of LR patient 10009476 in the final dataset of Beta Blocking agents.

	COD_REG	classe_pharma	death	labelOUT	timeOUT	age_in	gender
1	10009476	DIU	1	DEAD	973	82	F
	tot_hosp	comorbidity	tot_procedures	ADERENTE	ADERENZA	PDC	PDC_CLA
1	5	1	1	1	295	0.8082192	4
	curvaMG	day_1		day_365	dose_1		dose_365
1	1	1	...	295	40	...	11800

Table 2.19: Row of LR patient 10009476 in the final dataset of Diuretics.

2.2 Friuli Venezia Giulia Region dataset

In the Friuli Venezia Giulia Region (FVGR) dataset data about patients hospitalized for HF from 2009 to 2016 are collected. It is composed by 1,083,691 events related to 20,435 patients.

2.2.1 Variables

For each patient, identified by its unique anonymous ID code, we have some characteristics like the date of birth, the gender and the date of death, if it occurs before the end of the study. Moreover patients are classified as *Worsening* HF or *De Novo* on the base of the presence of at least one HF hospitalization in the five years preceding the first admission for HF. All these variables are reported in Table 2.20.

Each record in the dataset is related to an event, which can be a hospitalization or a pharmacological event. In the first case we have three types of hospitalizations: admission in hospital for HF, all-cause readmission in hospital or Integrated Home Care (IHC) service. For each hospitalization, we know the dates of admission and of discharge. In the second case, each event represents a pharmacological prescription characterized by the date of purchase, ATC and AIC codes and the total number of purchased boxes. All these variables are reported in Table 2.21.
Moreover, further information is available about patients' medical history, such as a list of comorbidities, one of particular procedures and laboratory data, as shown in Tables 2.22, 2.23 and 2.24, respectively.

Contrary to what happens in the Lombardy Region case, it is important to observe that the duration of each prescription is not available so the presence of AIC code, given by `FARMACI_MINSAN10` in Table 2.21, is fundamental for our analysis. Moreover, in LR dataset pharmacological prescriptions occur only after the reference date, whereas in this case they can also occur earlier. In particular we consider informative only those events dated six months before the first discharge for HF: prior ones are not appropriate for our analyses since too old with respect to the observation period. These further information allows us to integrate patient's data related to the observation period with his/her past medical history, adding knowledge about types of assumed drug and related dosages. For example, if a patient took BB in the six months before the reference date but he/she does not present BB prescriptions during his/her observation period, we assume that during the whole periods of hospitalizations or IHC services the patient takes also BB in the last previous recorded dose.

Variable	Description
`KEY_ANAGRAFE`	Anonymous ID code of each patient
`tipo`	Worsening - De Novo flag
`DATA_NASCITA`	Patient's date of birth
`ANA_SESSO`	Patient's gender
`deceduto`	Binary flag which marks if the patient is dead by the end of the study
`ana_data_decesso`	Date of death

Table 2.20: Patient's information in FVGR dataset.

Variable	Description	
Type of event	Pharma	Hospitalization
`stato`	NA	Type of hospitalization
`farma`	Binary flag which marks if the current event is pharma	
	TRUE	FALSE
`data_inizio_tot`	Date of purchase	Date of admission
`data_fine_tot`	NA	Date of discharge
`FARMACI_ATC`	ATC code	NA
`FARMACI_MINSAN10`	AIC code	NA
`FARMAPRESCR_PEZZI`	Number of purchased boxes	NA

Table 2.21: Event information in FVGR dataset.

Variable	Binary flag which marks the presence of
MI_CH	myiocardial infarction
CONG_HEART_FAIL_CH	congestive heart failure
PERIPH_VASC_DIS_CH	peripheric vascular disease
CEREBROVASC_DIS_CH	cerebrovascular disease
dementia	dementia
CHRONIC_PULM_DIS_CH	chronic pulmonary disease
RHEUMATIC_DIS_CH	rheumatic disease
PEPTIC_ULCER_DIS_CH	peptic ulcer disease
MILD_LIVER_CH	mild liver disease
diab_no_compl_ch	diabetes without compliance
diab_compl_c	diabetes with compliance
hem_parapl_ch	hemiplagia
renal_dis_ch	renal disease
met_solid_tum_ch	solid tumour
any_malig_ch	any malignity
mod_sev_liv_dis_ch	severe liver disease
aids_hiv_ch	AIDS/HIV
anam_cong_heart_fail_ch	congestive heart failure in anamnesis
anam_periph_vasc_dis_ch	peripheric vascular disease in anamnesis
anam_cerebrovasc_dis_ch	cerebrovascular disease in anamnesis
anam_dementia	dementia in anamnesis
anam_chronic_pulm_dis_ch	chronic pulmonary disease in anamnesis
anam_rheumatic_dis_ch	rheumatic disease in anamnesis
anam_peptic_ulcer_dis_ch	peptic ulcer disease in anamnesis
anam_mild_liver_ch	mild liver disease in anamnesis
anam_diab_no_compl_ch	diabetes without compliance in anamnesis
anam_diab_compl_ch	diabetes with compliance in anamnesis
anam_hem_parapl_ch	hemiplagia in anamnesis
anam_renal_dis_ch	renal disease in anamnesis
anam_met_solid_tum_ch	solid tumour in anamnesis
anam_any_malig_ch	any malignity in anamnesis
anam_mod_sev_liv_dis_ch	severe liver disease in anamnesis
anam_icd9_scc	ICD-9CM in anamnesis

Table 2.22: Variables for hospitalization comorbidities in FVGR dataset. All these variables are NA for pharmacological events.

Variable	Description	
Type of event	Pharma	Hospitalization
`fa`	NA	Binary flag which marks if patient has received an Atrial Fibrillation
`crtd`	NA	Binary flag which marks if patient has received a Cardiac Resynchronization Therapy Defibrillator
`crt`	NA	Binary flag which marks if patient has received a Cardiac Resynchronization Therapy
`coro`	NA	Binary flag which marks if patient has received a Coronary Angiography

Table 2.23: Variables for hospitalization procedures in FVGR dataset.

Variable	Description	
Type of event	Pharma	Hospitalization
`HBGL_MEDIAN_new`	NA	Median value of Health-Based Guidance Level
`CREA_MEDIAN_new`	NA	Median value of Creatinine
`BNP_MEDIAN_new`	NA	Median value of B-type Natriuretic Peptide
`EMO_MEDIAN_new`	NA	Median value of Hemochrome

Table 2.24: Variables for laboratory data in FVGR dataset.

2.2.2 Patients and events selection

The first step of our analysis consists in selecting the proper cohort of patients. First of all we perform data cleaning removing patients with inaccurate or incorrect records and we keep only non-pediatric patients. As in the previous case, we decide to consider as reference time the date of the first HF discharge and not of admission in order to exclude those patients who died during the first hospitalization. Therefore we select patients who survived at least one year because, as mentioned in Section 1.5, we are interested in investigating one year of adherence.

Regarding events, we keep only pharmacological events related to ACE, ARB, BB, AA and DIU and hospitalizations. To select only these specific pharmacological classes we consider the ATC codes, as we have explained in Section 1.3.1 (Table 1.3). In order to do that we introduce another categorical feature, named `classe_pharma`, which indicate the type of drug (*ACE-Inhibitors, Angiotensin Receptor Blockers, Beta-Blocking agents, Anti-Aldosterone agents* and *Diuretics*). Then, since we want a follow up period of one year, we select only those events within 365 days and we set a period of six months for pharmacological events occurred before the reference date (this means that we remove prior ones). We consider only those patients with at least one hospitalization and one pharmacological event. At the end we select only patients with at least one event after the reference date and, among events dated before the first discharge for HF, we keep only the last one for each type of drug.

We end up with a final dataset of 218,843 events corresponding to 13,619 patients. Figure 2.4 reports details about each step of the procedure.

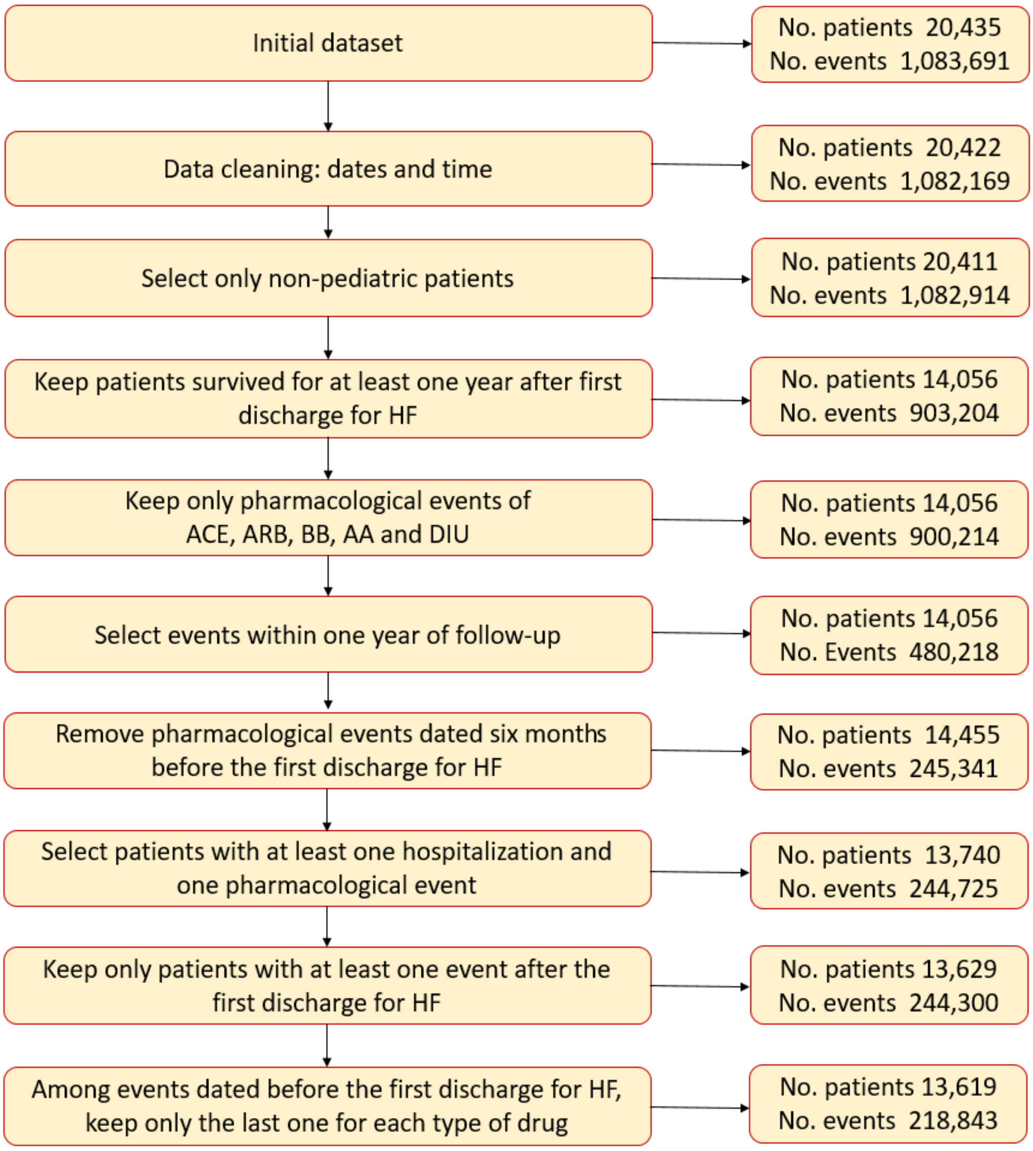

Initial dataset
No. patients 20,435
No. events 1,083,691
Data cleaning: dates and time
No. patients 20,422
No. events 1,082,169
Select only non-pediatric patients
No. patients 20,411
No. events 1,082,914
Keep patients survived for at least one year after first discharge for HF
No. patients 14,056
No. events 903,204
Keep only pharmacological events of ACE, ARB, BB, AA and DIU
No. patients 14,056
No. events 900,214
Select events within one year of follow-up
No. patients 14,056
No. Events 480,218
Remove pharmacological events dated six months before the first discharge for HF
No. patients 14,455
No. events 245,341
Select patients with at least one hospitalization and one pharmacological event
No. patients 13,740
No. events 244,725
Keep only patients with at least one event after the first discharge for HF
No. patients 13,629
No. events 244,300
Among events dated before the first discharge for HF, keep only the last one for each type of drug
No. patients 13,619
No. events 218,843

Figure 2.4: Patients and events selection procedure for FVGR dataset.

2.2.3 Adding auxiliary variables

At this point we need to introduce some variables which are necessary to develop our analysis.

First of all we rename some variables, as shown in Table 2.25, and we introduce `data_rif_ev` and `age_in`, which represent respectively the date of first discharge for HF and the patient's age at the beginning of the study. Then, since our final aim is to perform a survival analysis (see Chapter 3), we need the follow up time, named `timeOUT`, given by the difference between the date of death/censoring and the date of the first discharge for HF. For this purpose, we introduce `labelOUT`, which indicates the patient's status at the end of the study, and `data_studio_out`, which corresponds to the date of death or censoring, that is December 31st, 2016 ("2016-12-31"). All these variables are summarized in Table 2.26.

$$\texttt{labelOUT} := \begin{cases} \textit{dead} & \text{if } \texttt{death} = 1 \\ \textit{trucated} & \text{if } \texttt{death} = 0 \end{cases}$$

$$\texttt{data\_studio\_out} := \begin{cases} \texttt{ana\_data\_decesso} & \text{if } \texttt{labelOUT} = \text{dead} \\ \textit{"2016-12-31"} & \text{if } \texttt{labelOUT} = \text{truncated} \end{cases}$$

$$\texttt{timeOUT} := \texttt{data\_studio\_out} - \texttt{data\_rif\_ev}$$

For hospitalization events, added variables are reported in Table 2.27. We insert two other variables, `hosp` and `IHC`, which respectively represent the index of hospitalization and the fact that the patient was in IHC at least once or not. We also compute the Length of Stay in hospital as:

$$\texttt{LOS} := \texttt{data\_fine\_tot} - \texttt{data\_inizio\_tot}$$

For pharmacological events, all the added variables are shown in Table 2.28. We insert variables `pharm` and `classe_pharma`, for the index of pharmacological event and for the type of drug respectively. Using ATC codes and WHO's website as explained in Section 1.3.1, we add the Defined Daily Dose (`DDD`) which is not available for drugs containing molecules of several pharmacological classes (i.e. combinations). Therefore we also introduce the binary flag `COMBO` which marks if the drug is a combination or not. Moreover, using AIC code as explained in Section 1.3.2, we recover milligrams of drug in one tablet (`QTA_MG_CPR`), milligrams of the first active principle in one tablet (`QTA_MG_IFAR`) and the number of tablets in one box (N_CPR). Finally, for each drug prescription we determine the assumed dose (DOSE). We hypothesize that the patient takes DDD, if the drug is not a combination, *mg* of one tablet, if the drug is a combination and it is composed by only one active principle, or *mg* of the first active principle in one tablet, if the drug is a combination and it is composed by more active principles.

$$\texttt{DOSE} := \begin{cases} \texttt{DDD} & \text{if } \texttt{COMBO} = 0 \\ \texttt{QTA\_MG\_CPR} & \text{if } \texttt{COMBO} = 1 \text{ and only one active principle} \\ \texttt{QTA\_MG\_IFAR} & \text{if } \texttt{COMBO} = 1 \text{ and more active principles} \end{cases}$$

Initial variable	Renamed variable
ANA_SESSO	gender
deceduto	death
FARMACI_ATC	ATC
FARMACI_MINSAN10	AIC
FARMAPRESCR_PEZZI	N_PEZZI

Table 2.25: Renamed variables in FVGR dataset.

Variable	Description
data_rif_ev	Date of first discharge for HF
age_in	Patient's age at the first discharge for HF
labelOUT	Status at the end of the study
data_studio_out	Date of death/censoring
timeOUT	Follow up time [days]

Table 2.26: Patients' added variable in FVGR dataset.

Variable	Description
hosp	Index of hospitalization
IHC	Binary flag which marks if the patient was in IHC at least once
LOS	Length of stay

Table 2.27: Added ariables for hospitalization events in FVGR dataset.

Variable	Description
pharm	Index of pharmacological event
classe_pharma	Type of drug
DDD	Defined Daily Dose (NA for combinations)
COMBO	Binary flags which marks if the drug is a combination
QTA_MG_CPR	Milligrams in one tablet
QTA_MG_IFAR	Milligrams of the first active principle in one tablet (NA if only one active principle)
QTA_MG_CPR_CAR	String that indicates the milligrams of each active principle in one tablet (NA if only one active principle)
N_CPR	Number of tablets in one box
DOSE	Current dose

Table 2.28: Added variables for pharmacological events in FVGR dataset.

2.2.4 Duration of prescriptions

In FVG dataset duration of prescriptions are not available so we compute it using approach introduced in Section 1.4.

First of all we calculate the total milligrams of drug contained in one box (QTA_BOX) and the coverage days of a single box (qt_pharma_BOX).
In particular, we use Equation (1.1) with DDD information for event with COMBO = 0 and we suppose that a patient takes a table a day for event with COMBO = 1, as follows:

$$\texttt{QTA\_BOX} := \begin{cases} \texttt{QTA\_MG\_CPR} \cdot \texttt{N\_CPR} & \text{if only one active principle} \\ \texttt{QTA\_MG\_IFAR} \cdot \texttt{N\_CPR} & \text{if more active principles} \end{cases}$$

$$\texttt{qt\_pharma\_BOX} := \begin{cases} \texttt{QTA\_BOX/DDD} & \text{if } \texttt{COMBO} = 0 \\ \texttt{N\_CPR} & \text{if } \texttt{COMBO} = 1 \end{cases}$$

However we observe that, using this procedure, in 18.9% of pharmacological events this computation is unrealistic. In fact, coverage days period of a single box is too short or too long. In particular, this happens for pharmacological events with COMBO = 0 and a big difference in term of numeric values between DDD and mg of one tablet (QTA_MG_CPR or QTA_MG_IFAR). We report two examples:

- *Too short coverage days period*
  A patient has a pharmacological event with DDD = $40mg$ (COMBO = 0), QTA_MG_CPR = $2.5mg$ and N_CPR = 10.

The coverage days of a single box is:

$$\texttt{qt\_pharma\_BOX} := \frac{\texttt{QTA\_BOX}}{\texttt{DDD}} = \frac{2.5 \cdot 10}{40} = 0.625 \simeq 0 \text{ days}$$

- *Too long coverage days period*
  A patient has a pharmacological event with $\texttt{DDD} = 40mg$ ($\texttt{COMBO} = 0$), $\texttt{QTA\_MG\_CPR} = 500mg$ and $\texttt{N\_CPR} = 20$. The coverage days of a single box is:

$$\texttt{qt\_pharma\_BOX} := \frac{\texttt{QTA\_BOX}}{\texttt{DDD}} = \frac{500 \cdot 20}{40} = 250 \text{ days}$$

In order to come up with coverage days periods for all patients avoiding unrealistic cases, we decide to modify our procedure and we consider two possible alternatives:

1. **"One tablet a day" approach**
   We impose that all patients assume one tablet a day for each pharmacological event:

$$\texttt{qt\_pharma\_BOX} := \texttt{N\_CPR}$$

   Also the current doses are modified accordingly:

$$\texttt{DOSE} := \begin{cases} \texttt{QTA\_MG\_CPR} & \text{if only one active principle} \\ \texttt{QTA\_MG\_IFAR} & \text{if more than one active principle} \end{cases}$$

2. **Mixed approach**
   We consider as unrealistic cases those one with $\texttt{COMBO} = 0$ and a coverage days period less than 7 days or greater than 100 days. Therefore we divide patients into three groups:

$$\begin{aligned} C_0 =& \{\text{events with } \texttt{COMBO} = 0 \text{ and } 7 \leq \texttt{qt\_pharma\_BOX} \leq 100\} \\ \widetilde{C}_0 =& \{\text{events with } \texttt{COMBO} = 0 \text{ and } \texttt{qt\_pharma\_BOX} < 7\} \cup \\ & \{\text{events with } \texttt{COMBO} = 0 \text{ and } \texttt{qt\_pharma\_BOX} > 100\} \\ C_1 =& \{\text{events with } \texttt{COMBO} = 1\} \end{aligned}$$

   We use DDD information only for those events belong to C_0 (74.1% of the total pharmacological events), which we consider realistic computations, whereas we impose the assumption of one tablet a day for events belong to $\widetilde{C}_0$ (unrealistic ones, 18.9%) and C_1 (7%). This means that we change the computation of the coverage days only for events belong to $\widetilde{C}_0$:

$$\texttt{qt\_pharma\_BOX} := \begin{cases} \texttt{QTA\_BOX/DDD} & \text{if event} \in C_0 \\ \texttt{N\_CPR} & \text{if event} \in C_1 \text{ or } \widetilde{C}_0 \end{cases}$$

Also the current doses are modified accordingly:

$$\texttt{DOSE} := \begin{cases} \texttt{DDD} & \text{if event} \in C_0 \\ \texttt{QTA\_MG\_CPR} & \text{if event} \in C_1 \text{ or } \widetilde{C}_0 \\ & \quad \text{and only one active principle} \\ \texttt{QTA\_MG\_IFAR} & \text{if event} \in C_1 \text{ or } \widetilde{C}_0 \text{ and} \\ & \quad \text{more than one active principles} \end{cases}$$

Finally, in both approaches we compute coverage days of each prescriptions multiplying the coverage days of one box by the number of purchased box:

$$\texttt{qt\_pharma} := \texttt{qt\_pharma\_BOX} \cdot \texttt{N\_PEZZI}$$

At this point, given qt_pharma, we are finally able to compute the end date of each prescription:

$$\texttt{data\_fine\_tot} := \texttt{data\_inizio\_tot} + \texttt{qt\_pharma}$$

All the introduced variables are summarized in Table 2.29.

Variable	Description
QTA_BOX	Total milligrams of drug contained in one box
qt_pharma_BOX	Coverage days of one box
qt_pharma	Coverage days of each prescription
data_fine_tot	End date of each prescription

Table 2.29: Variables related to the duration of prescriptions for pharmacological events in FVGR dataset.

2.2.5 Adherence variables

As done in Section 2.1.4, we insert some adherence variables, summarized in Table 2.30, in order to establish if the drug was taken continuously during all the follow up period. We remind that the reference date of each patient is the one of the first discharge from hospital so, in adherence computation, we do not consider the first hospitalization.

First of all, for each patient we compute the number of distinct coverage days during an observation period of 365 days and we call it ADERENZA. Dividing this last data by 365 (number of days in the observation periods) like in Equation (1.2), we obtain the Proportion of Days Covered (PDC), which is a number between 0

and 1.
Finally, as explained in Section 1.5, we determine adherent patients and adherence levels, respectively named ADERENTE and PDC_CLA.

Variable	Description
ADERENZA	Coverage days during observation period
PDC	Proportion of Days Covered
ADERENTE	Binary flags which marks if a patient is adherent
PDC_CLA	Adherence class (or level)

Table 2.30: Adherence variables for FVGR dataset.

2.2.6 Curve of cumulative days covered by drug assumption

At this point, we determine the curves of cumulative days covered by drug assumption, as done in Section 2.1.5. Potentially there are five different curves for each patient, one for each type of drug (ACE, ARB, BB, AA and DIU) depending on which drugs he/she assumes.

As we did for LR dataset, we set an observation period of 365 days and, in case of overlapping periods, we consider only distinct days. We assume that all the prescribed types of drug are assumed by patients during the whole period of hospitalization or IHC service. Finally, we do not consider the first hospitalization because the reference date of each patient is the one of the first discharge from hospital.

In Figure 2.5 we report as example the resultant curves of cumulative days covered by drug assumption for a female patient identified by KEY_ANAGRAFE = 3192939. She assumes three different types of drug: ACE (red lines), BB (green lines) and DIU (blue lines). Left panel shows curves computed with mixed approach, whereas right panel shows curves computed with "one tablet a day" approach. We observe that using different approaches can lead to different results.

We decide to focus our analyses on curves computed with mixed approach. Therefore, in Chapter 4 (Section 4.2) we will report the results obtained from mixed approach data and we underline the differences with respect to "one tablet a day" approach, if present.

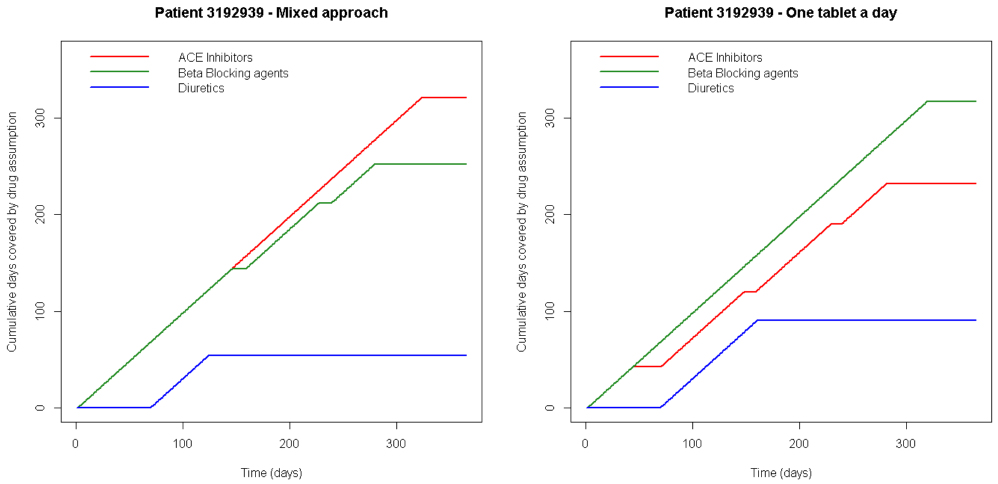


Figure 2.5: Curves of cumulative days covered by drug assumption of patient 3192939 using mixed approach (left panel) and "one tablet a day" approach (right panel). The patient assumes ACE (red lines), BB (green lines) and DIU (blue lines).

2.2.7 Curve of assumed dose

As done in Section 2.1.6, starting from curves of accumulated days, we generate curves total assumed dose at time t. Also in this case each patient might have five different curves, one for each type of drug (ACE, ARB, BB, AA and DIU) depending on which drugs he/she assumes.

As we did before, we set the observation period at 365 days and, in case of overlapping periods, we consider only distinct days. We assume again that all the prescribed types of drug are assumed by patients during the whole period of hospitalization or of IHC and we do not consider the first hospitalization because the reference date of each patient is the one of the first discharge from hospital.

On the contrary of what happened in LR case, each pharmacological event has an attributed dose, which is given by `DOSE`, as explained in Sections 2.2.3 and 2.2.4. Therefore, we need only one additional hypothesis for hospitalizations and IHC services (13.7% of total events):

1. to hospitalizations and IHC services we assign the dose of the previous pharmacological prescription (if present), otherwise the median of doses (`DOSE`)

Using this last hypothesis, a dose is attributed to each event so we are able to compute the curves of assumed dose for each patient, in contrast to LR dataset.

In Figure 2.6 we report as example the resultant of curves of assumed dose for the female patient identified by `KEY_ANAGRAFE` = 3192939. She assumes three different types of drug: ACE (red lines), BB (green lines) and DIU (blue lines). Left panel shows curves computed with mixed approach, whereas right panel shows curves computed with "one tablet a day" approach. Also in this case we observe that the resultant curves are different using the different approaches.

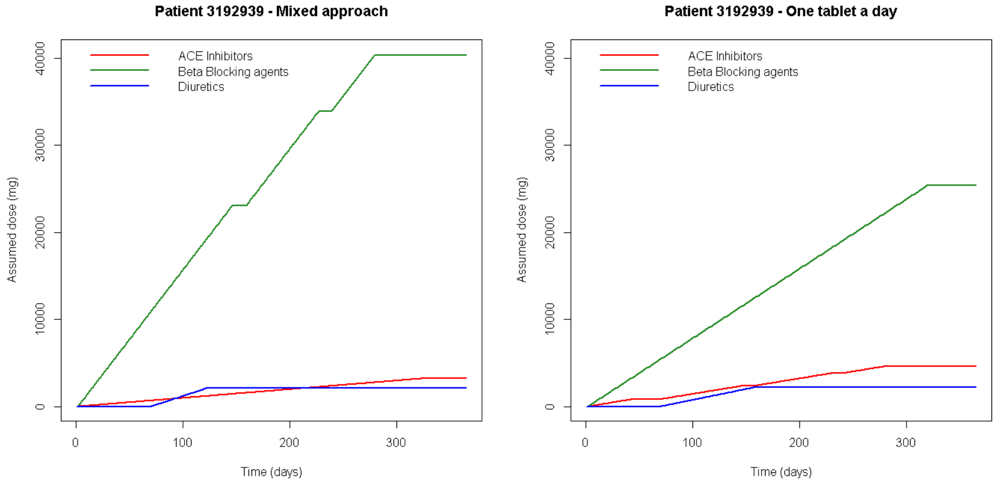


Figure 2.6: Curves of assumed dose of patient 3192939 using mixed approach (left panel) and "one tablet a day" approach (right panel). The patient assumes ACE (red lines), BB (green lines) and DIU (blue lines).

2.2.8 Final datasets

At the end of the procedure described in the previous Sections, for each pharmacological class we assemble a final dataset selecting a list of patient's features. For each type of drug and for each patient, we decide to maintain some characteristics and to modify other ones creating three new variables:

- total number of patient's hospitalizations during follow up
- total number of patient's procedures at the first HF hospitalization
- Charlson comorbidity index at the first HF hospitalization (see Appendix D for details about computation)

We also keep the curves of cumulative coverage days and of assumed dose over time for each patient. The resulting variables are summarized in Table 2.31.

Variable	Description
KEY_ANAGRAFE	Anonymous ID code of each patient
data_rif_ev	Date of the first discharge for HF
data_studioout	Date of death/censoring
labelOUT	Status at the end of the study
timeOUT	Follow up time [days]
death	Binary flag which marks if the patient is dead by the end of the study
tipo	Worsening - De Novo flag
gender	Patient's gender
age_in	Patient's age at the first discharge for HF
IHC	Binary flag which marks if the patient was in IHC
CHARLSON	Charlson comorbidity index
tot_procedures	Total number of procedures the patient underwent
tot_hosp	Total number of patient's hospitalizations
classe_pharma	Type of drug (ACE, ARB, BB, AA, DIU) patient assumes
ADERENTE	Binary flag which marks if a patient is adherent or not
ADERENZA	Coverage days during observation period for the assuming drug
PDC	Proportion of Days Covered
PDC_CLA	Adherence class (or level)
day_t	Value of the curve of accumulated days at time t
dose_t	Value of the curve of assumed dose at time t

Table 2.31: Variables retained for each patients of the FVGR final datasets.

We underline that our initial dataset is in a long format, in fact it contains several rows for each patient, one for each event. After all data rearrangements, for each approach (mixed and "one tablet a day") we end up with five final datasets, one for each pharmacological class (ACE, AA, ARB, BB and DIU) with only one row per patient. Therefore, each final dataset contains as many rows as there are patients who follow the treatment.

As example we report the female patient, corresponding to KEY_ANAGRAFE = 3192939, that we have considered in Sections 2.2.6 and 2.2.7 for the computation of the curves of cumulative days covered by drug assumption and of assumed dose, respectively. For this patient, the initial dataset contains 18 events, corresponding to the first hospitalization and 17 pharmacologiacal events (seven for ACE, eight for BB and two for DIU). Since she follows three different treatments (ACE, BB and DIU), she has been inserted in the final datasets of ACE, BB and

DIU. In particular, starting from FVGR dataset computed with mixed approach:

- Table 2.32 shows her retained variables for ACE treatment
- Table 2.33 shows her retained variables for BB treatment
- Table 2.34 shows her retained variables for DIU treatment

	KEY_ANAGRAFE	data_rif_ev	data_studio_out	labelOUT	timeOUT	death
1	3192939	2009-08-05	2010-10-28	DEAD	449	1
	tipo	gender	age_in	IHC	CHARLSON	tot_procedures
1	denovo	F	84	0	3	0
	tot_hosp	classe_pharma	ADERENTE	ADERENZA	PDC	PDC_CLA
1	1	ACE	1	321	0.8794521	4
	day_1		day_365	dose_1		dose_365
1	0	...	321	0	...	3210

Table 2.32: Row of FVGR patient 3192939 in the final dataset of ACE Inhibitors computed through mixed approach.

	KEY_ANAGRAFE	data_rif_ev	data_studio_out	labelOUT	timeOUT	death
1	3192939	2009-08-05	2010-10-28	DEAD	449	1
	tipo	gender	age_in	IHC	CHARLSON	tot_procedures
1	denovo	F	84	0	3	0
	tot_hosp	classe_pharma	ADERENTE	ADERENZA	PDC	PDC_CLA
1	1	BB	0	252	0.690411	3
	day_1		day_365	dose_1		dose_365
1	0	...	252	0	...	40320

Table 2.33: Row of FVGR patient 3192939 in the final dataset of Beta Blocking agents computed through mixed approach.

	KEY_ANAGRAFE	data_rif_ev	data_studio_out	labelOUT	timeOUT	death
1	3192939	2009-08-05	2010-10-28	DEAD	449	1
	tipo	gender	age_in	IHC	CHARLSON	tot_procedures
1	denovo	F	84	0	3	0
	tot_hosp	classe_pharma	ADERENTE	ADERENZA	PDC	PDC_CLA
1	1	DIU	0	54	0.1479452	1
	day_1		day_365	dose_1		dose_365
1	0	...	54	0	...	2160

Table 2.34: Row of FVGR patient 3192939 in the final dataset of Diuretics computed through mixed approach.

Once described the LR and FVGR datasets, in the next Chapter we will introduce statistical methodologies we will use for our analyses.

Chapter 3

Methodologies

In this Chapter we describe statistical methodologies used in order to analyse the effect of drugs on patients' survival probabilities. In Section 3.1 we present the fundamental characteristics of survival analysis and we introduce the Proportional-Hazard Cox model. In Section 3.2 we describe a methodology to deal with time-dependent variables such as the curves described in the previous Chapter: Joint Modelling of longitudinal and time-to-event data.

3.1 Survival analysis

Survival analysis, as mentioned in [2] [32] [19], is an important field of statistics dealing with time-to-event data analysis for which the outcome variable of interest, that is called *survival time*, is the amount of time elapsed since a so-called origin event until an event of interest. By time, we mean years, months, weeks or days from the beginning of follow up of an individual until an event occurs. By event, we mean death, disease incidence, relapse from remission, recovery or any designated experience of interest that may happen to an individual.
Many examples exist in several research fields such as demography, industry and socio-economic sciences. We will concentrate on applications in clinical research, with particular attention to hospitalization and pharmacological prescription processes. In our case time correspond to days, the origin event is the first discharge for HF and the event of interest is the death of the patient.

3.1.1 Censoring

Most survival analyses must consider a key analytical problem called *censoring.* In essence, censoring occurs when we have some information about individual survival time, but we do not know the survival time exactly: we usually know that a person was alive up to a certain time but do not know exactly when it failed or would fail. There are generally three reasons why censoring may occur:

1. a person does not experience the event before the study ends;
2. a person is lost to follow up during the study period;
3. a person withdraws from the study because a reason different to the event of interest.

Therefore time to event data are often only partially observed and come as a mixture of complete and incomplete observations that constitutes a big difference compared to most other statistical data. Hence censoring complicates all of the technical issues involved in analysing the data.

For each patient i, let T_i^* be the non-negative random variable denoting the failure time and C_i be a random variable that denotes the time at which a censoring mechanism kicks in. What we actually observe in time-to-event studies is the *failure time* that is either the event time T_i^* or, whichever is smaller, the censoring time C_i:

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

In addition, we usually get information on whether T_i is an actual event time or a censored observation defining an indicator random variable δ_i for non-censoring

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

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

Our definition of censoring is the most common form of censoring, also known as *right censoring*: we are interested in the longevity of our subjects but we only have censored observations because for some patients we will never know when they die but only that at censoring time they were still alive.

3.1.2 Survival function and hazard rate

Let T denote the non-negative random variable of failure time with probability density function $f(t)$ and distribution function $F(t) = Pr(T \leq t)$.

Definition 3.1.1 (Survival function). *The survival function at time t is defined as the the complement of the distribution function:*

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t) \tag{3.3}$$

Time to event data must always be non-negative and they are usually regarded as continuous. The survival function, also called the survival curve, is a non-increasing function: at the time origin $S(0) = 1$ because everybody is alive and as t gets large, $S(t)$ tends to 0 because everything/everybody eventually breaks down.

Definition 3.1.2 (Hazard function)**.** *The hazard function is the instantaneous risk of failure at time t, conditional on survival to that time:*

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < T + \Delta t | T \geq t)}{\Delta t} \tag{3.4}$$

Technically, if a waiting time T has as density function $f(t)$, the hazard function is

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T < T + \Delta t | T \geq t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{Pr(T \in [t, t + \Delta t) \cap T \geq t)/Pr(T \geq t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{Pr(T \in [t, t + \Delta t))}{\Delta t} \cdot \frac{1}{Pr(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{\int_t^{t+\Delta t} f(u)du}{\Delta t} \cdot \frac{1}{Pr(T > t)} \\
&= \frac{f(t)}{S(t)} \\
&= -\frac{d}{dt} \log S(t)
\end{aligned}$$

The third equality follows because $T \in [t, t + \Delta t)$ implies $T \geq t$, the fifth equality follows from the Fundamental Theorem of Calculus and the definition of a derivative and the last one from the fact that $-f(t)$ is the derivative of $S(t)$.

Integrating from 0 to t and we using the boundary condition $S(0) = 1$, we obtain a formula for the probability of surviving to duration t as a function of the hazard:

$$S(t) = \exp\left\{ - \int_0^t \lambda(u)du \right\} \tag{3.5}$$

We can also define the cumulative or integrated hazard function, that, in some cases, turns out to be much easier to estimate than the hazard one.

Definition 3.1.3 (Cumulative hazard function)**.** *The cumulative hazard function at time t is:*

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{3.6}$$

The goal of the studies in survival analysis is usually to estimate the hazard function and to assess how the covariates affect it.

3.1.3 Kaplan-Meier estimator

In 1958 Kaplan and Meier [17] proposed the Kaplan-Meier estimator, also known as the *product limit estimator*, which is a non-parametric statistic used to estimate the survival function from lifetime data. An important advantage of the Kaplan-Meier curve is that the method can deal with censored data.

Suppose to have data of the type (T,δ) and let $t_{(1)} < t_{(2)} < ... < t_{(m)}$ denote the unique event times, that are the distinct ordered times of death (not counting censoring times). For each t_i, we define d_i as the number of observed events at t_i and r_i as the number of patients still at risk at that moment. The Kaplan-Meier estimator or *product limit* of the survival function is:

$$\widehat{S}_{KM}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.7}$$

Several estimators are used to approximate its variance. One of the most common estimators is Greenwood's formula:

$$\widehat{Var}(\widehat{S}_{KM}(t)) = (\widehat{S}_{KM}(t))^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{3.8}$$

The Kaplan-Meier estimate is a step function with discontinuities or jumps at the observed death times. If there is no censoring, it coincides with the empirical survival function.

3.1.4 Nelson-Aalen estimator

The Nelson–Aalen estimator, introduced in [1] in 1978, is a non-parametric estimator of the cumulative hazard rate function in case of censored data or incomplete data

As in the previous case let $t_{(1)} < t_{(2)} < ... < t_{(m)}$ denote the unique event times, that are the distinct ordered times of death (not counting censoring times). For each t_i, we define d_i as the number of observed events at t_i and r_i as the number of patients still at risk at that moment. The Nelson-Aalen estimator of the cumulative hazard rate is:

$$\widehat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i} \tag{3.9}$$

It can be intuitively interpreted as the ratio of the number of deaths to the number exposed.

Breslow in 1972 suggested then the following estimator for the survival function:

$$\widehat{S}_B(t) = \exp\{-\widehat{\Lambda}(t)\} = \prod_{i:t_i \leq t} \exp\left\{-\frac{d_i}{n_i}\right\} \tag{3.10}$$

The variance of $\widehat{\Lambda}(t)$ can be approximated by $Var(-\log(\widehat{S}_B(t)))$ which can be obtained from Greenwood's formula.

The Breslow estimator and the Kaplan-Meier estimator are asymptotically equivalent and they usually are quite close to each other, particularly when the number of deaths is small relative to the number exposed. However, in general the Breslow estimator has uniformly lower variance than the Kaplan-Meier, though it is biased, especially when $\widehat{S}(t)$ is close to zero.

3.1.5 The Proportional-Hazard Cox model

A popular model used in survival analysis that can be used to assess the importance of various covariates in the survival times of individuals is the Cox model, which describes the relationship of covariates to a survival or other censored outcome.

Let $\mathbb{X}$ be the covariate matrix where X_{ij} denotes the j-th covariate of the i-th person, with $i = 1, ..., n$ and $j = 1, ..., p$. Suppose also that our covariates are fixed over time. For an individual i with covariate vector $\boldsymbol{X}_i$ (that corresponds to the i-th row of the covariate matrix) the Cox model assumes a hazard function for the survival time of the form

$$\lambda_i(t) = \lambda_0(t) \exp\{\boldsymbol{\beta}^T \boldsymbol{X}_i\} \tag{3.11}$$

where $\lambda_0(t)$ is an unspecified non-negative function of time called *baseline hazard* and $\boldsymbol{\beta}$ is the column vector of coefficients that we want to estimate.
For each patient i with covariate vector $\boldsymbol{X}_i$, the corresponding survival function are is:

$$S_i(t|\boldsymbol{X}_i) = [S_0(t)]^{\exp(\boldsymbol{X}_i^T \boldsymbol{\beta})} \tag{3.12}$$

where $S_0(t)$ is the survival function of the baseline population, which is

$$S_0(t) = \exp\left\{-\int_0^t \lambda_0(u) du\right\} \tag{3.13}$$

This kind of model for censored survival data then specifies that covariates have a proportional effect on the hazard function of the life-time distribution of an individual. Indeed, the hazard ratio HR for two subjects with fixed covariate vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_k$

$$HR = \frac{\lambda_i(t)}{\lambda_k(t)} = \frac{\lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X}_i)}{\lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X}_k)} = \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{X}_i)}{\exp(\boldsymbol{\beta}^T\boldsymbol{X}_k)} = \exp\{\boldsymbol{\beta}^T(\boldsymbol{X}_i - \boldsymbol{X}_k)\}$$

is constant over time. For this reason this model is also known as *proportional hazard model.*

The estimation of Cox's regression coefficients is not straightforward because of the semiparametric nature of the model. Since it is impossible to take advantage of the ordinary likelihood methods, it is necessary to use a *partial likelihood.* The term "partial" is used because the likelihood formula considers probabilities only for those subjects who died, and does not explicitly consider probabilities for those subjects who are censored.

Let T_j be the observed event time and suppose that $T_1 < T_2 < ....$ The partial likelihood for $\boldsymbol{\beta}$ is:

$$L(\boldsymbol{\beta}) = \prod_{T_j} \pi(i_j|T_j) = \prod_{T_j} \frac{Y_{i_j}(T_j)\exp(\boldsymbol{\beta}^T\boldsymbol{X}_{i_j}(T_j))}{\sum_{l=1}^{n} Y_l(T_j)\exp(\boldsymbol{\beta}^T\boldsymbol{X}_l(T_j))} \tag{3.14}$$

where

- i_j is the index of the individual who experiences an event at T_j
- $Y_i(t)$ is an indicator function which assumes the value 1 if i-th patients is at risk for the event of interest just before time t and the value 0 otherwise

Introducing the notation $R_j = \{l \in \{1 : n\}|Y_l(T_j) = 1\}$ for the risk set at T_j, we can simplify the expression (3.14) with the following:

$$L(\boldsymbol{\beta}) = \prod_{T_j} \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{X}_{i_j}(T_j))}{\sum_{l\in R_j} \exp(\boldsymbol{\beta}^T\boldsymbol{X}_l(T_j))} \tag{3.15}$$

The maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that maximizes the expression (3.15).

3.1.6 Time-dependent covariates

Covariates which change their values over the course of the study, such as the curves introduced in Chapter 2, are named time-dependent. Kalbfieish and Prentice in [16] define different types of time-dependent covariates, in particular: *external* (exogenous) covariates or *internal* (endogenous) covariates. The external covariates are unaffected by the process and their value over time is established from the beginning, while the internal covariates are related to the behaviour of the individuals over time and their time paths are jointly determined with the responses of interest.

Time-dependent covariates are often of interest when dealing with clinical databases. Examples of internal time-dependent covariates are number of comorbidities and values of biochemical markers measured on the individuals during follow up. Examples of external time-dependent covariates are age and number of procedures of individuals during follow up. Pharmacological treatments are a special case of time-dependent covariates because they can be considered both as endogenous and exogenous. Indeed, they might be perceived as external if their values are prescribed at the beginning of the study, or as internal if the treatment is modified according to the disease progression.

In the following Sections we describe Joint Models, which have been proposed for considering properly time-dependent covariates. In these cases our time-dependent variables consist of the pharmacological treatment curves described in Chapter 2. Usually pharmacological treatments are involved in the analysis as binary and fixed time covariate. In this work we are interested in representing pharmacological information as a time-varying covariate, which is a more realistic representation. For this reason, modelling drug assumption with a time-dependent variable is a new and original approach to represent patients' adherence over time.

3.2 Joint Modelling

In 2010 Rizopoulos proposed a joint model approach for dealing with internal time-dependent covariates [26] and wrote the associated R package [27]. We use this approach in order to investigate how patients' time-to-event outcome are influenced by longitudinal data, which in our case are represented by the pharmacological treatment curves.

3.2.1 Submodels specification

Let pairs $\{(T_i, \delta_i), i = 1, ..., n\}$ be the observed data for time-to-event outcome, as defined in Section 3.1.1. Let $y_i(t)$ denote the value for the longitudinal outcome at time point t for the i-th subject. We do not actually observe $y_i(t)$ at all time points, but only at the very specific occasions t_{ij} at which measurements were taken. Thus, the observed longitudinal data consist of the measurements $y_{ij} = \{y_i(t_{ij}), j = 1, ..., n_i\}$. We also denote the true and unobserved value of the longitudinal outcome at time t as $m_i(t)$, that differs from $y_i(t)$ because the latter is contaminated with with measurement error value of the longitudinal outcome at time t.

To quantify the effect of $m_i(t)$ on the risk for an event, Rizopoulos introduces the following *relative risk model*:

$$\begin{aligned} \lambda_i(t|\mathcal{M}_i(t), \boldsymbol{X}_i) &= \lim_{dt \to 0} \frac{Pr\{t \leq T_i^* < t + dt | T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{X}_i\}}{dt} \\ &= \lambda_0(t) \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta} + \alpha m_i(t)\} \end{aligned} \tag{3.16}$$

where

- $\mathcal{M}_i(t) = \{m_i(u), 0 \leq u < t\}$ denotes the history of the true unobserved longitudinal process up to time point t
- $\lambda_0(\cdot)$ denotes the baseline risk function
- $\boldsymbol{X}_i$ is a vector of baseline covariates
- $\boldsymbol{\beta}$ is the vector of regression coefficients
- α is a parameter that quantifies the effect of the underlying longitudinal outcome on the risk for an event of interest to happen

The baseline risk function can be left unspecified or can be approximated using step functions or spline-based approaches.

In the above definition of the survival model Rizopoulos used the true unobserved value of the underlying longitudinal covariate $m_i(t)$. In order to quantify the effect of this covariate to the risk for an event, it is necessary to estimate $m_i(t)$ and successfully reconstruct the complete longitudinal history $\mathcal{M}_i(t)$. To achieve this they use the available measurements $y_{ij} = \{y_i(t_{ij}), j = 1, ..., n_i\}$ of each subject and a set of modelling assumptions. In particular, they focus on normal data and they postulate a linear mixed effects model to describe the subject-specific longitudinal evolutions:

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= \widetilde{X}_i^T(t)\gamma + Z_i^T(t) b_i + \varepsilon_i(t) \qquad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{3.17}$$

where

- γ is the vector of the unknown fixed effects parameters
- b_i is the vector of random effects
- $\widetilde{X}_i(t)$ denote row vectors of the design matrices for the fixed effects
- $Z_i(t)$ denote row vectors of the design matrices for the random effects
- $\varepsilon_i(t)$ is the measurement error term with variance σ^2

Finally, the random effects b_i are assumed independent of $\varepsilon_i(t)$ and normally distributed with $b_i \sim \mathcal{N}(0, D)$.

3.2.2 Maximum likelihood estimation

Rizopoulos based the maximum likelihood estimation for joint models on the maximization of the log-likelihood corresponding to the joint distribution of the time-to-event and longitudinal outcomes $\{T_i, \delta_i, y_i\}$. To define this joint distribution he assumes that the vector of time-independent random effects b_i underlies both the longitudinal and survival processes. This means that these random effects account for both the association between the longitudinal and event outcomes, and the correlation between the repeated measurements in the longitudinal process (conditional independence):

$$p(T_i, \delta_i, y_i | b_i; \boldsymbol{\theta}) = p(T_i, \delta_i | b_i; \boldsymbol{\theta}) p(y_i | b_i; \boldsymbol{\theta}) \tag{3.18}$$

$$p(y_i | b_i; \boldsymbol{\theta}) = \prod_j p\{y_i(t_{ij}) | b_i; \boldsymbol{\theta}\} \tag{3.19}$$

where

- $\boldsymbol{\theta} = (\theta_t^T, \theta_y^T, \theta_b^T)^T$ is the parameter vector, with:
  - θ_t denoting the parameters for the event time outcome
  - θ_y denoting the parameters for the longitudinal outcomes
  - θ_b denoting the unique parameters of the random-effects covariance matrix
- y_i is the $n_i \times 1$ vector of longitudinal responses of the i-th subject
- $p(\cdot)$ denotes an appropriate probability density function

Under the modelling assumptions presented in the previous Section and the above conditional independence assumptions, the joint log-likelihood contribution for the i-th subject can be formulated as

$$\begin{aligned}\log p(T_i, \delta_i, y_i; \boldsymbol{\theta}) = \log \int & p(T_i, \delta_i | b_i; \theta_t, \gamma) \\ & \times \left[\prod_j p\{y_i(t_{ij}) | b_i; \theta_y\} \right] p(b_i; \theta_b) db_i\end{aligned} \tag{3.20}$$

where

- $p(T_i, \delta_i | b_i; \theta_t, \gamma)$ is the likelihood of the survival part
- $p\{y_i(t_{ij}) | b_i; \theta_y\}$ is the univariate normal density for the longitudinal responses
- $p(b_i; \theta_b)$ is the multivariate normal density of the random effects

Furthermore, the likelihood of the survival part is written as

$$p(T_i, \delta_i | b_i; \theta_t, \gamma) = \{\lambda_i(T_i | \mathcal{M}_i(T_i); \theta_t, \beta)\}^{\delta_i} \mathcal{S}_i(T_i | \mathcal{M}_i(T_i); \theta_t, \gamma) \tag{3.21}$$

with $\lambda_i(\cdot)$ given by (3.16) and

$$\begin{aligned}\mathcal{S}_i(t | \mathcal{M}_i(t), \boldsymbol{X}_i; \theta_t, \gamma) &= Pr(T_i^* > t | \mathcal{M}_i(t), \boldsymbol{X}_i; \theta_t, \gamma) \\ &= \exp\left\{ - \int_0^t \lambda_i(s | \mathcal{M}_i(s); \theta_t, \gamma) ds \right\}\end{aligned} \tag{3.22}$$

Maximization of the log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_i \log p(T_i, \delta_i, y_i; \boldsymbol{\theta}) \tag{3.23}$$

with respect to $\boldsymbol{\theta}$ is a computationally challenging task, due to the fact that the integral with respect to the random effects in (3.19), and the integral in the definition of the survival function in (3.22) do not have an analytical solution. In order to approximate them numerical integration techniques are needed.

For our analysis, among the options available in `JM` package [27], we assume a relative risk model (3.16) with a piecewise-constant baseline risk function of the form:

$$\lambda_0(t) = \sum_{q=1}^{Q} \xi_q I(v_{q-1} < t \leq v_q) \tag{3.24}$$

where $0 = v_0 < v_1 < ... < v_Q$ denotes a split of the time scale, with v_Q being larger than the largest observed time, and ξ_q denotes the value of the hazard in the interval $(v_{q-1}, v_q]$. Moreover, among all the techniques cited in [26] [27] [28], we use the Gauss-Hermite integration rule to approximate integral (3.20), being this approach the most suitable to our case.

3.2.3 Expected Survival

Based on a joint model fitted on a sample of size n, it is possible to predict survival probabilities for a new subject i that has provided a set of longitudinal measurements $\mathcal{Y}_i(t) = y_i(s); 0 \leq s \leq t$, where $y_i(t)$ represents an endogenous time-dependent covariate. In the calculation of subject-specific survival probabilities, $y_i(t)$ is directly related to the failure mechanism and providing longitudinal measurements up to time point t implies survival up to time t. Hence, it is more relevant to focus on the conditional probability of surviving for a time $u > t$, given survival up to t, that is:

$$\pi_i(u|t) = Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n; \boldsymbol{\theta}) \tag{3.25}$$

where $\mathcal{D}_n = \{T_i, \delta_i, y_i; i = 1, ..., n\}$ denotes the sample on which the joint model was fitted on and on which we wish to base our predictions. Using a Bayesian formulation and assumption (3.18), Equation (3.25) can be rewritten as:

$$\begin{aligned} \pi_i(u|t) &= Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n; \boldsymbol{\theta}) \\ &= \int Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}_n) d\boldsymbol{\theta} \end{aligned} \tag{3.26}$$

In the same way the first part of the integral can be rewritten as:

$$\begin{aligned} &Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) = \\ &= \int Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t), b_i; \boldsymbol{\theta}) p(b_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) db_i \\ &= \int Pr(T_i^* \geq u | T_i^* > t, b_i; \boldsymbol{\theta}) p(b_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) db_i \\ &= \int \frac{S_i\{u|\mathcal{M}_i(u, b_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}}{S_i\{t|\mathcal{M}_i(t, b_i, \boldsymbol{\theta})} p(b_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}; \boldsymbol{\theta}\}) db_i \end{aligned} \tag{3.27}$$

where $S_i(\cdot)$ is given by (3.22). Furthermore they note that the longitudinal history $\mathcal{M}_i(\cdot)$, as approximated by the linear mixed effects model, is a function of both the random effects and the parameters. For the second part Rizopolous assumes that $\{\boldsymbol{\theta}|\mathcal{D}_n\}$ can be well approximated by a multivariate normal distribution with mean $\widehat{\boldsymbol{\theta}}$, the maximum likelihood estimates, and covariance matrix $\widehat{\mathcal{H}} = \widehat{var}(\widehat{\boldsymbol{\theta}})$.

Combining (3.26) with (3.27) and $\widehat{\boldsymbol{\theta}} \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}, \widehat{\mathcal{H}})$, a Monte Carlo estimate of $\pi_i(u|t)$ can be obtained using the following simulation scheme:

1. Draw $\boldsymbol{\theta}^{(\ell)} \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}, \widehat{\mathcal{H}})$

2. Draw $b_i^{(\ell)} \sim \{b_i | T_i^*, \mathcal{Y}_i(t), \boldsymbol{\theta}^{(\ell)}\}$

3. Compute
$$\pi_i^{(\ell)}(u|t) = \frac{S_i\{u|\mathcal{M}_i(u, b_i^{(\ell)}, \boldsymbol{\theta}^{(\ell)}); \boldsymbol{\theta}^{(\ell)}\}}{S_i\{t|\mathcal{M}_i(t, b_i^{(\ell)}, \boldsymbol{\theta}^{(\ell)}); \boldsymbol{\theta}^{(\ell)}\}}$$

4. Repeat Steps 1–3 for each subject, $\ell = 1, ..., L$ times, where L denotes the number of Monte Carlo samples.

Steps 1 and 3 are straightforward to perform. On the contrary, the posterior distribution of the random effects given the observed data in Step 2 is of non-standard form, and thus a more sophisticated approach is required to sample from it. For this purpose Rizopolous makes use of a Metropolis–Hastings algorithm with independent proposals from a multivariate t-distribution centred at the empirical Bayes estimates $\widehat{b}_i = \text{argmax}_b\{\log p(T_i^*, \mathcal{Y}_i(t), b; \widehat{\boldsymbol{\theta}})\}$, with scale matrix $\widehat{var}(\widehat{b}_i) = \{\partial^2 \log p(T_i^*, \mathcal{Y}_i(t), b; \widehat{\boldsymbol{\theta}})/\partial b^T \partial b|_{b=\widehat{b}_i}\}^{-1}$ and four degrees of freedom.

The realizations $\{\pi_i^{(\ell)}(u|t), \ell = 1, ..., L\}$ can be used to derive estimates of $\pi_i(u|t)$, such as
$$\widehat{\pi}_i(u|t) = \text{median}\{\pi_i^{(\ell)}(u|t), \ell = 1, ..., L\} \tag{3.28}$$

or

$$\widehat{\pi}_i(u|t) = \frac{1}{L}\sum_{\ell=1}^{L} \pi_i^{(\ell)}(u|t) \tag{3.29}$$

and confidence intervals using the Monte Carlo sample percentiles.

3.2.4 Types of residuals

To assess the fit of the model we need to check the residual plot. Rizopolous ([26], Appendix B) introduces different types of residuals, bot for longitudinal and event processes. We remind to the paper for a deeper technical discussion and we retain here only the parts which are useful to the following analyses.

For the longitudinal part of the joint model two frequently used types of residuals are the standardized marginal and standardized subject-specific residuals, which are defined as
$$r_i^{(ym)} = \widehat{V}_i^{-1/2}(y_i - \widetilde{X}_i\widehat{\gamma}) \tag{3.30}$$
$$r_i^{(ys)}(t_{ij}) = \{y_i(t_{ij}) - \widetilde{X}_i^T(t_{ij})\widehat{\gamma} - Z_i^T(t_{ij})\widehat{b}_i\}/\widehat{\sigma} \tag{3.31}$$
where

- $\widehat{\gamma}$, $\widehat{\sigma}$ and $\widehat{D}$ denote the maximum likelihood estimates under model (3.17)

- $\widehat{b}_i$ are the empirical Bayes estimates for the random effects
- $\widehat{V}_i = Z_i \widehat{D} Z_i^T + \widehat{\sigma}^2 I$, with I denoting the identity matrix of appropriate dimensions

The marginal residuals $r_i^{(ym)}$ predict the marginal errors $y_i - \widetilde{X}_i \gamma = Z_i b_i + \varepsilon_{yi}$ and can be used to investigate misspecification of the mean structure $\widetilde{X}_i \gamma$ as well as to validate the assumptions for the within-subjects covariance structure V_i. The subject-specific residuals $r_i^{(ys)}(t_{ij})$ predict the conditional errors $\varepsilon_i(t)$ and can be used for checking the homoscedasticity and normality assumptions of the linear mixed models.

For the survival part of the joint model, a standard type of residuals is the Cox-Snell residuals. These are calculated as the value of cumulative risk function evaluated at the observed event times T_i:

$$r_i^{(tcs)} = \int_0^{T_i} \lambda_i(s|\widehat{\mathcal{M}}_i(s); \widehat{\boldsymbol{\theta}}) ds \tag{3.32}$$

If the assumed model fits the data well, we expect $r_i^{(tcs)}$ to have a unit exponential distribution; however, when T_i is censored, $r_i^{(tcs)}$ will be censored as well. To take censoring into account in checking the fit of the model, we can compare graphically the Kaplan-Meier estimate of the survival function of $r_i^{(tcs)}$ with the survival function of the unit exponential distribution.

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

Chapter 4

Applications and Results

In this Chapter we describe the results of the statistical analyses performed applying models and methods described in Chapter 3 to the datasets described in Chapter 2. In particular, in Section 4.1 we illustrate the main steps of the analysis performed on the data from Lombardy Region, whereas in Section 4.2 we report the same details about data arising from Friuli Venezia Giulia Region. We remind that our final aim is to compare the use of a dichotomized binary variable for adherence in a Cox's model or of a time-dependent variable for coverage using a JM approach.

Analyses are carried out using the R software [24]. For all models different levels of significance follow the same notation for p-value: '***' p-value $<$ 0.001, '**' 0.001 $<$ p-value $<$ 0.01, '*' 0.01 $<$ p-value $<$ 0.05, '.' 0.05 $<$ p-value $<$ 0.1, ' ' 0.1 $<$ p-value $<$ 1.

4.1 Lombardy Region (LR)

In this Section we start with a descriptive analysis of the LR dataset. We proceed with a Functional K-means technique in order to cluster the time-varying curves introduced in Sections 2.1.5 and 2.1.6. Then we perform a Cox Proportional Hazards regression analysis using only survival data including the binary variable for adherence. Finally, we apply the Joint Models method described in Section 3.2 for the analysis of both longitudinal and survival data.

4.1.1 Descriptive analysis

In this part we would like to introduce a statistical description of the LR database, remembering that 4,406 patients are present with a total of 94,151 events, as described in Section 2.1.2.

First of all, we can notice that in LR dataset about 2,401 males (55.5%) and 2,005 females (44.5%) are collected and that the number of dead patients is 1,149 (26.1%). Table 4.1 reports also the number (%) of patients lost to follow up or censored. Figure 4.1 show the same information through a barplot. The follow up time ranges from one to seven years, with a mean of about four years (1,460 days), as summarized in Table 4.2.

Pts	All	*Dead* (%)	*Lost* (%)	*Censored* (%)
All	4,406	1,149 (26.1%)	9 (0.2%)	3,248 (73.7%)
Female	2,005	549 (27.4%)	4 (0.2%)	1,452 (72.4%)
Male	2,401	600 (25.0%)	5 (0.2%)	1,796 (74.8%)

Table 4.1: Total number of dead, lost and truncated patients of LR dataset.

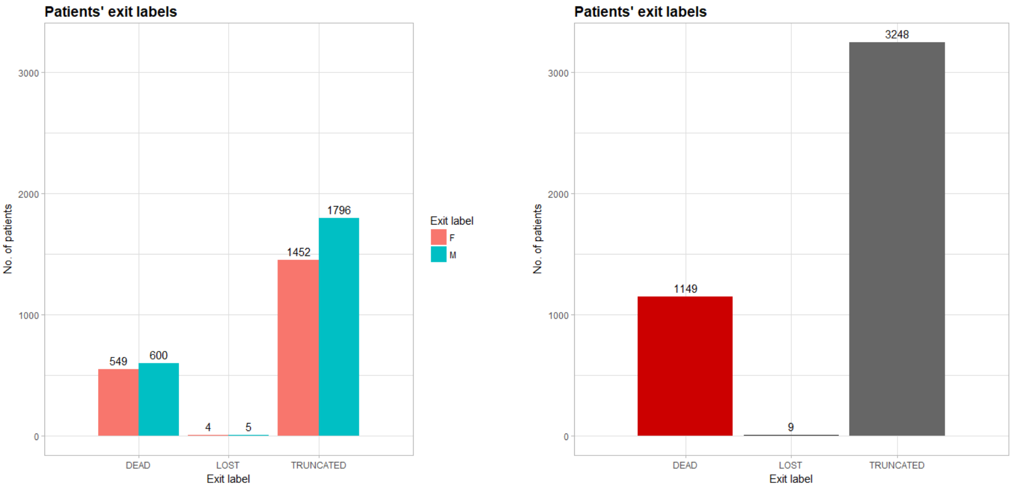


Figure 4.1: Barplots of LR exit status: patients divided by gender (left panel) and all patients (right panel).

Follow up time [days]						
Pts	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All	366	995.2	1,505	1,506	2,041	2,555
Female	367	974	1,453	1,485	2,011	2,555
Male	366	1,010	1,537	1,523	2,052	2,554

Table 4.2: Summary of the LR patients' follow up time (expressed in days).

Age at the first hospitalization [years]						
Pts	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All	18	67	75	72.97	81	98
Female	24	70	78	75.74	83	98
Male	18	65	72	70.66	79	97

Table 4.3: Summary of the LR patients' age at the first hospitalization.

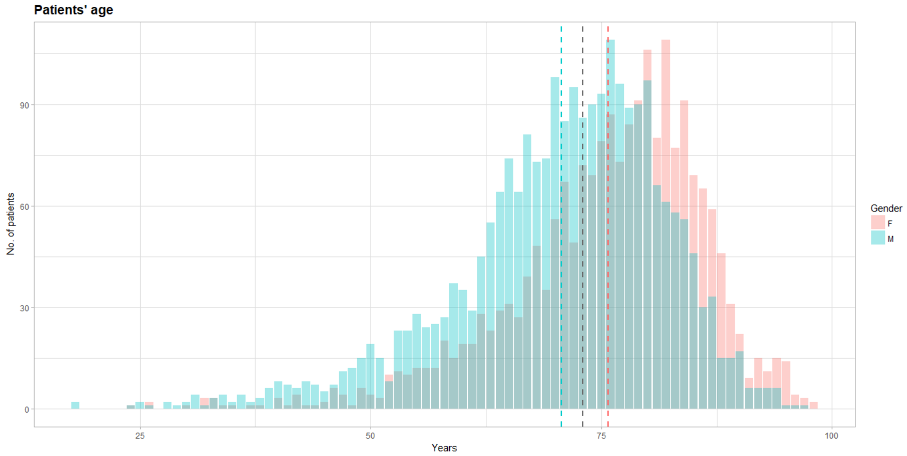


Figure 4.2: Snapshot of the LR patients' age distribution at their first hospitalization. Gray, blue and pink dotted lines represent the mean age of all patients, male patients and female patients, respectively.

As it can be seen in Table 4.3 and Figure 4.2, the age of patients ranges from 18 to 98 years, being females older than males (p-value of one-side Wilcoxon test is $< 2{\cdot}10^{-16}$). In particular 2,357 (58.8%) patients are between 67 and 81 years old.

The average number of hospitalizations per patient is equal to 2.29 and 1,771 (40.2%) patients have only the first hospitalization, which is not considered in the computation of coverage days, so the adherence information related to those patients came from pharmacological events only, as described in Sections 2.1.4 and 2.2.5. The average number of comorbidities at the first hospitalization per patient is equal 2.11. In particular, 118 (2.7%) patients have no comorbidity, 3,843 (87.2%) have one, two or three comorbidities and 445 (10.1%) patients have more than three comorbidities. Moreover, the average number of procedures at the first hospitalization per patient is 0.11. In particular, only 462 (10.5%) patients underwent to at least one procedure, whereas 3,944 (89.5%) did not undergo to any procedure. All these data are reported in Table 4.4 and Figures 4.3 and 4.4.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Hospitalizations	1	1	2	2.29	3	14
Comorbidities	0	1	2	2.11	3	8
Procedures	0	0	0	0.11	0	3

Table 4.4: Summary of the LR patients' numbers of hospitalizations, comorbidities and procedures.

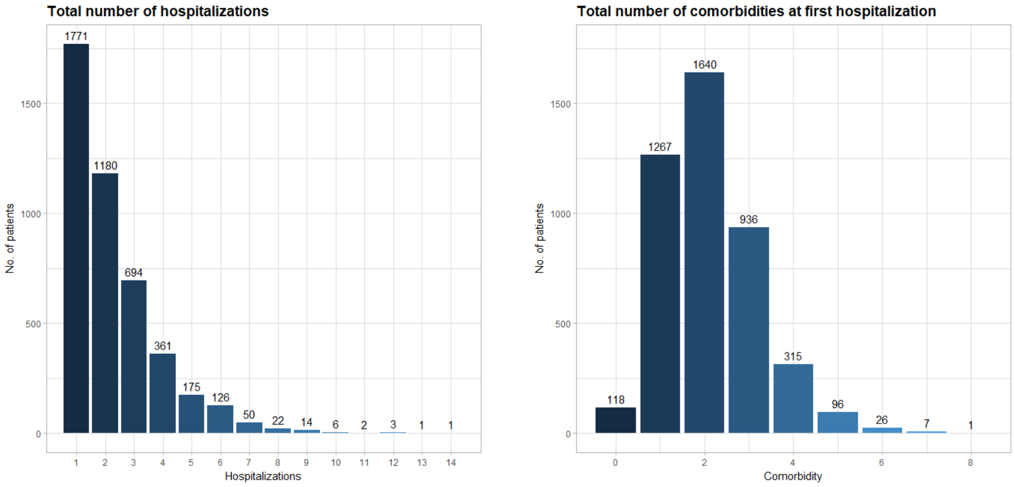


Figure 4.3: Barplots of the LR patients' total number of hospitalizations (left panel) and value of comorbidities at the first hospitalization (right panel).

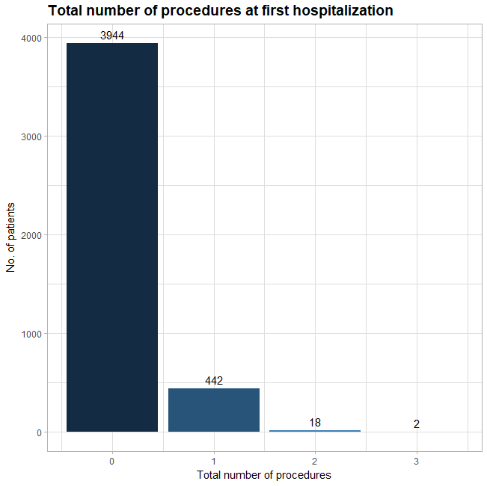


Figure 4.4: Barplot of the LR patients' total number of procedures at the first hospitalization.

In Figure 4.5 (left panel) we observe that 2,916 (66.2%) patients assumes ACE Inhibitors (ACE), 2,006 (45.5%) Anti-Aldosterone agents (AA), 1,473 (33.4%) Angiotensin Receptor Blockers (ARB), 2,890 (65.6%) Beta-Blocking agents (BB) and 3,399 (77.1%) Diuretics (DIU). Moreover, we underline that a patient can follow different pharmacological treatments, i.e., monotherapy (only one drug assumed), bytherapy (a combination of two drugs is assumed), tritherapy (a combination of three drugs is assumed) and, in general, "n" therapy (n types of drug are assumed). In particular, 446 (10.2%) patients follow a monotherapy, 1,166 (24.5%) a bitherapy, 1,474 (33.5%) a tritherapy, 1,116 (25.3%) assume four different types of drugs and 204 (4.6%) take all the pharmacological classes over the observation period (1 year), as reported in Figure 4.5 (right panel).

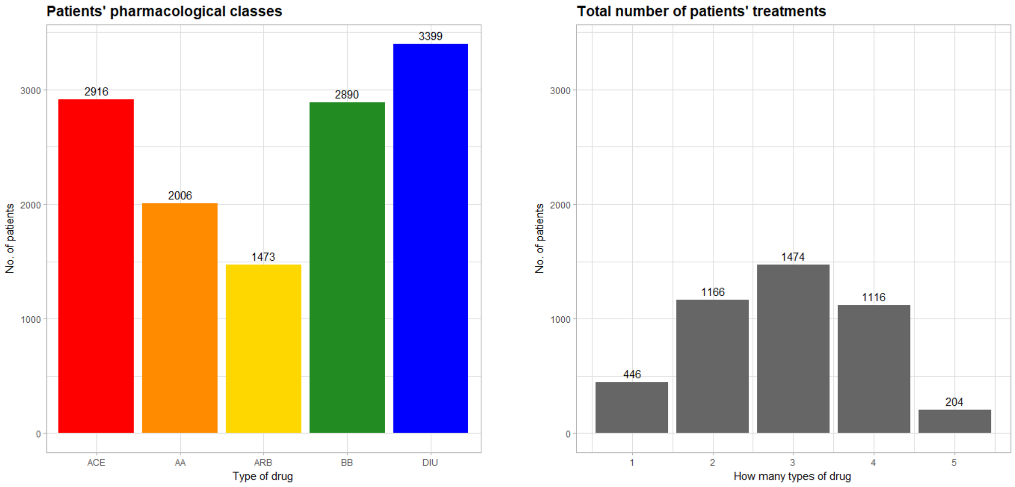


Figure 4.5: Barplots of LR patients' pharmacological classes (left panel) and total number of therapies (right panel).

For each type of drug, as mentioned in Section 2.1.7, we assemble a final dataset selecting a list of patients' peculiar features (see Table 2.15). A summary of these five final datasets is reported in Table 4.5 and in Appendix A.

At this point we proceed with the analyses on the five pharmacological datasets. For brevity we report here only the results related to ACE Inhibitors. In particular, from Table 4.5, we observe that 152 (5.2%) ACE patients present `curvaMG` $= 0$. Therefore in ACE final datasets 2,916 curves of cumulative days covered by drug assumption and 2,764 (94.5%) curves of assumed dose are collected.

Variable	Value	ACE-Inhibitors	AR Blockers	BB agents	AA agents	Diuretics
No. patients		2,916	1,473	2,890	2,006	3,399
`gender`	*Male* (%)	1,681 (57.6%)	775 (52.6%)	1,675 (58%)	1,040 (51.8%)	1,821 (53.6%)
	Female (%)	1,235 (42.4%)	698 (47.4%)	1,215 (42%)	966 (48.2%)	1,578 (46.4%)
`age_in`	mean (sd)	72.17 (11.44)	73.01 (10.20)	71.04 (11.35)	73.61 (10.60)	73.82 (10.60)
`labelOUT`	*Died* (%)	718 (24.6%)	356 (24.2%)	633 (21.9%)	545 (27.2%)	955 (28.1%)
	Truncated (%)	2,189 (75.1%)	1,115 (75.7%)	1,459 (72.7%)	1,119 (72.8%)	2,438 (71.7%)
	Lost (%)	9 (0.3%)	2 (0.1%)	6 (0.2%)	2 (0.1%)	6 (0.2%)
`death`	0 (%)	2,198 (75.4%)	1,117 (75.8%)	2,257 (78.1%)	1,461 (72.8%)	2,444 (71.9%)
	1 (%)	718 (24.6%)	356 (24.2%)	633 (21.9%)	545 (27.2%)	955 (28.1%)
`timeOUT`	mean (sd)	1,543.45 (613.46)	1,521.05 (609.92)	1,516.60 (615.30)	1,473.91 (623.35)	1,477.81 (613.68)
`tot_hosp`	mean (sd)	2.35 (1.61)	2.32 (1.63)	2.40 (1.64)	2.45 (1.67)	2.40 (1.65)
`comorbidity`	mean (sd)	2.04 (1.07)	2.20 (1.10)	2.01 (1.09)	2.06 (1.05)	2.16 (1.13)
`tot_procedures`	mean (sd)	0.13 (0.36)	0.10 (0.30)	0.14 (0.37)	0.13 (0.35)	0.11 (0.34)
`PDC`	mean (sd)	0.72 (0.28)	0.64 (0.29)	0.38 (0.23)	0.38 (0.22)	0.61 (0.28)
`ADERENTE`	0 (%)	1,358 (46.6%)	880 (59.8%)	2,676 (92.6%)	1,900 (94.7%)	2,279 (67%)
	1 (%)	1,558 (53.4%)	593 (40.2%)	214 (7.4%)	106 (5.3%)	1,120 (33%)
`PDC_CLA`	[0 ; 0.25) (%)	286 (9.8%)	202 (13.7%)	1,051 (36.4%)	666 (33.2%)	473 (13.9%)
	[0.25 ; 0.50) (%)	422 (14.4%)	286 (19.4%)	1,044 (36.1%)	778 (38.8%)	695 (20.4%)
	[0.50 ; 0.75) (%)	506 (17.4%)	288 (19.6%)	514 (17.8%)	424 (21.1%)	948 (27.9%)
	[0.75 ; 1] (%)	1,702(58.4%)	697 (47.3%)	281 (9.7%)	138 (6.9%)	1,238 (37.8%)
`curvaMG`	0 (%)	152 (5.2%)	201 (13.6%)	5 (0.2%)	996 (49.7%)	0 (0%)
	1 (%)	2,764 (94.8%)	1,272 (86.4%)	2,885 (99.8%)	1,010 (50.3%)	3,399 (100%)

Table 4.5: Final RL datasets summaries for each pharmacological class.

4.1.2 Functional K-mean

In order to verify if the time-varying curves introduced in Sections 2.1.5 and 2.1.6 highlight some differences in terms of adherence, we perform a Functional K-mean on curves describing the drug assumption, as described in [15].

The results for $k = 2$ (best option according to the silhouette plot) are reported in Figures 4.6 and 4.7, respectively for curves of cumulative days covered by drug assumption and for curves of assumed dose. In particular, Figure 4.6 (left panel) shows the final centroids (orange and red solid lines) and the final clusters for curves of cumulative days (right panel), whereas Figure 4.7 shows the final centroids (left panel, yellow and pink solid lines) and the final clusters for curves of assumed dose (right panel).

Since we are interested in evaluating adherence, we consider adherent patients (Figure 4.8) and adherence levels (Figure 4.9) for each cluster. It can be stated that the curves of cumulative days covered by drug assumption fully represent adherence characteristics: in fact the orange cluster represents patients with highest PDC. On the contrary, the curves of assumed dose do not express that information correctly. In fact, as it can evinced by the barplot in Figure 4.9, left panel shows that as PDC class decreases as the number of patients of the orange cluster decreases and the number of patients of the red cluster increases, whereas in right panel the number of patients of the two clusters (yellow and pink) is equally distributed into PDC classes in proportion of their dimensionalities. In particular, left panel shows that the highest PDC class (75-100%) is composed by 1,702 curves, all belonging to the orange cluster, whereas right panel shows that the highest PDC class (75-100%) is composed by 1,644 curves, 1,166 (70.2%) belonging to the pink cluster and 438 (29.1%) belonging to the yellow one. For this reason from now on we consider only the curves of cumulative days covered by drug assumption.

In Figure 4.10 (left panel) we observe that the presence of male and female is equally distributed in the clusters in proportion of their dimensionalities. The patients' age (right panel) in the two clusters are different (p-value of Wilcoxon test is equal to $8.475 \cdot 10^{-5}$).
In Figure 4.11, we compare hospitalizations (left panel) and comorbidities (right panel) in the two clusters. In particular we observe that 773 (39.1%) patients of orange cluster and 346 (36.8%) of red cluster have only the first hospitalization. Moreover, at the first hospitalization 65 (3.3%) patients of orange cluster and 26 (2.8%) of red cluster have no comorbidity, 1,750 (88.5%) patients of orange cluster and 820 (87.3%) of red cluster have one, two or three comorbidities and 162 (8.2%) patients of orange cluster and 93 (9.9%) of red cluster have more than

three comorbidities. Therefore the number of hospitalizations and of comorbidities is equally distributed in the two clusters in proportion of their dimensionalities.

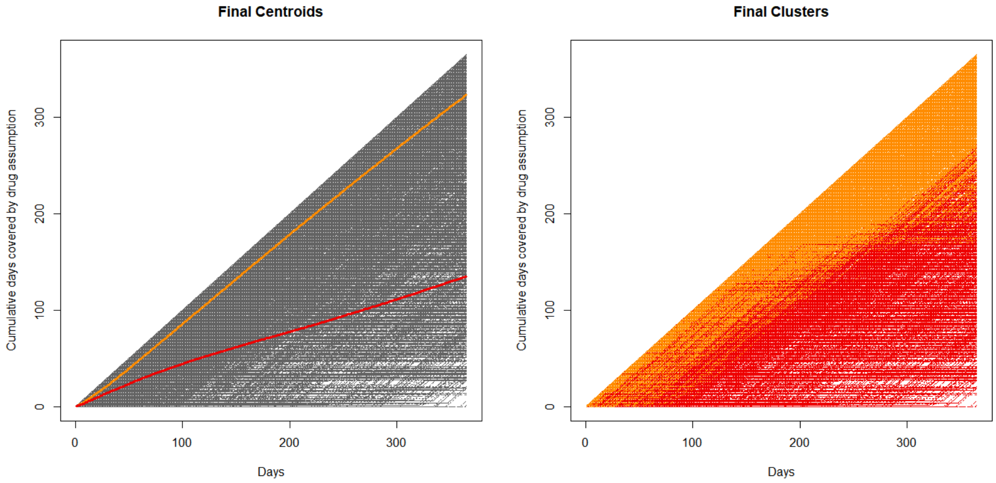


Figure 4.6: Matplots of curve of cumulative days covered by drug assumption for functional 2-mean. Left panel shows the final centroids (orange and red solid lines) and right panel shows the final clusters.

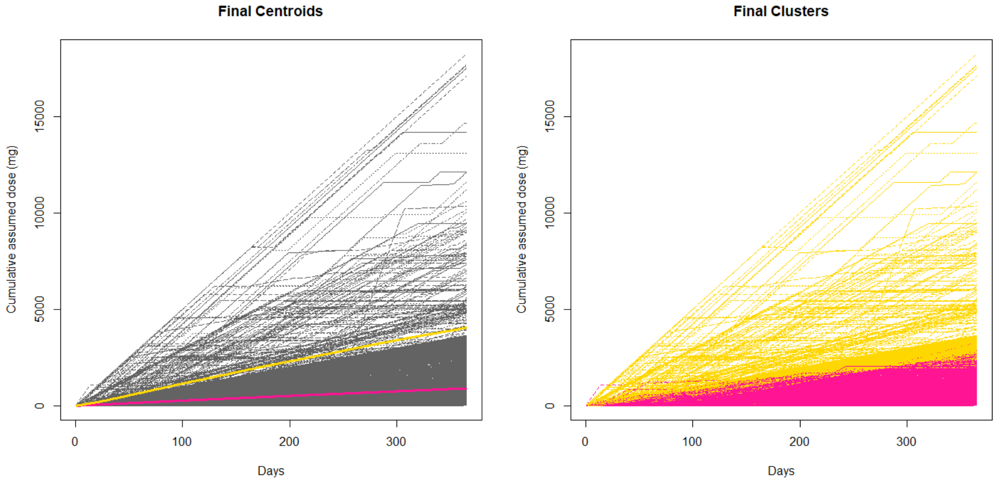


Figure 4.7: Matplots of curve of cumulative assumed dose for functional 2-mean. Left panel shows the final centroids (yellow and pink solid lines) and right panel shows the final clusters.

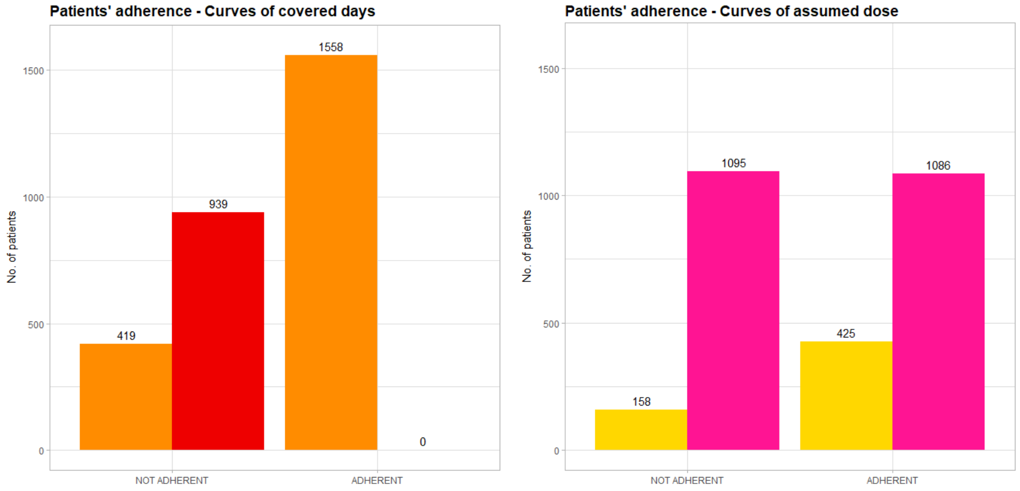


Figure 4.8: Barplots of adherent patients divided into clusters. Left panel shows the number of adherent patients related to curves of cumulative days covered by drug assumption. Right panel shows the number of adherent patients related to curves of assumed dose.

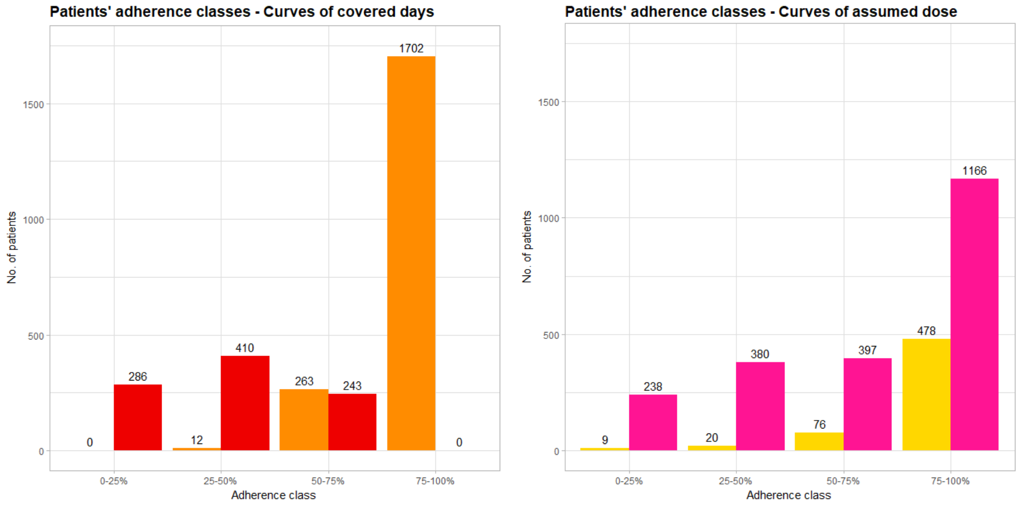


Figure 4.9: Barplots of adherence levels divided into clusters. Left panel shows the number of patients related to curves of cumulative days covered by drug assumption and divided into adherence levels. Right panel shows the number of patients related to curves of assumed dose and divided into adherence levels.

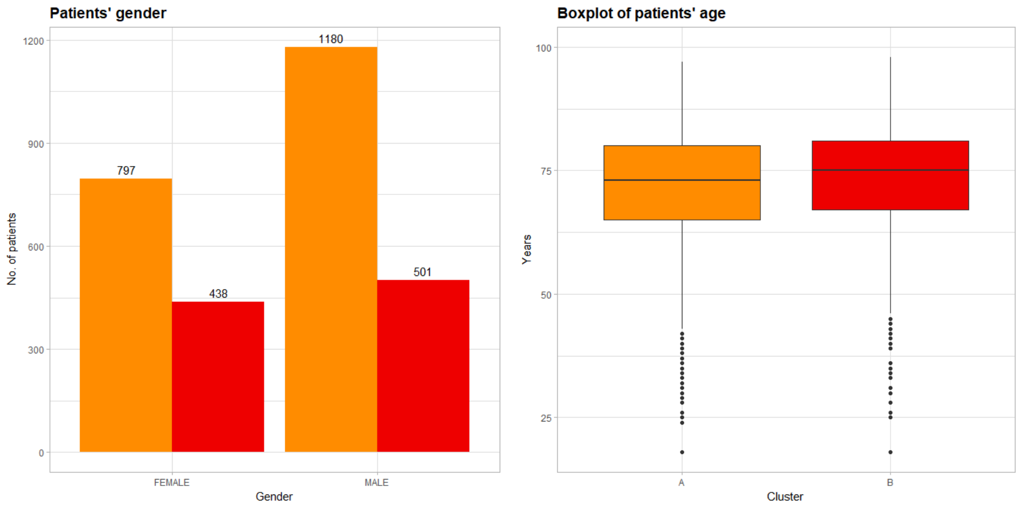


Figure 4.10: Left panel shows barplot of male and female divided into clusters. Right panel shows boxplots of patients' age divided into clusters

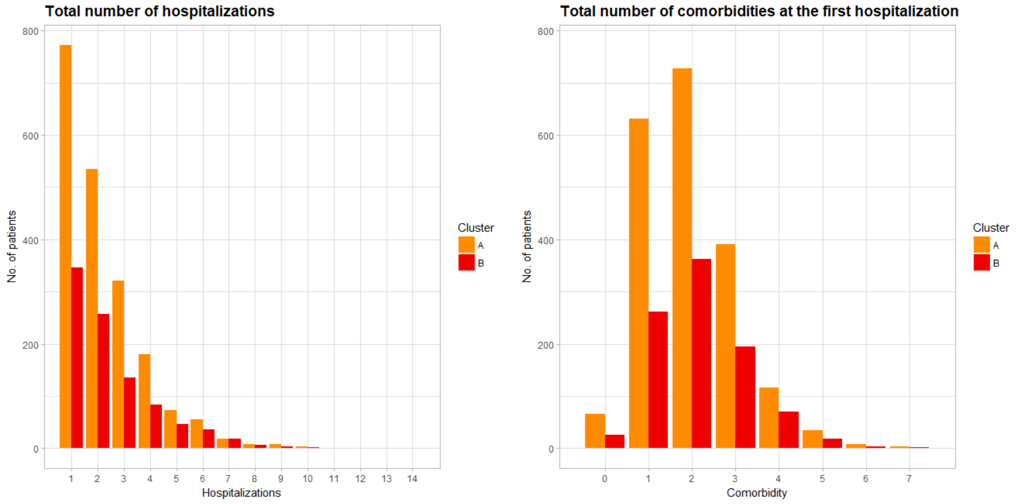


Figure 4.11: Left panel shows the barplot of the total number of hospitalizations stratified by the clusters pointed out by the K-mean procedure. Right panel shows the barplot of the total number of comorbidities at the first hospitalization stratified by the clusters pointed out by the K-mean procedure.

4.1.3 Cox's PH Model with adherence binary variable

In order to assess the role of available covariates with respect to the overall survival time of a patient, we perform a Cox's regression models.
Following notation introduced in Chapter 3, we choose five fixed covariates

$$\boldsymbol{X}_j\ (j = 1, ..., 5) = \{\texttt{age\_in},\ \texttt{gender},\ \texttt{tot\_hosp},\ \texttt{comorbidity},\ \texttt{ADERENTE}\}$$

Specifically:

- age_in is the patient's age at the first hospitalization
- gender is the gender of the patient
- tot_hosp is the total number of hospitalizations
- comorbidity is the number of comorbidities at the first hospitalization
- ADERENTE is the binary variable indicating whether a patient is adherent or non adherent, computed as specified in Section 2.1.4

Table 4.6 summaries the main features of the aforementioned covariates and of death and timeOUT covariates, which respectively represent the binary variable that indicates whether a patient is dead or not at the end of the study and the follow up time [days] of the patient. In particular, death is our output variable.

Table 4.7 reports the summary of the Cox's model. From p-values we note that all the covariates are statistically significant at confidence level $\alpha = 5\%$, though there is not a strong evidence for a gender effect. These results are also confirmed by stratified log-rank tests for which p-values are reported in Table 4.8. In fact, in this case p-value for gender is 0.109.

The covariate age_in is strongly significant and being younger leads to a higher survival probability, as it was reasonable to expect. Figure 4.12 also confirms this results. In fact, it represents survival stratified by age. Categories are selected according to the following criteria: we consider female patients with two comorbidities, two hospitalizations and aged 45, 65, 80 and 90 years. We observe that the higher the age, the lower the survival. Moreover, having a higher age leads to larger confidence intervals over time so the uncertainty about the prediction of the survival outcome increases. The p-value of the log-rank test (Table 4.8) for age stratified by *junior* ($\texttt{age\_in} < 65$), *senior* ($65 \leq \texttt{age\_in} \leq 85$), *old senior* ($\texttt{age\_in} > 85$) is $< 2 \cdot 10^{-16}$.

Variable	Value	
No. patients		2,916
death	0 (%) 1 (%)	2,198 (75.4%) 718 (24.6%)
timeOUT	mean (sd)	1,543.45 (613.46)
age_in	mean (sd)	72.17 (11.44)
gender	*Male* (%) *Female* (%)	1,681 (57.6%) 1,235 (42.4%)
tot_hosp	mean (sd)	2.35 (1.61)
comorbidity	mean (sd)	2.04 (1.07)
ADERENTE	0 (%) 1 (%)	1,358 (46.6%) 1,558 (53.4%)

Table 4.6: ACE Inhibitors dataset for Cox's model with fixed covariates.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
age_in	0.067860	1.070215	0.004668	14.536	< 2e-16	***
genderM	0.177182	1.193849	0.077617	2.283	0.022443	*
tot_hosp	0.113542	1.120239	0.020826	5.452	4.98e-08	***
comorbidity	0.169536	1.184755	0.032814	5.167	2.38e-07	***
ADERENTE1	-0.258985	0.771834	0.075390	-3.435	0.000592	***

Table 4.7: Summary of the Cox's model for overall survival time with fixed covariates only.

Variable	Stratification	Value	Pts	p-value
Age	*junior* *senior* *old senior*	age_in <65 $65 \leq$ age_in ≤ 85 age_in >85	624 2,047 245	$< 2 \cdot 10^{-16}$

Gender	*female*
	male
		gender = F
		gender = M
			1,235
			1,681
				0.109

Hospitalizations	*only one* *few* *many*	tot_hosp = 1 $2 \leq$ tot_hosp ≤ 3 tot_hosp ≥ 4	1,119 1,249 548	$1.86 \cdot 10^{-4}$
Comorbidities	*low* *medium* *high*	comorbidity <2 $2 \leq$ comorbidity ≤ 3 comorbidity ≥ 4	984 1,677 255	$1.71 \cdot 10^{-13}$

Adherence	*non adherent*
	adherent
		ADERENTE = 0
		ADERENTE = 1
			1,358
			1,558
				$1.24 \cdot 10^{-6}$

Table 4.8: P-values of stratified log-rank tests.

The covariate `ADERENTE` is significant and the Hazard Ratio (HR) for adherent patients, given by the exponentiated coefficient in Table 4.7, is $0.772 < 1$. This means that, as we expected, being adherent increases the survival probability. Figure 4.13 shows this results through the KM of survival stratified by adherent and non adherent patients. Categories are selected according to the following criteria: we consider female patients aged 72 years, with two comorbidities and two hospitalizations. The p-value of the log-rank test (Table 4.8) stratified by *adherent* and *non adherent* patients is $1.24 \cdot 10^{-6}$ providing strong evidence for this covariate to be significant.

The covariate `tot_hosp` is strongly significant and a higher number of hospitalizations corresponds to a lower survival probability. This result is also confirmed by Figure 4.14 which shows KM of survival stratified by the total number of hospitalizations. Categories are selected according to the following criteria: we consider adherent female patients aged 72 years, with two comorbidities and 1, 5 and 10 hospitalizations, respectively. We observe that the higher the number, the lower the survival. This is due to the fact that a higher number of hospitalizations, especially in elderly patients, is probably an index of deterioration in patient's state of health. Moreover, being hospitalized more times leads to larger confidence intervals over time so the uncertainty about the prediction of the survival outcome increases. The p-value of the log-rank test (Table 4.8) for number of hospitalizations stratified by *only one* ($\texttt{tot\_hosp} = 1$), *few* ($2 \leq \texttt{tot\_hosp} \leq 3$), *many* ($\texttt{tot\_hosp} \geq 4$) is $1.86 \cdot 10^{-4}$.

Similarly the covariate `comorbidity` is strongly significant and a higher number of comorbidities at the first hospitalization corresponds to a lower survival probability. Figure 4.15 also confirms this results. It represents survival stratified by the total number of comorbidities at the first hospitalization. We consider adherent female patients aged 72 years, with two hospitalizations and 0, 4 and 8 comorbidities, respectively. Also in this case we observe that the higher the number, the lower the survival. This is somehow expected, since a higher comorbidity often indicates more critical clinical prognosis. Moreover, having more comorbidities leads to larger confidence intervals over time so the uncertainty about the prediction of the survival outcome increases. The p-value of the log-rank test (Table 4.8) for comorbidities stratified by *low* ($\texttt{comorbidity} < 2$), *medium* ($2 \leq \texttt{comorbidity} \leq 3$), *high* ($\texttt{comorbidity} \geq 4$) is $1.71 \cdot 10^{-13}$.

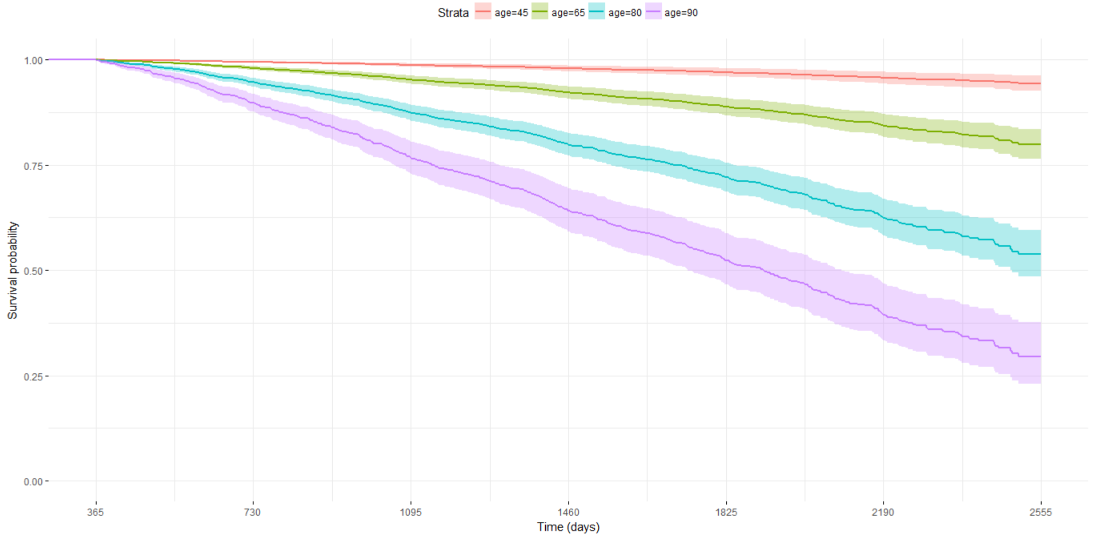


Figure 4.12: Survival probability plot for adherent female patients with two comorbidities, two hospitalizations and that are 45, 65, 80 and 90 years old.

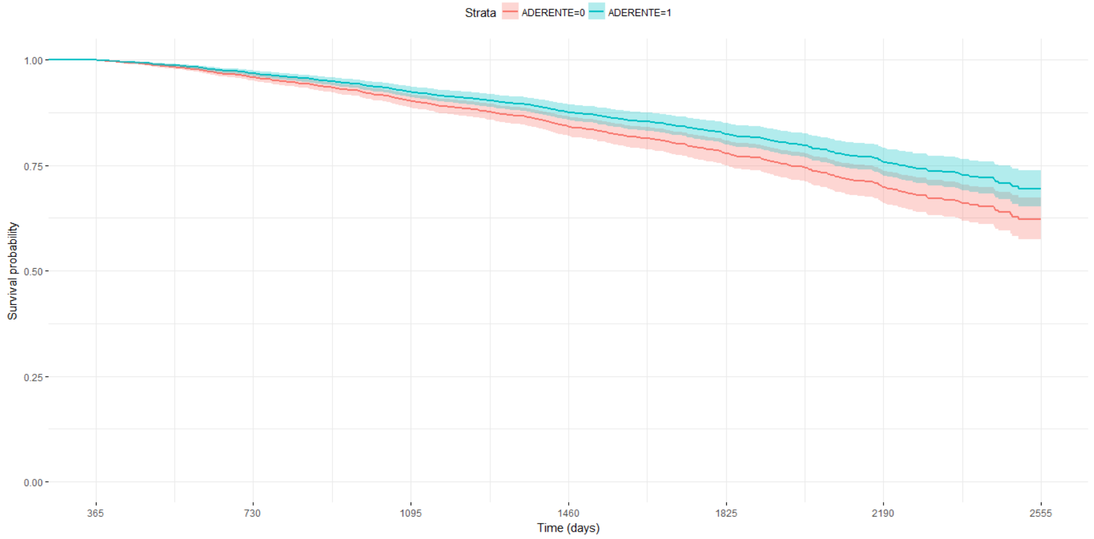


Figure 4.13: Survival probability plot for adherent vs non adherent female patients with 72 years old, two comorbidities and two hospitalizations.

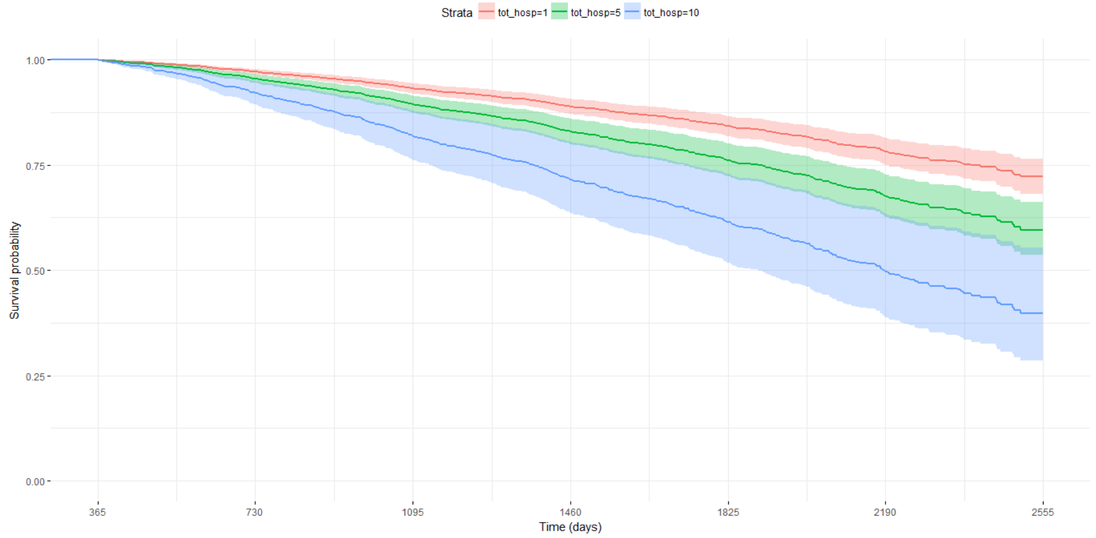


Figure 4.14: Survival probability plot for adherent female patients with 72 years old, two comorbidities and 1, 5 and 10 hospitalizations.

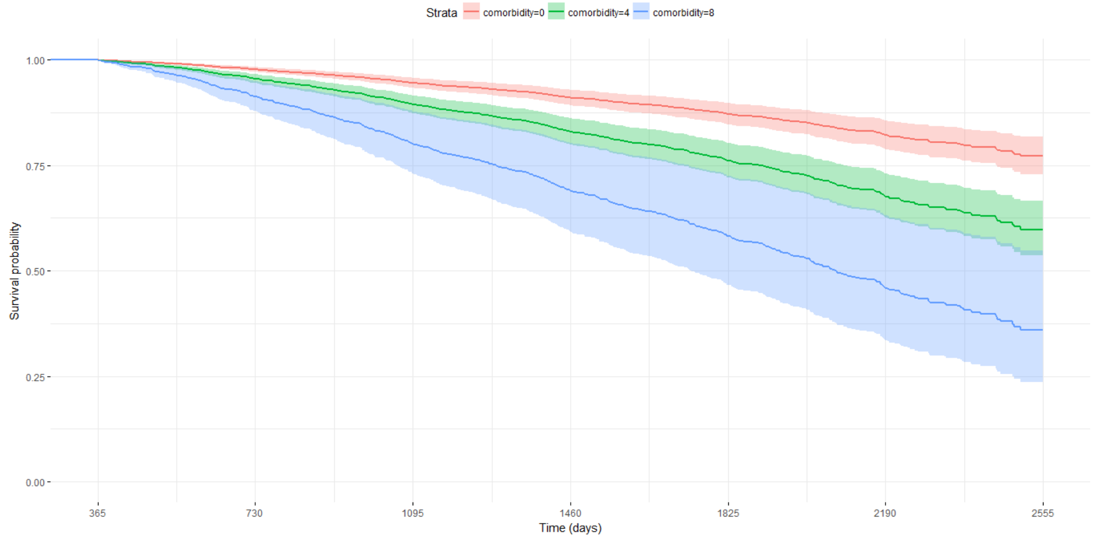


Figure 4.15: Survival probability plot for adherent female patients with 72 years old, two hospitalizations and 0, 4 and 8 comorbidities.

4.1.4 Joint Modelling of re-hospitalization and drug consumption for HF patients

Since we aim at making up time-dependent covariates to be inserted and treated in a survival model with innovative statistical approach, we proceed with the joint models technique introduced in Section 3.2 and in [26].

For the survival part of the model, we consider the same covariates used in Section 4.1.3 but, instead of dichotomized adherence, we study the effect of curves of cumulative days covered by drug assumption as a secondary joint process. Therefore, in our analysis the longitudinal process in given by these curves, whereas the event process is thought as dependent on age, gender, number of hospitalizations and comorbidities. In particular, following notation introduced in Section 3.2.1, we set:

- $m_i(t)$ = square root of the value of cumulative days curve[1]
- $\boldsymbol{X}_i = (\texttt{age\_in},\ \texttt{gender},\ \texttt{tot\_hosp},\ \texttt{comorbidity})_i$
- $\lambda_0(\cdot)$ = piecewise-constant baseline risk function given by (3.24)

We also remind that we use the Gauss-Hermite integration rule to approximate integral (3.20).

The summaries of both longitudinal and event processes are shown in Tables 4.9 and 4.10, respectively. We note that all the covariates are significant at confidence level $\alpha = 5\%$, except for `gender`. The parameter labeled as `Assoct` corresponds to parameter α in Equation (3.16), which measures the effect of $m_i(t)$ on the risk of death.

We proceed by checking the fit of the model using residuals plots (see Section 3.2.4). From the residuals for the longitudinal process in Figure 4.16, it evinces that the hypotheses of normally distributed random effects b_i and measurement error terms $\varepsilon_i(t)$ are not fully satisfied. On the contrary, Figure 4.17, about Cox-Snell residuals for the event process, shows that an appropriate functional form for covariates is used in the model.

Despite the non optimality of goodness of fit results, we decide to go on further for getting insights of the predictions provided by the JM tool.

[1] For computational reasons, as explained in [26], it is necessary to perform an ad hoc data transformation, here represented by the square root.

	Value	Std.Err	z-value	p-value	
(Intercept)	2.6937	0.0028	952.6757	<0.0001	***
obstime	0.0419	0.0000	3137.5414	<0.0001	***

Table 4.9: Summary of the JM longitudinal process for ACE Inhibitors.

	Value	Std.Err	z-value	p-value	
age_in	0.0610	0.0045	13.4563	<0.0001	***
genderM	0.1293	0.0774	1.6704	0.0948	.
tot_hosp	0.1137	0.0207	5.4853	<0.0001	***
comorbidity	0.1669	0.0328	5.0881	<0.0001	***
Assoct	0.0052	0.0021	2.4569	0.0140	*

Table 4.10: Summary of the JM event process for ACE Inhibitors.

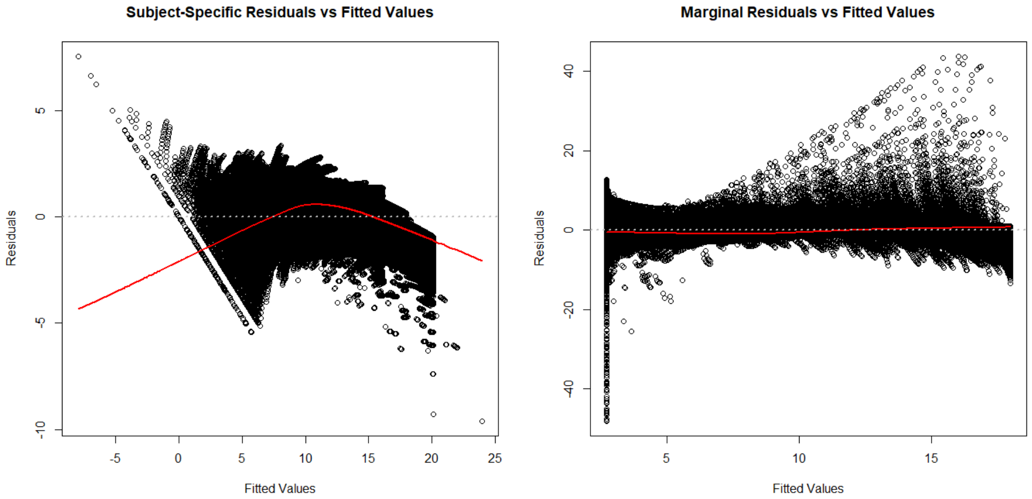


Figure 4.16: Diagnostic plots for the fitted joint model. The left panel depicts the subject-specific residuals for the longitudinal process versus their corresponding fitted values. The right panel depicts the standardized marginal residuals for the longitudinal process versus their corresponding fitted values.

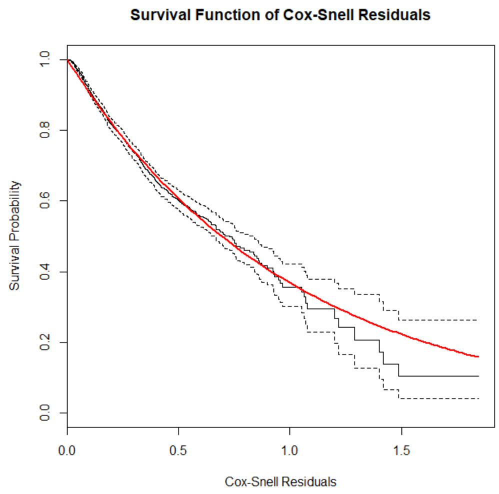


Figure 4.17: Kaplan-Meier estimate of the Cox-Snell residuals for the event process. The dashed lines denote the 95% confidence intervals.

In particular, we focused on the calculation of expected survival probabilities. We compute $\pi_i(u|t)$ for patients in the dataset who have not died by the censored time, using $L = 200$ Monte Carlo samples. For each significant covariate we consider a set of four real patients, presenting different values for the covariate of interest and similar/equal values for the remaining ones. Patients selection is reported in Tables 4.11, 4.12, 4.13 and 4.14.

COD_REG	timeOUT	age_in	gender	tot_hosp	comorbidity	PDC	day_365
14836448	1360	43	M	1	2	0.991	362
11267026	2422	57	M	1	2	0.997	364
14615697	1126	71	M	1	2	1.00	365
21874674	1620	89	M	1	2	1.00	365

Table 4.11: Patients' data used for JM survival plots on different age values.

COD_REG	timeOUT	age_in	gender	tot_hosp	comorbidity	PDC	day_365
19634104	1651	66	M	1	2	0.997	364
21595517	776	66	M	2	2	1.00	365
21506299	2033	67	M	4	2	0.989	361
21338024	2520	68	M	6	2	0.997	364

Table 4.12: Patients' data used for JM survival plots on different hospitalization values.

COD_REG	timeOUT	age_in	gender	tot_hosp	comorbidity	PDC	day_365
21180034	632	80	M	2	0	0.973	355
13211878	2194	80	M	2	1	0.986	360
18893559	2278	81	M	2	2	0.981	358
12415236	1502	80	M	2	4	0.997	364

Table 4.13: Patients' data used for JM survival plots on different comorbidity values.

COD_REG	timeOUT	age_in	gender	tot_hosp	comorbidity	PDC	day_365
11045618	1438	80	F	1	2	0.115	42
10966162	798	80	F	1	2	0.274	100
11525460	858	80	F	1	2	0.833	304
21251850	635	80	F	1	2	0.967	353

Table 4.14: Patients' data used for JM survival plots on different curves of cumulative days covered by drug assumption.

Figure 4.18 shows results of predictions related to the four patients in Table 4.11, i.e., when age varies given all the other covariates as "fixed" in the sense explained above. Moving from left to right panels, survival for elder people is shown. We observe that being younger corresponds to an higher survival probability, as we could have expected.

Figure 4.19 shows results of predictions related to the four patients in Table 4.12, i.e., when the total number of hospitalizations varies given all the other covariates as "fixed" in the sense explained above. Moving from left to right panels, survival for increasing number of hospitalizations is shown. We observe that an higher number of hospitalizations corresponds to a lower survival probability.

Figure 4.20 shows results of predictions related to the four patients in Table 4.13, i.e., when the number of comorbidities at the first hospitalization varies given all the other covariates as "fixed" in the sense explained above. Moving from left to right panels, survival for increasing number of comorbidities is shown. We observe that an higher number of comorbidities corresponds to a lower survival probability.

Figure 4.21 shows results of predictions related to the four patients in Table 4.14, i.e., considering different curves of cumulative days covered by drug assumption given all the other covariates as "fixed" in the sense explained above. We observe that having a curve of days covered with an higher final value, and so an higher PDC, correspond to an higher survival probability, as we could have expected. In fact, moving from left to right panels, survival for increasing PDCs is shown. Furthermore, having a lower PDC leads to larger confidence intervals over time, as reported in Tables 4.15, 4.16 and 4.17, so the uncertainty about the prediction of the survival outcome increases.

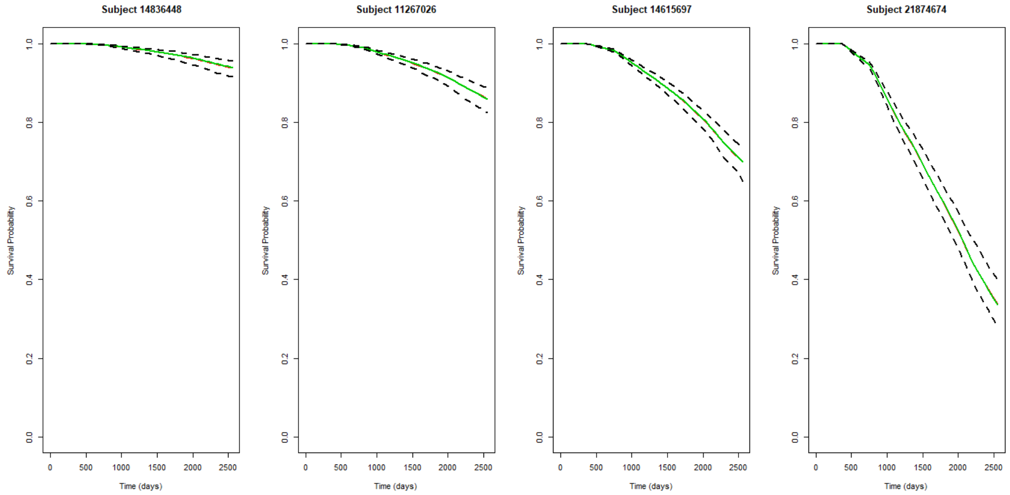


Figure 4.18: Survival probability plots for male patients with one hospitalizations, two comorbidities and PDC greater than 0.99. From the left panel patients are 43, 57, 71 and 89 years old.

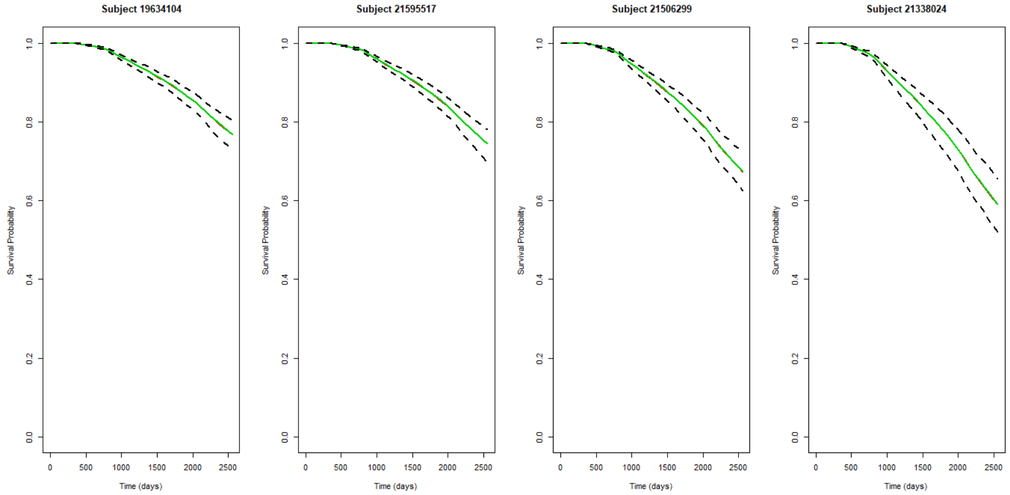


Figure 4.19: Survival probability plots for male patients with 66/67/68 years old, two comorbidities and PDC greater than 0.98. From the left panel patients have 1, 2, 4 and 6 hospitalizations.

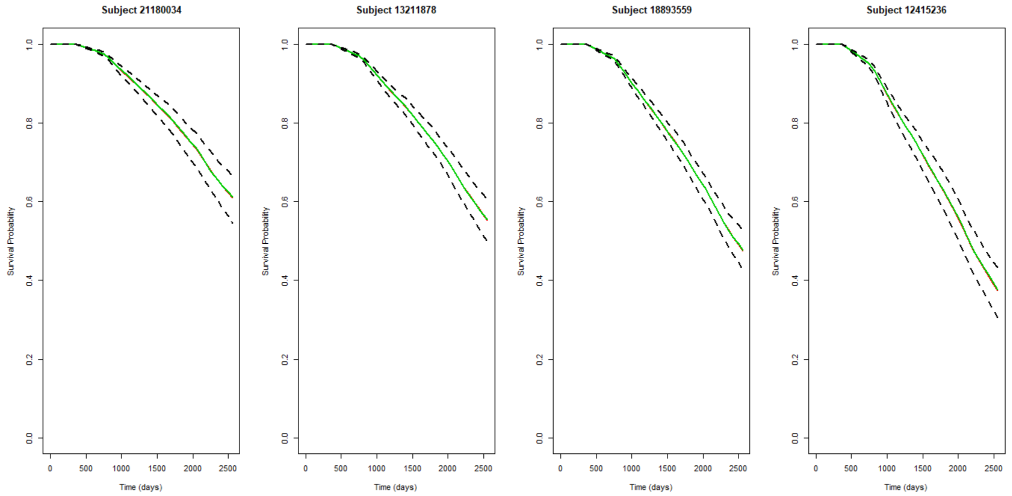


Figure 4.20: Survival probability plots for male patients with 80/81 years old, two hospitalizations and PDC greater than 0.97. From the left panel patients have 0, 1, 2 and 4 comorbidities.

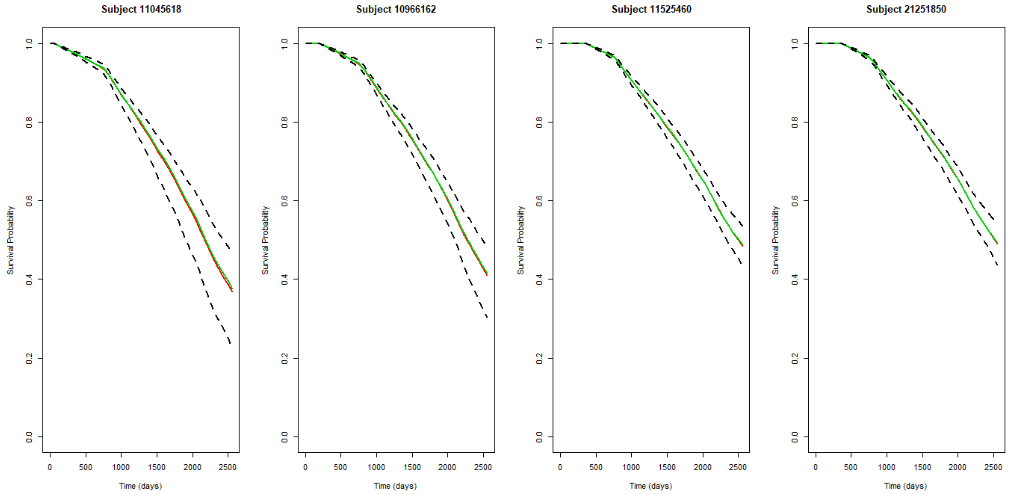


Figure 4.21: Survival probability plots for female patients with 80 years old, one hospitalization and two comorbidities. From the left panel patients have a PDC of 0.115, 0.274, 0.833 and 0.967.

COD_REG	PDC	Time	Mean	Median	Lower	Upper
11045618	0.115	1 year	0.9350	0.9357	0.9217	0.9463
10966162	0.274	1 year	0.9484	0.9488	0.9389	0.9572
11525460	0.833	1 year	0.9641	0.9643	0.9579	0.9703
21251850	0.967	1 year	0.9640	0.9642	0.9577	0.9702

Table 4.15: Mean and median values of patients' survival probabilities one year after the end of the follow up with the respective PDC and 95% confidence intervals.

COD_REG	PDC	Time	Mean	Median	Lower	Upper
11045618	0.115	3 years	0.7422	0.7472	0.6828	0.7773
10966162	0.274	3 years	0.7662	0.7682	0.7303	0.7930
11525460	0.833	3 years	0.7967	0.7981	0.7706	0.8205
21251850	0.967	3 years	0.7972	0.7989	0.7713	0.8207

Table 4.16: Mean and median values of patients' survival probabilities three years after the end of the follow up with the respective PDC and 95% confidence intervals.

COD_REG	PDC	Time	Mean	Median	Lower	Upper
11045618	0.115	5 years	0.4991	0.5055	0.3808	0.5815
10966162	0.274	5 years	0.5370	0.5421	0.4675	0.5934
11525460	0.833	5 years	0.5962	0.5953	0.5569	0.6304
21251850	0.967	5 years	0.5994	0.5995	0.5605	0.6330

Table 4.17: Mean and median values of patients' survival probabilities five years after the end of the follow up with the respective PDC and 95% confidence intervals.

Given these analyses, we can conclude that modelling the drug assumption process as time-varying covariates in a joint model setting is a richer inferential instrument than a simple Cox's model. In fact, it is an interpretative and forecasting tool for exploring the effects of pharmacological treatments on survival. For example, it allows us to confirm some pharmacoepidemiological intuition as the fact that medication nonadherence is commonly associated with adverse health conditions [18] in a more suitable way. Moreover, such a model enables a tailored prediction for different patients profiles.

4.2 Friuli Venezia Giulia Region (FVGR)

In this Section we start with a descriptive analysis of the FVGR dataset. We proceed with a Functional K-means technique in order to cluster the time-varying curves introduced in Sections 2.2.6 and 2.2.7. Then we perform a Cox Proportional Hazards regression analysis using only survival data including the binary variable for adherence. Finally, we apply the Joint Models method described in Section 3.2 for the analysis of both longitudinal and survival data.

4.2.1 Descriptive analysis

In this part we would like to introduce a statistical description of the FVGR database, remembering that 13,619 patients are present with a total of 218,843 events, as described in Section 2.2.2.

First of all, we can notice that in FVGR dataset information about 6,324 males (46.4%) and 7,295 females (53.6%) are collected and that the number of dead patients is 6,661 (48.9%). Table 4.18 also reports the number (%) of patients censored at the end to follow up. Figure 4.22 shows the same information through a barplot. The follow up time ranges from one to eight years, with a mean of about three years (1,095 days), as summarized in Table 4.19.

As it can be seen in Table 4.20 and Figure 4.23, the age of patients ranges from 26 to 106 years, being females older than males (p-value of one-side Wilcoxon test is $< 2{\cdot}10^{-16}$). In particular 7,286 (53.5%) patients are between 75 and 87 years old.

Pts	All	*Dead* (%)	*Censored* (%)
All	13,619	6,661 (48.9%)	6,958 (51.1%)
Female	7,295	3,657 (50.1%)	3,638 (49.9%)
Male	6,324	3,004 (47.5%)	3,320 (52.5%)

Table 4.18: Total number of dead and truncated patients of FVGR dataset.

Follow up time [days]						
Pts	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All	365	695.5	1,111	1,274	1,760	2,917
Female	365	696	1,113	1,272	1,752	2,912
Male	365	694	1,108	1,277	1,769	2,917

Table 4.19: Summary of the FVGR patients' follow up time (expressed in days).

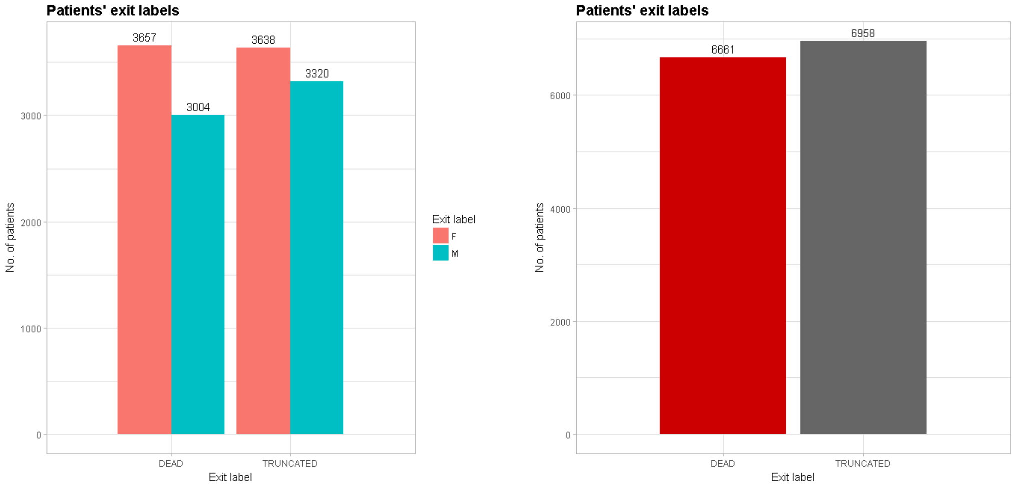


Figure 4.22: Barplots of FVGR exit status: patients divided by gender (left panel) and all patients (right panel).

Age at the first hospitalization [years]						
Pts	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All	26	75	82	79.99	87	106
Female	37	78	84	82.6	88	106
Male	26	71	78	76.99	84	102

Table 4.20: Summary of the FVGR patients' age at the first hospitalization.

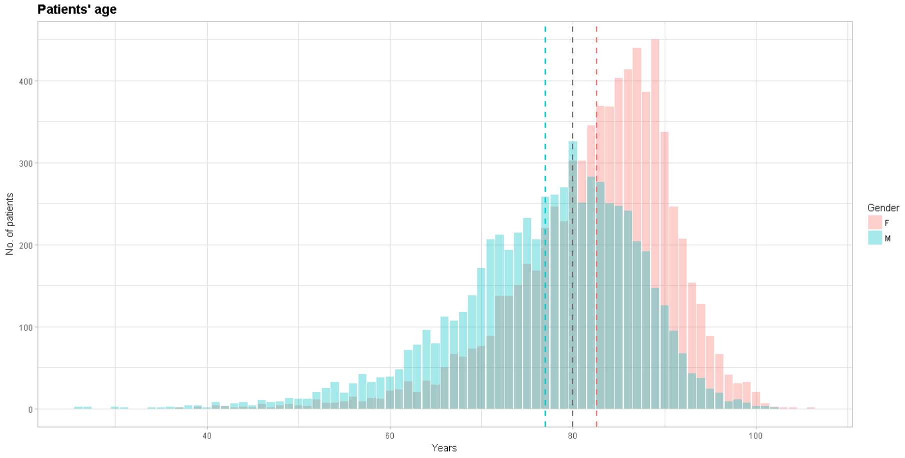


Figure 4.23: Snapshot of the FVGR patients' age distribution at their first hospitalization. Gray, blue and pink dotted lines represent the mean age of all patients, male patients and female patients, respectively.

The average number of hospitalizations per patient is equal to 2.05 and 6,298 (46.2%) patients have only the first hospitalization, which is not considered in the computation of coverage days, so the adherence information related to those patients came from pharmacological events only, as described in Sections 2.1.4 and 2.2.5. The Charlson comorbidity indices at the first hospitalization per patient is equal 2.08. In particular, 2,120 (15.5%) patients have an index equal to zero, 9,119 (67%) have an index of one, two or three and 2,380 (17.5%) patients have an index greater that three. Moreover, the average number of procedures at the first hospitalization per patient is 0.41. In particular, only 5,524 (40.6%) patients underwent to at least one procedure, whereas 8,095 (59.4%) did not undergo to any procedure. All these data are reported in Table 4.21 and Figures 4.24 and 4.25. In Figure 4.25 (right panel) we also observe that 11,885 (87.3%) patients did not go in IHC, whereas 1,734 (12.7%) went in IHC at least once.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Hospitalizations	1	1	2	2.05	3	19
Charlson index	0	1	2	2.08	3	15
Procedures	0	0	0	0.41	1	3

Table 4.21: Summary of the FVGR patients' numbers of hospitalizations, comorbidities and procedures.

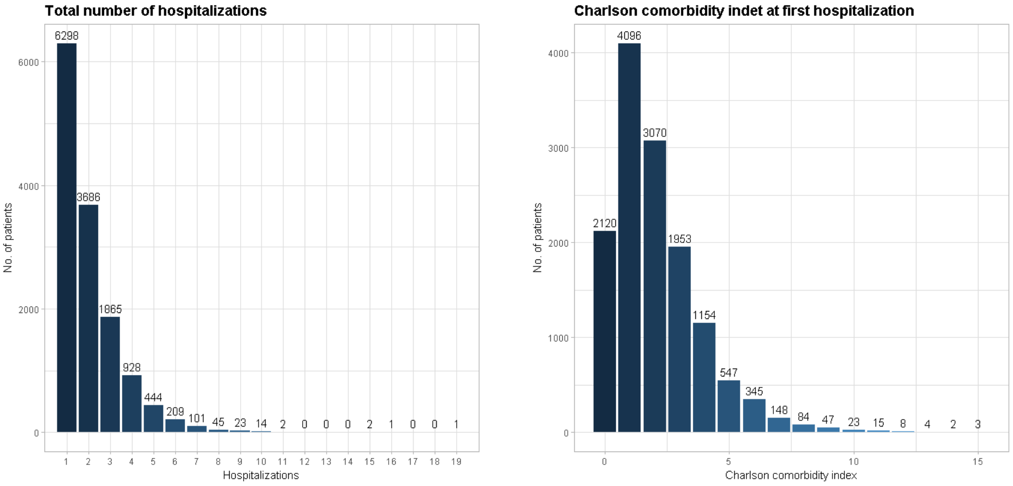


Figure 4.24: Barplots of the FVGR patients' total number of hospitalizations (left panel) and value of comorbidities at the first hospitalization (right panel).

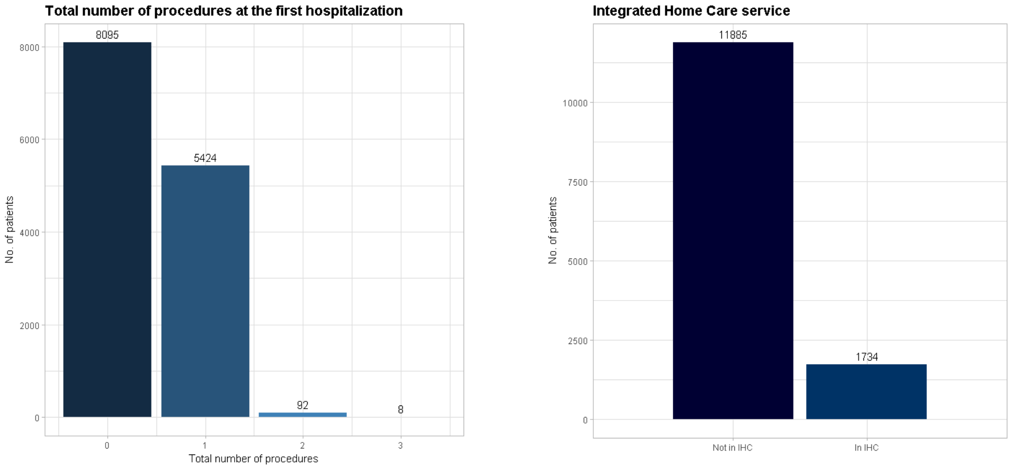


Figure 4.25: Barplot of the FVGR patients' total number of procedures at the first hospitalization (left panel) and patients' IHC service.

In Figure 4.26 (left panel) we observe that 8,481 (62.3%) patients assumes ACE Inhibitors (ACE), 6,025 (44.2%) Anti-Aldosterone agents (AA), 4,004 (29.4%) Angiotensin Receptor Blockers (ARB), 9,341 (68.6%) Beta-Blocking agents (BB) and 12,387 (91%) Diuretics (DIU). Moreover, we underline that a patient can follow different pharmacological treatments. In particular, 1,005 (7.4%) patients follow a monotherapy, 3,174 (23.3%) a bitherapy, 5,389 (39.5%) a tritherapy, 3,537 (26%) assume four different types of drugs and 514 (3.8%) take all the pharmacological classes over the observation period (1 year), as reported in Figure 4.26 (right panel).

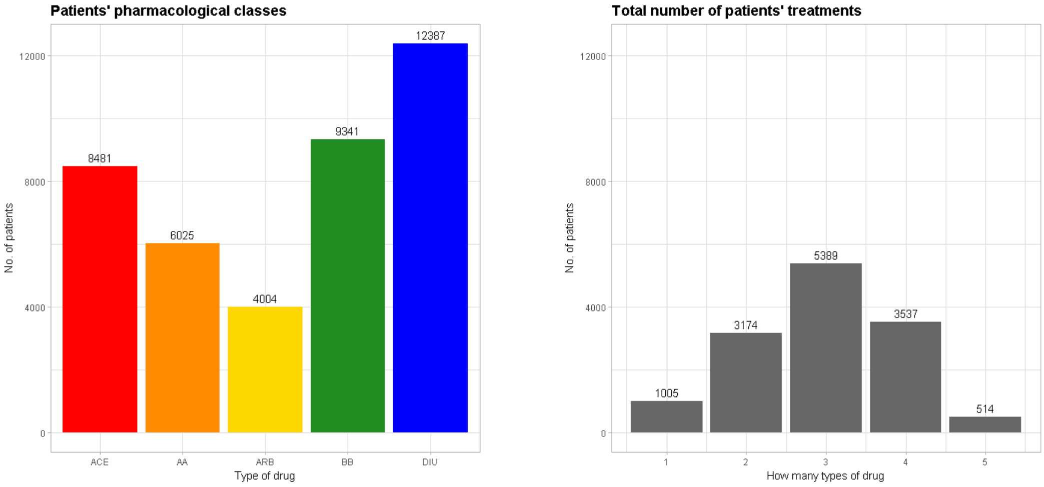


Figure 4.26: Barplots of FVGR patients' pharmacological classes (left panel) and total number of therapies (right panel).

For each approach (mixed or "one tablet a day") and for each type of drug, as mentioned in Section 2.2.8, we assemble a final dataset selecting a list of patients' peculiar features (see Table 2.31). A summary of these five final datasets is reported in Appendix E.

We remind that in the following Sections we report only the results obtained using the mixed approach (Section 2.2.4), highlighting the differences of those obtained with "one tablet a day" approach, if any.

4.2.2 Functional K-means

In order to verify if the time-varying curves introduced in Sections 2.2.6 and 2.2.7 highlight some differences in terms of adherence, we perform a Functional K-mean on curves describing the drug assumption, as described in [15].

The results for $k = 2$ (best option according to the silhouette plot) are reported in Figures 4.27 and 4.28, respectively for curves of cumulative days covered by drug assumption and for curves of assumed dose. In particular, Figure 4.27 (left panel) shows the final centroids (orange and red solid lines) and the final clusters for curves of cumulative days (right panel), whereas Figure 4.28 shows the final centroids (left panel, yellow and pink solid lines) andthe final clusters for curves of assumed dose (right panel).

Since we are interested in evaluating adherence, we consider adherent patients (Figure 4.29) and adherence levels (Figure 4.30) for each cluster. It can be stated that the curves of cumulative days covered by drug assumption fully represent adherence characteristics: in fact the orange cluster represents patients with highest PDC. On the contrary, the curves of assumed dose do not express that information correctly. In fact, as it can evinced by the barplot in Figure 4.30, left panel shows that as PDC class decreases as the number of patients of the orange cluster decreases and the number of patients of the red cluster increases, whereas in right panel the number of patients of the two clusters (yellow and pink) is equally distributed into PDC classes in proportion of their dimensionalities. In particular, we observe that the highest PDC class (75-100%) is composed by 1,917 curves. For cumulative days covered by drug assumption (left panel) all these 1,917 curves belong to the orange cluster, whereas for assumed doses (right panel), 1,516 (79.1%) curves belong to the pink cluster and 401 (20.9%) belong to the yellow one. For this reason from now on we consider only the curves of cumulative days covered by drug assumption.

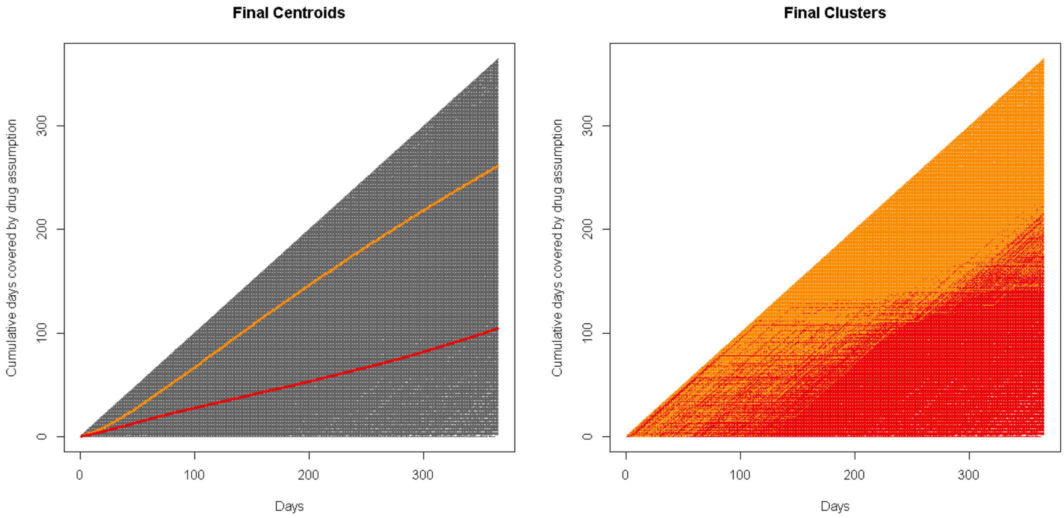


Figure 4.27: Matplots of curve of cumulative days covered by drug assumption for functional 2-mean. Left panel shows the final centroids (orange and red solid lines) and right panel shows the final clusters.

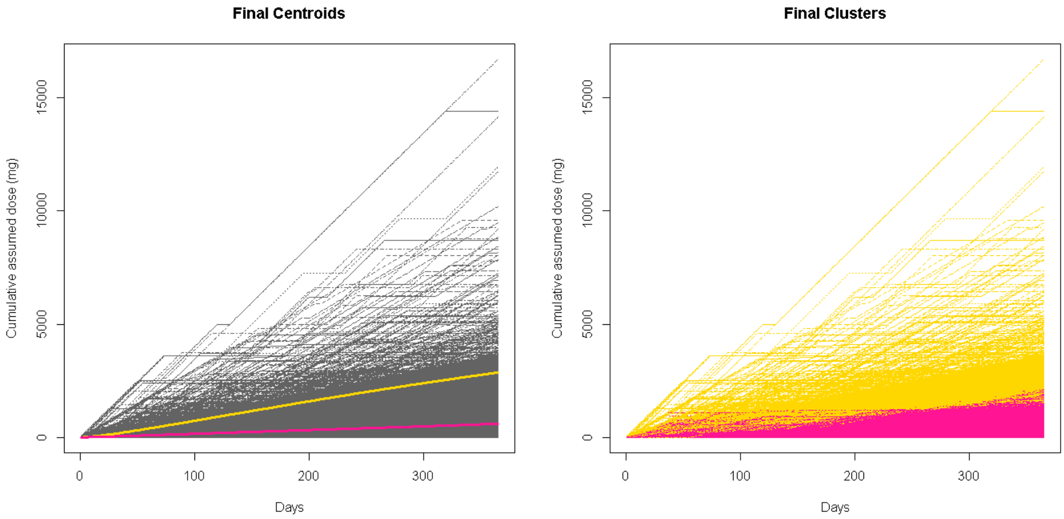


Figure 4.28: Matplots of curve of cumulative assumed dose for functional 2-mean. Left panel shows the final centroids (yellow and pink solid lines) and right panel shows the final clusters.

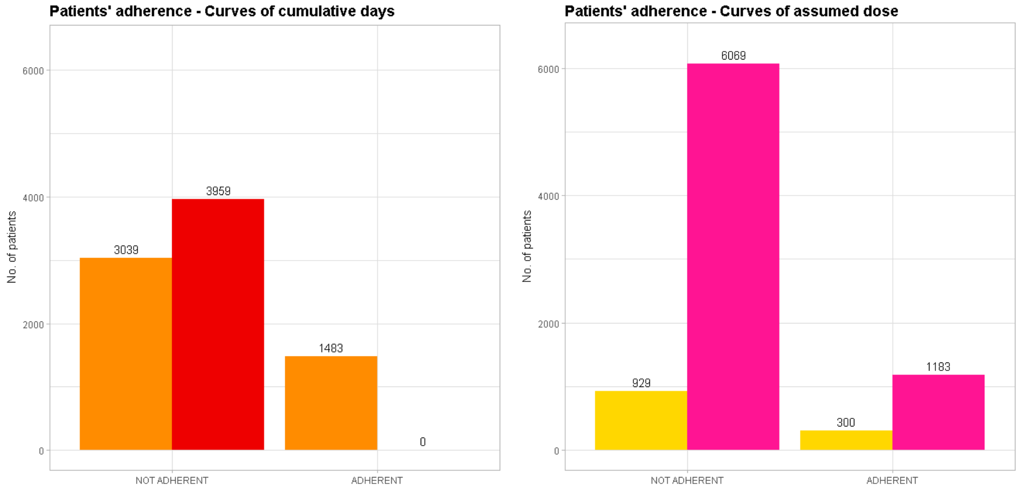


Figure 4.29: Barplots of adherent patients divided into clusters. Left panel shows the number of adherent patients related to curves of cumulative days covered by drug assumption. Right panel shows the number of adherent patients related to curves of assumed dose.

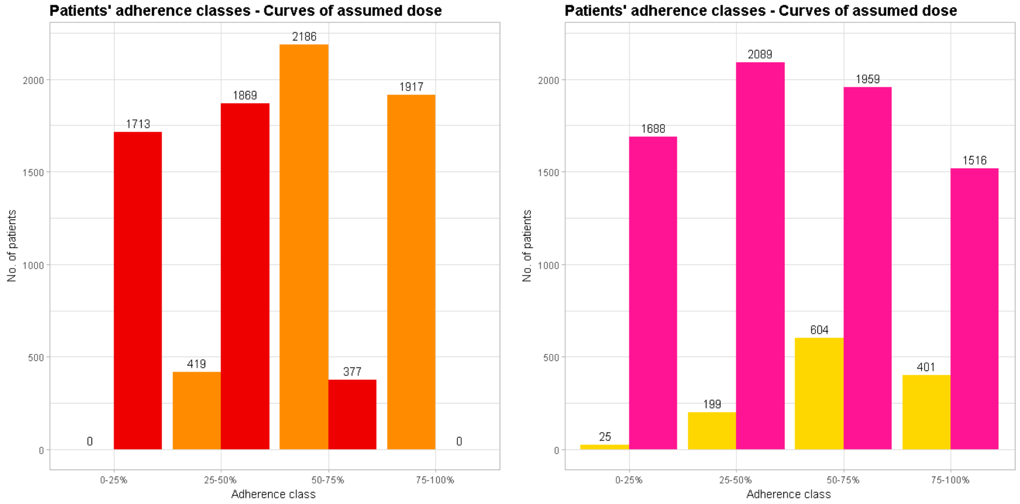


Figure 4.30: Barplots of adherence levels divided into clusters. Left panel shows the number of patients related to curves of cumulative days covered by drug assumption and divided into adherence levels. Right panel shows the number of patients related to curves of assumed dose and divided into adherence levels.

In Figure 4.31 (left panel) we observe that the presence of male and female is equally distributed in the clusters in proportion of their dimensionalities. The patients' age (right panel) in the two clusters are different (p-value of Wilcoxon test is equal to $7.814 \cdot 10^{-11}$).

In Figure 4.32, we compare hospitalizations (left panel) and Charlson comorbidity indices (right panel) in the two clusters. In particular we observe that 2,095 (46.3%) patients of orange cluster and 1,532 (38.7%) of red cluster have only the first hospitalization. Moreover, at the first hospitalization 685 (15.4%) patients of orange cluster and 599 (15.1%) of red cluster have a Charlson index of zero, 3,139 (69.4%) patients of orange cluster and 2,617 (66.1%) of red cluster have an index of one, two or three and 698 (15.4%) patients of orange cluster and 743 (18.8%) of red cluster have more than three comorbidities.
Therefore the number of hospitalizations and the values of Charlson comorbidity indices are equally distributed in the two clusters in proportion of their dimensionalities.

In Figure 4.33 we compare IHC service in the two clusters. We observe that 550 (12.2%) patients of orange cluster and 556 (14%) of red cluster went in IHC during the observation period, whereas 3,972 (87.8%) of orange cluster and 3,403 (86%) of red one did not go in IHC during the observation period. Therefore, also the dichotomized values of IHC are equally distributed in the two clusters in proportion of their dimensionalities.

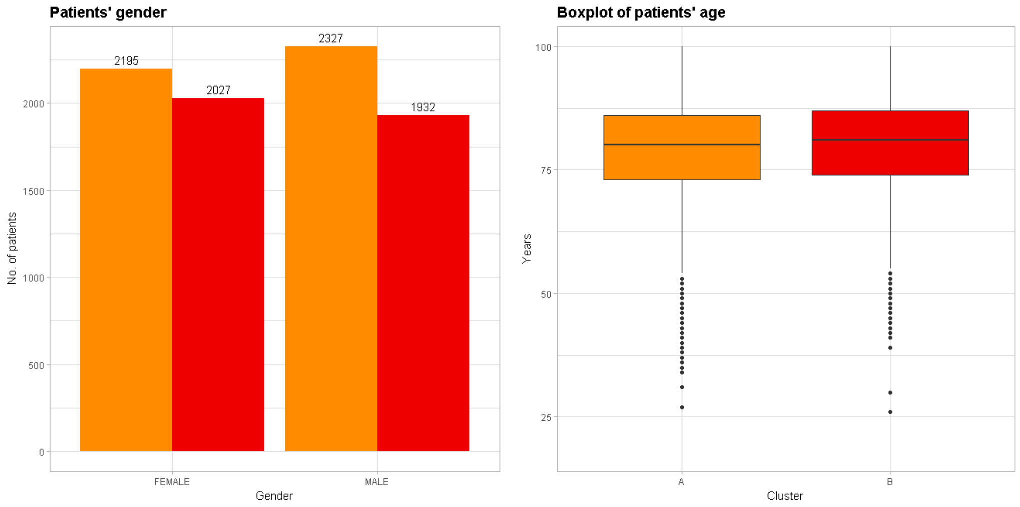


Figure 4.31: Left panel shows barplot of male and female divided into clusters. Right panel shows boxplots of patients' age divided into clusters

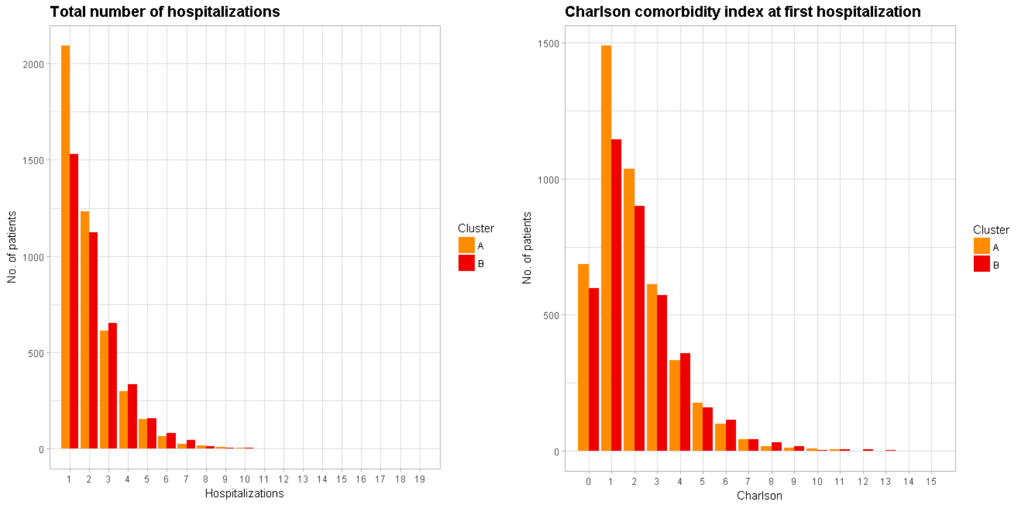


Figure 4.32: Left panel shows the barplot of the total number of hospitalizations stratified by the clusters pointed out by the K-mean procedure. Right panel shows the barplot of the Charlson comorbidity index at the first hospitalization stratified by the clusters pointed out by the K-mean procedure.

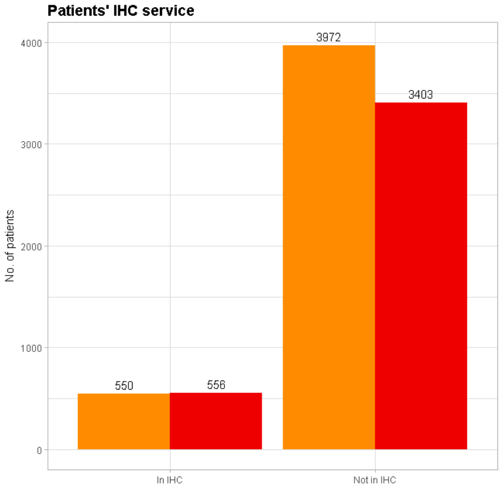


Figure 4.33: Barplot of IHC service stratified by the clusters pointed out by the K-mean procedure.

The same results are obtained using ACE final dataset computed with "one tablet a day" approach.

4.2.3 Cox's PH Model with adherence binary variable

In order to assess the role of available covariates with respect to the overall survival time of a patient, we perform a Cox's regression models.
Following notation introduced in Chapter 3, we choose six fixed covariates

$$\boldsymbol{X}_j\ (j = 1, ..., 6) = \{\texttt{age\_in},\ \texttt{gender},\ \texttt{tot\_hosp},\ \texttt{CHARLSON},\ \texttt{IHC},\ \texttt{ADERENTE}\}$$

Specifically:

- `age_in` is the patient's age at the first hospitalization
- `gender` is the gender of the patient
- `tot_hosp` is the total number of hospitalizations
- `CHARLSON` is the Charlson comorbidity index at the first hospitalization
- `IHC` is the binary flag which marks if the patient was in IHC
- `ADERENTE` is the binary variable indicating whether a patient is adherent or non adherent, computed as specified in Section 2.2.5

Table 4.22 summaries the main features of the aforementioned covariates and of `death` and `timeOUT` covariates, which respectively represent the binary variable that indicates whether a patient is dead or not at the end of the study and the follow up time [days] of the patient. In particular, `death` is our output variable.

Variable	Value	
No. patients		8,481
`death`	0 (%)	4,389 (51.8%)
	1 (%)	4,092 (48.2%)
`timeOUT`	mean (sd)	1,309.25 (692.72)
`age_in`	mean (sd)	79.15 (9.82)
`gender`	*Male* (%)	4,259 (50.2%)
	Female (%)	4,222 (49.8%)
`tot_hosp`	mean (sd)	2.15 (1.44)
`CHARLSON`	mean (sd)	2.06 (1.76)
`IHC`	0 (%)	7,375 (87%)
	1 (%)	1,106 (13%)
`ADERENTE`	0 (%)	6,998 (82.5%)
	1 (%)	1,483 (17.5%)

Table 4.22: ACE Inhibitors dataset for Cox's model with fixed covariates.

Table 4.23 reports the summary of the Cox's model. From p-values we note that all the covariates are statistically significant at confidence level $\alpha = 5\%$. These results are also confirmed by stratified log-rank tests for which p-values are reported in Table 4.24.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
`age_in`	0.065080	1.067244	0.002103	30.941	< 2e-16	***
`genderM`	0.192373	1.212123	0.033163	5.801	6.60e-09	***
`tot_hosp`	0.126694	1.135070	0.009347	13.555	< 2e-16	***
`CHARLSON`	0.111369	1.117807	0.008060	13.817	< 2e-16	***
`IHC1`	0.242102	1.273924	0.042943	5.638	1.72e-08	***
`ADERENTE1`	-0.138725	0.870467	0.043619	-3.180	0.00147	**

Table 4.23: Summary of the Cox's model for overall survival time with fixed covariates only.

Variable	Stratification	Value	Pts	p-value
Age	*junior*	`age_in` < 65	2,422	$< 2 \cdot 10^{-16}$
	senior	$65 \leq$ `age_in` ≤ 85	379	
	old senior	`age_in` > 85	5,680	
Gender	*female*	`gender` = F	4,222	$4.71 \cdot 10^{-3}$
	male	`gender` = M	4,259	
Hospitalizations	*only one*	`tot_hosp` = 1	3,627	$< 2 \cdot 10^{-16}$
	few	$2 \leq$ `tot_hosp` ≤ 3	3,625	
	many	`tot_hosp` ≥ 4	1,229	
Charlson index	*low*	`CHARLSON` < 2	3,918	$< 2 \cdot 10^{-16}$
	medium	$2 \leq$ `CHARLSON` ≤ 3	3,122	
	high	`CHARLSON` ≥ 4	1,411	
IHC service	*not in IHC*	`IHC` = 0	7,375	$< 2 \cdot 10^{-16}$
	in IHC	`IHC` = 1	1,106	
Adherence	*non adherent*	`ADERENTE` = 0	6,998	$1.56 \cdot 10^{-10}$
	adherent	`ADERENTE` = 1	1,483	

Table 4.24: P-values of stratified log-rank tests.

The covariate `age_in` is strongly significant and being younger leads to a higher survival probability, as it was reasonable to expect. Figure 4.34 also confirms this results. In fact, it represents survival stratified by age. Categories are selected according to the following criteria: we consider female patients that went in IHC, with two hospitalizations, a Charlson index of 2 and aged 45, 65, 80 and 90 years. We observe that the higher the age, the lower the survival. The p-value of the log-rank test (Table 4.24) for age stratified by *junior* (`age_in` < 65), *senior* (65

$\leq$ age_in ≤ 85), *old senior* (age_in > 85) is $< 2 \cdot 10^{-16}$.

The covariate tot_hosp is strongly significant and a higher number of hospitalizations corresponds to a lower survival probability. This result is also confirmed by Figure 4.35 which shows KM of survival stratified by the total number of hospitalizations. Categories are selected according to the following criteria: we consider adherent female patients aged 79 years, that went in IHC, with a Charlson index of 2 and 1, 5 and 10 hospitalizations, respectively. We observe that the higher the number, the lower the survival. This is due to the fact that a higher number of hospitalizations, especially in elderly patients, is probably an index of deterioration in patient's state of health. The p-value of the log-rank test (Table 4.24) for number of hospitalizations stratified by *only one* (tot_hosp $= 1$), *few* ($2 \leq$ tot_hosp ≤ 3), *many* (tot_hosp ≥ 4) is $< 2 \cdot 10^{-16}$.

Similarly the covariate CHARLSON is strongly significant and a higher Charlson comorbidity index at the first hospitalization corresponds to a lower survival probability. Figure 4.36 also confirms this results. It represents survival stratified by the Charlson index at the first hospitalization. We consider adherent female patients aged 79 years, that went in IHC, with two hospitalizations and Charlson indices of 0, 4, 8 and 12, respectively. Also in this case we observe that the higher the number, the lower the survival. This is somehow expected, since a higher comorbidity often indicates more critical clinical prognosis. The p-value of the log-rank test (Table 4.24) for Charlson comorbidity indices stratified by *low* (CHARLSON $<$ 2), *medium* ($2 \leq$ CHARLSON ≤ 3), *high* (CHARLSON ≥ 4) is $< 2 \cdot 10^{-16}$.

The covariate IHC is strongly significant and being gone in IHC service corresponds to a lower survival probability. This result is also confirmed by Figure 4.37 which shows KM of survival stratified by IHC binary values. Categories are selected according to the following criteria: we consider adherent female patients aged 79 years, with two hospitalizations and a Charlson index of 2. This result is related to the fact that being gone in IHC at least once, especially in elderly patients, is probably an index of deterioration in patient's state of health. The p-value of the log-rank test (Table 4.24) stratified by *not in IHC* (IHC $= 0$) and *in IHC* (IHC $=$ 1) is $< 2 \cdot 10^{-16}$.

The covariate gender is strongly significant and being a male corresponds to a lower survival probability. This result is also confirmed by Figure 4.38 which shows KM of survival stratified by gender. Categories are selected according to the following criteria: we consider adherent patients aged 79 years, that went in IHC, with two hospitalizations and a Charlson index of 2. In particular, in Table

4.25, we observe that the 51.6% of dead patients are females, whereas the 48.4% are males. Moreover, the p-value of the log-rank test (Table 4.24) stratified by *female* and *male* is $4.71 \cdot 10^{-3}$.

The covariate `ADERENTE` is significant and the Hazard Ratio (HR) for adherent patients, given by the exponentiated coefficient in Table 4.23, is $0.87 < 1$. This means that, as we expected, being adherent increases the survival probability. Figure 4.39 shows this results through the KM of survival stratified by adherent and non adherent patients. Categories are selected according to the following criteria: we consider female patients aged 79 years that went in IHC, with two hospitalizations and a Charlson index of 2. The p-value of the log-rank test (Table 4.24) stratified by *adherent* and *non adherent* patients is $1.56 \cdot 10^{-10}$ providing strong evidence for this covariate to be significant.

	Pts	*Female* (%)	*Male* (%)
Dead	4,092	2,113 (51.6%)	1,979 (48.4%)
Censored	4,389	2,109 (48.4%)	2,280 (51.9%)

Table 4.25: Total number of dead and truncated patients of ACE mixed approach FVGR final dataset stratified by gender.

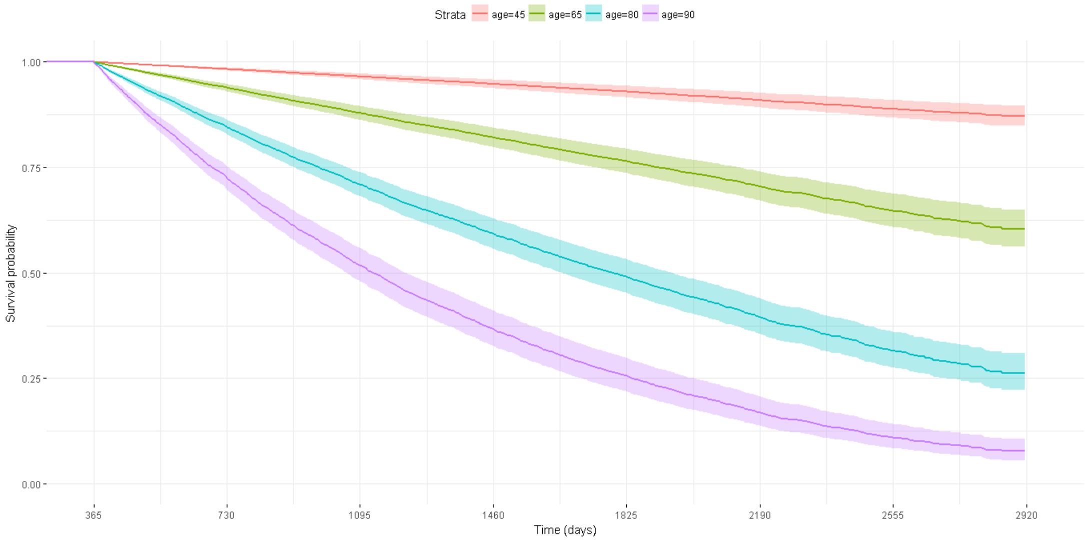


Figure 4.34: Survival probability plot for adherent female patients with two comorbidities, a Charlson index of 2 and that went in IHC and are aged 45, 65, 80 and 90 years.

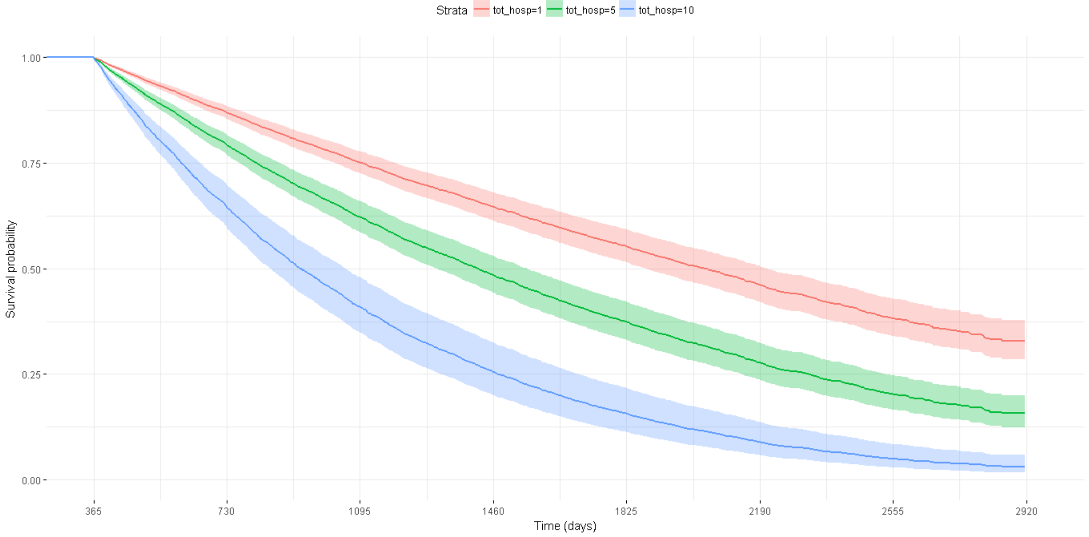


Figure 4.35: Survival probability plot for adherent female patients aged 79 years that went in IHC and that present a Charlson index of 2 and 1, 5 and 10 hospitalizations.

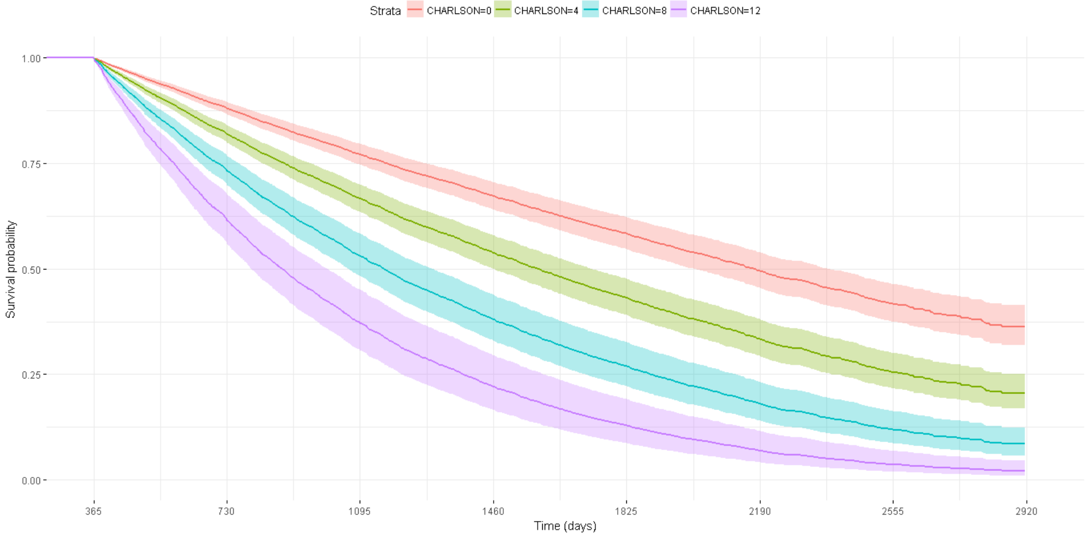


Figure 4.36: Survival probability plot for adherent female patients aged 79 years, with two hospitalizations, that went in IHC and that present a Charlson index of 0, 4, 8 and 12.

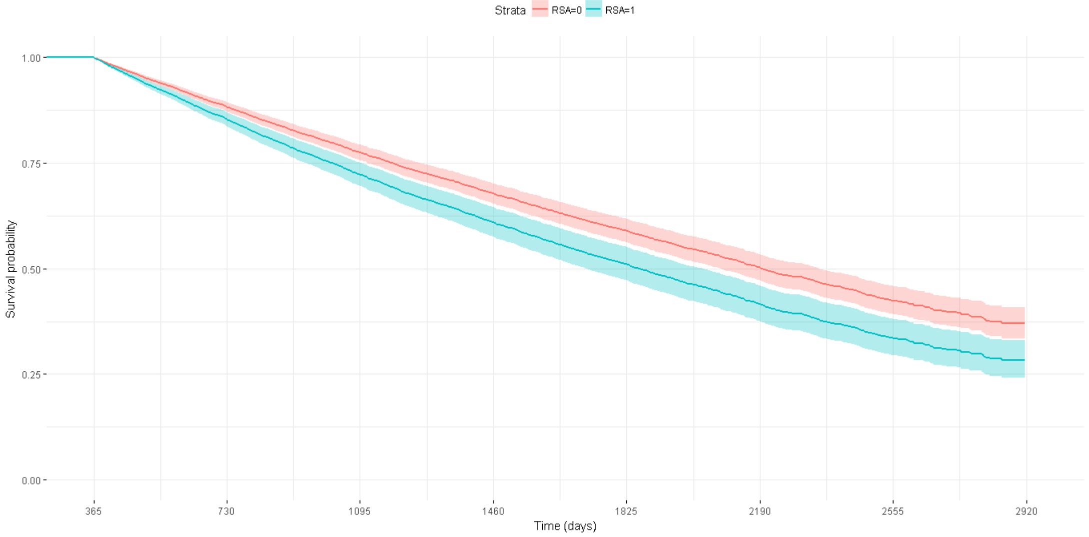


Figure 4.37: Survival probability plot for adherent female patients with 79 years old, two hospitalizations and a Charlson index of 2 stratified by patients that went in IHC service or not.

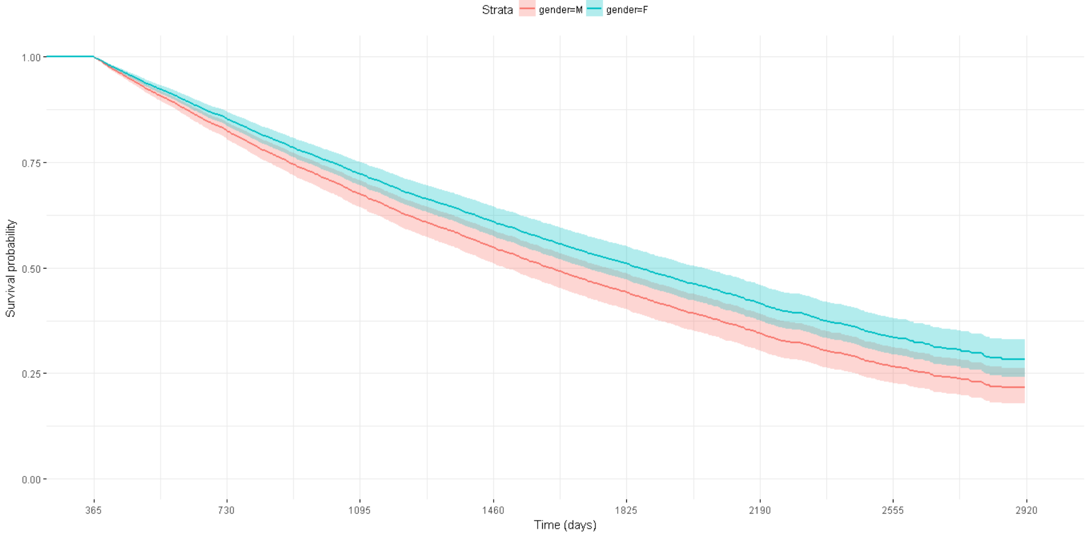


Figure 4.38: Survival probability plot for female vs male adherent patients with 79 years old, two hospitalizations, a Charlson index of 2 and that went in IHC.

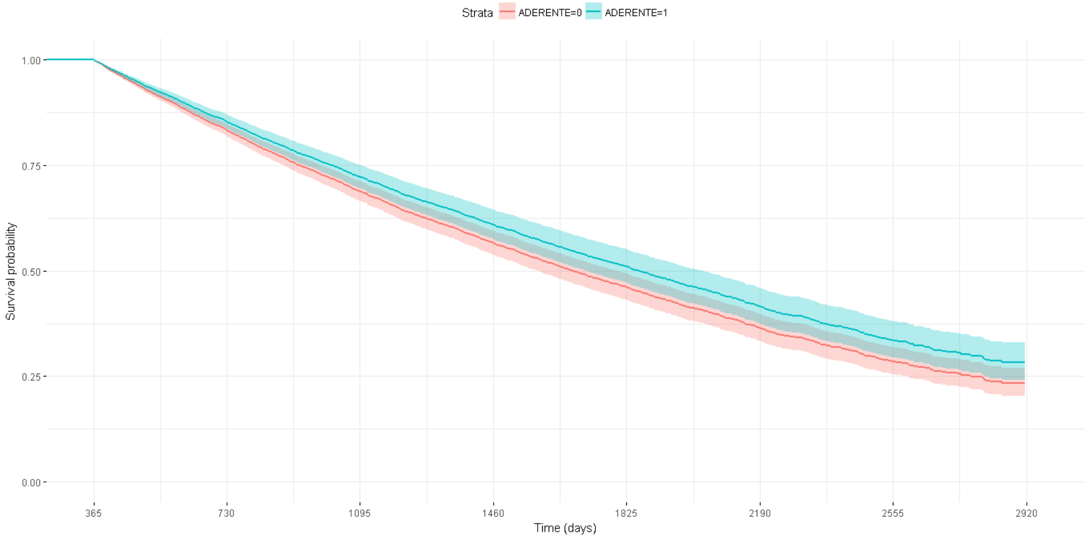


Figure 4.39: Survival probability plot for adherent vs non adherent female patients with 79 years old, two hospitalizations, a Charlson index of 2 and that went in IHC.

Cox PH model with data computed using "one tablet a day" approach

Using ACE Inhibitors final datasets computed through "one tablet a day" approach, we reached to the same conclusions, except for covariate `ADERENTE`.
In fact, in Table 4.26, that reports the summary of the Cox's model, we observe that all the covariates are statistically significant at confidence level $\alpha = 5\%$, except for `ADERENTE`.
However, the p-value of the log-rank test stratified by *adherent* and *non adherent* patients is $3.3 \cdot 10^{-5}$, providing strong evidence to be significant as marginal covariate.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
`age_in`	0.065274	1.067452	0.002103	31.034	< 2e-16	***
`genderM`	0.190703	1.210100	0.033166	5.750	8.93e-09	***
`tot_hosp`	0.127793	1.136317	0.009320	13.712	< 2e-16	***
`CHARLSON`	0.111452	1.117900	0.008064	13.821	< 2e-16	***
`IHC1`	0.248890	1.282601	0.042891	5.803	6.52e-09	***
`ADERENTE1`	-0.091710	0.912370	0.059072	-1.553	0.121	

Table 4.26: Summary of the Cox's model for overall survival time with fixed covariates only for data computed through "one tablet a day".

4.2.4 Joint Modelling of re-hospitalization and drug consumption for HF patients

Since we aim at making up time-dependent covariates to be inserted and treated in a survival model with innovative statistical approach, we proceed with the joint models technique introduced in Section 3.2 and in [26].

For the survival part of the model, we consider the same covariates used in Section 4.2.3 but, instead of dichotomized adherence, we study the effect of curves of cumulative days covered by drug assumption as a secondary joint process. Therefore, in our analysis the longitudinal process in given by these curves, whereas the event process is thought as dependent on age, gender, number of hospitalizations, Charlson comorbidity index and dichotomized IHC. In particular, following notation introduced in Section 3.2.1, we set:

- $m_i(t)$ = cubic root of the value of cumulative days curve [2]
- $\boldsymbol{X}_i = (\texttt{age\_in},\ \texttt{gender},\ \texttt{tot\_hosp},\ \texttt{CHARLSON},\ \texttt{IHC})_i$
- $\lambda_0(\cdot)$ = piecewise-constant baseline risk function given by (3.24)

We observe that data transformation of the value of cumulative days curve is done using a cubic root function, whereas in LR case (see Section 4.1.4) a square root is used. This is due to the different distributions of the curves of cumulative days covered by drug assumption in the two datasets.
We also remind that we use the Gauss-Hermite integration rule to approximate integral (3.20).

The summaries of both longitudinal and event processes are shown in Tables 4.27 and 4.28, respectively. We note that all the covariates are significant at confidence level $\alpha = 5\%$. The parameter labeled as `Assoct` corresponds to parameter α in Equation (3.16), which measures the effect of $m_i(t)$ on the risk of death.

We proceed by checking the fit of the model using residuals plots (see Section 3.2.4). From the residuals for the longitudinal process in Figure 4.40, it evinces that the hypotheses of normally distributed random effects b_i and measurement error terms $\varepsilon_i(t)$ are not fully satisfied. On the contrary, Figure 4.41, about Cox-Snell residuals for the event process, shows that an appropriate functional form for covariates is used in the model.

[2] For computational reasons, as explained in [26], it is necessary to perform an ad hoc data transformation, here represented by the cubic root.

Despite the non optimality of goodness of fit results, we decide to go on further for getting insights of the predictions provided by the JM tool.

	Value	Std.Err	z-value	p-value	
(Intercept)	1.6690	0.0012	1391.357	<0.0001	***
obstime	0.0128	0.0000	2189.913	<0.0001	***

Table 4.27: Summary of the JM longitudinal process for ACE Inhibitors.

	Value	Std.Err	z-value	p-value	
age_in	0.0610	0.0021	29.4394	<0.0001	***
genderM	0.1668	0.0331	5.0349	<0.0001	***
tot_hosp	0.1237	0.0094	13.1941	<0.0001	***
CHARLSON	0.1103	0.0081	13.6956	<0.0001	***
IHC1	0.2460	0.0430	5.7256	<0.0001	***
Assoct	0.0065	0.0024	2.7559	0.0059	**

Table 4.28: Summary of the JM event process for ACE Inhibitors.

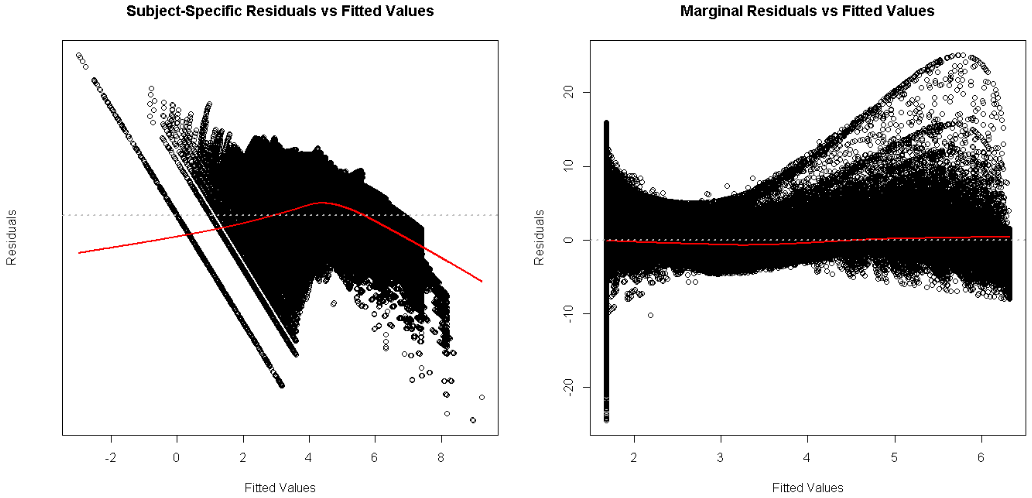


Figure 4.40: Diagnostic plots for the fitted joint model. The left panel depicts the subject-specific residuals for the longitudinal process versus their corresponding fitted values. The right panel depicts the standardized marginal residuals for the longitudinal process versus their corresponding fitted values.

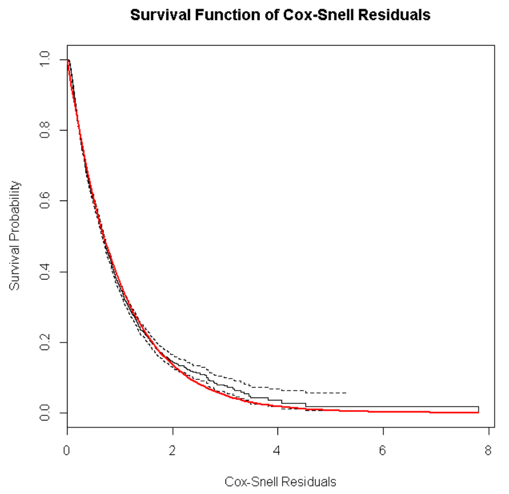


Figure 4.41: Kaplan-Meier estimate of the Cox-Snell residuals for the event process. The dashed lines denote the 95% confidence intervals.

In particular, we focused on the calculation of expected survival probabilities. We compute $\pi_i(u|t)$ for patients in the dataset who have not died by the censored time, using $L = 200$ Monte Carlo samples. For each significant covariate we consider a set of two/three/four real patients, presenting different values for the covariate of interest and similar/equal values for the remaining ones. Patients selection is reported in Tables 4.29, 4.30, 4.31, 4.32, 4.33 and 4.34.

Figure 4.42 shows results of predictions related to the four patients in Table 4.29, i.e., when age varies given all the other covariates as "fixed" in the sense explained above. Moving from left to right panels, survival for elder people is shown. We observe that being younger corresponds to an higher survival probability, as we could have expected.

KEY_ANAGRAFE	timeOUT	gender	age_in	IHC	CHARLSON	tot_hosp	PDC	day_365
1552430	1475	M	57	0	2	1	1	365
649999	2242	M	68	0	2	1	0.997	364
1114835	1387	M	75	0	2	1	0.989	361
914446	2585	M	84	0	2	1	0.997	364

Table 4.29: Patients' data used for JM survival plots on different age values.

Figure 4.43 shows results of predictions related to the two patients in Table 4.30, i.e., when gender varies given all the other covariates as "fixed" in the sense explained above. We observe that the confidence intervals of female (right panel) and male (left panel) patients present overlapping confidence intervals. However, from the Hazard Ration (HR) of the covariate `gender`, given by the exponential `Value` related to `gender` in Table 4.28, that is $\exp(0.1668) = 1.1815$, we have that being a male corresponds to a lower survival probability.

KEY_ANAGRAFE	timeOUT	gender	age_in	IHC	CHARLSON	tot_hosp	PDC	day_365
438197	1755	F	84	0	1	1	0.962	351
2640000	2487	M	82	0	1	1	0.978	357

Table 4.30: Patients' data used for JM survival plots on different gender.

Figure 4.44 shows results of predictions related to the four patients in Table 4.31, i.e., when the total number of hospitalizations varies given all the other covariates as "fixed" in the sense explained above. Moving from left to right panels, survival for increasing number of hospitalizations is shown. We observe that an higher number of hospitalizations corresponds to a lower survival probability.

KEY_ANAGRAFE	timeOUT	gender	age_in	IHC	CHARLSON	tot_hosp	PDC	day_365
1365066	1093	M	73	0	2	1	1	365
1407063	675	M	74	0	2	2	0.978	357
1514470	544	M	73	0	2	3	0.984	359
3008032	1219	M	72	0	2	4	0.962	351

Table 4.31: Patients' data used for JM survival plots on different hospitalization values.

Figure 4.45 shows results of predictions related to the two patients in Table 4.32, i.e., when the total dichotomized IHC varies given all the other covariates as "fixed" in the sense explained above. Patient of left panel did not go in IHC, whereas patient in right one went in IHC. We observe that being gone in IHC corresponds to a lower survival probability.

KEY_ANAGRAFE	timeOUT	gender	age_in	IHC	CHARLSON	tot_hosp	PDC	day_365
3080473	1509	F	80	0	1	2	0.888	324
1811218	1503	F	85	1	1	2	0.890	325

Table 4.32: Patients' data used for JM survival plots on different IHC values.

Figure 4.46 shows results of predictions related to the three patients in Table 4.33, i.e., when the Charlson comorbidity index at the first hospitalization varies given all the other covariates as "fixed" in the sense explained above. Moving from left to right panels, survival for increasing Charlson index is shown. We observe that an higher Charlson comorbidity index corresponds to a lower survival probability.

KEY_ANAGRAFE	timeOUT	gender	age_in	IHC	CHARLSON	tot_hosp	PDC	day_365
1665903	2247	M	72	0	0	2	0.995	363
1414462	2235	M	73	0	1	2	0.997	364
549427	2286	M	72	0	4	2	0.973	355

Table 4.33: Patients' data used for JM survival plots on different Charlson comorbidity indices.

Figure 4.47 shows results of predictions related to the four patients in Table 4.34, i.e., considering different curves of cumulative days covered by drug assumption given all the other covariates as "fixed" in the sense explained above. We observe that having a curve of days covered with an higher final value, and so an higher PDC, correspond to an higher survival probability, as we could have expected. In fact, moving from left to right panels, survival for increasing PDCs is shown. Furthermore, having a lower PDC leads to larger confidence intervals over time, as reported in Tables 4.35, 4.36 and 4.37, so the uncertainty about the prediction of the survival outcome increases.

KEY_ANAGRAFE	timeOUT	gender	age_in	IHC	CHARLSON	tot_hosp	PDC	day_365
1647184	2270	M	72	0	1	1	0.077	28
1246096	2081	M	72	0	1	1	0.255	93
2968114	1422	M	72	0	1	1	0.726	265
672100	837	M	72	0	1	1	0.863	315

Table 4.34: Patients' data used for JM survival plots on different curves of cumulative days covered by drug assumption.

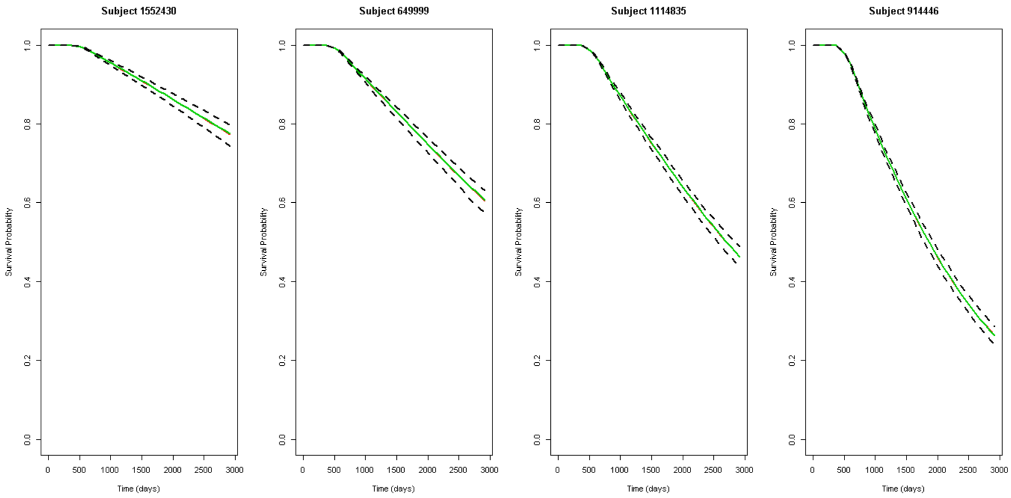


Figure 4.42: Survival probability plots for male patients that did not go in IHC, with one hospitalizations, a Charlson index of 2 and PDC greater than 0.989. From the left panel patients are 57, 68, 75 and 84 years old.

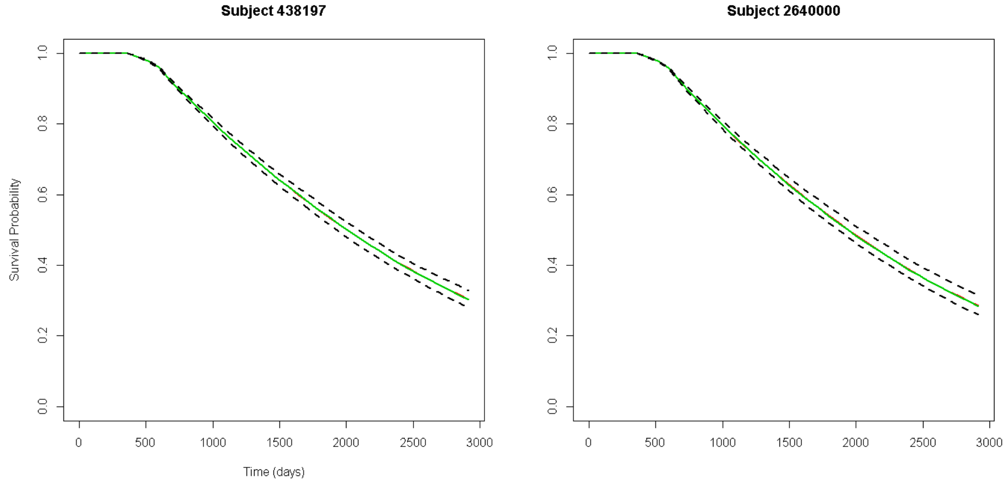


Figure 4.43: Survival probability plots for female (left panel) vs male (right panel) patients aged 84/82 years, that did not go in IHC, with a Charlson index of 1 and PDC greater than 0.96.

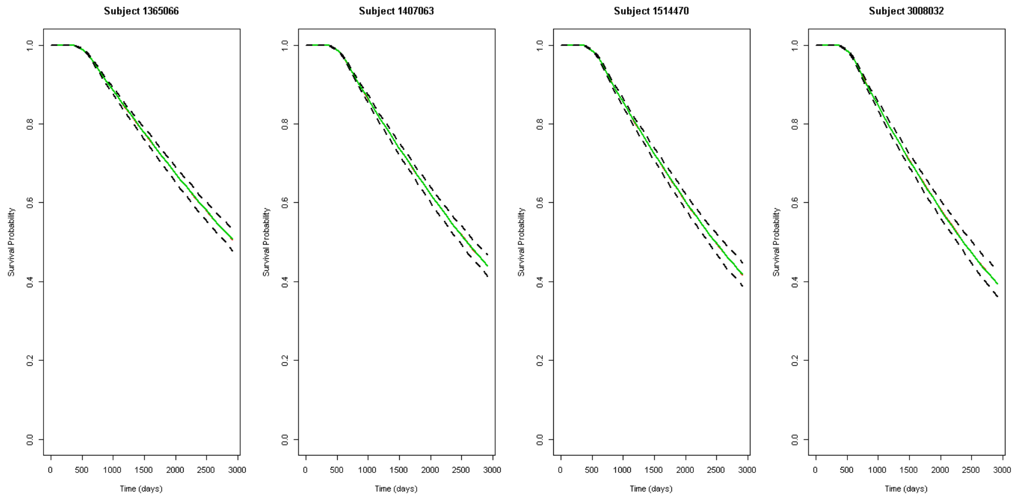


Figure 4.44: Survival probability plots for male patients that did not go in IHC, with 72/73/74 years old, a Charlson index of 2 and PDC greater than 0.96. From the left panel patients have 1, 2, 3 and 4 hospitalizations.

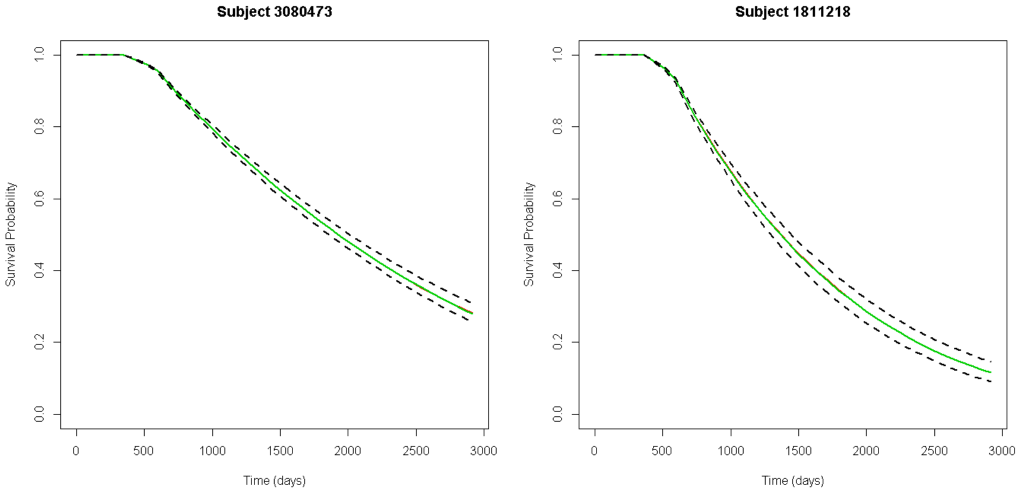


Figure 4.45: Survival probability plots for female patients aged 80/85 years, with two hospitalizations, a Charlson index of 1 and PDC greater than 0.88. Patient in left panel did not go in IHC, whereas patient in right panel went in IHC.

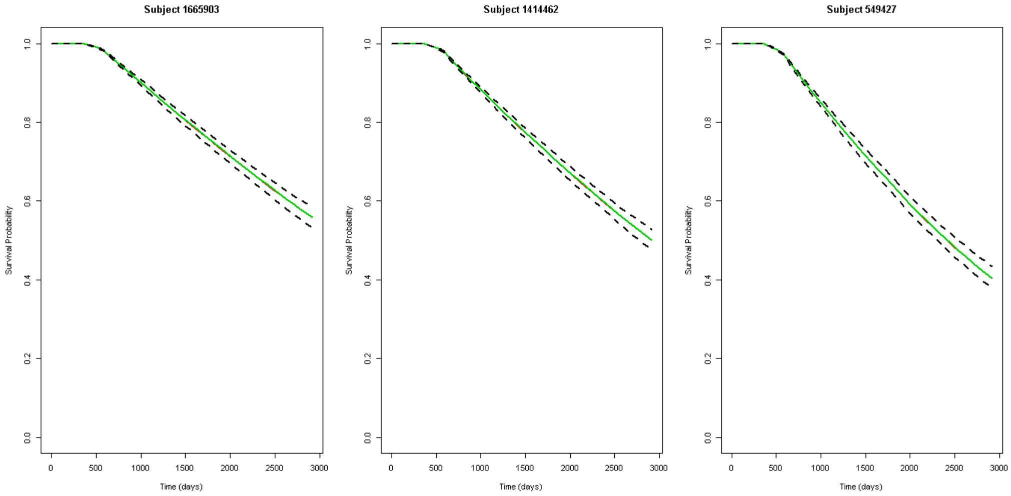


Figure 4.46: Survival probability plots for male patients that did not go in IHC, with 72/73 years old, two hospitalizations and PDC greater than 0.97. From the left panel patients have 0, 1 and 4 Charlson comorbidity indices.

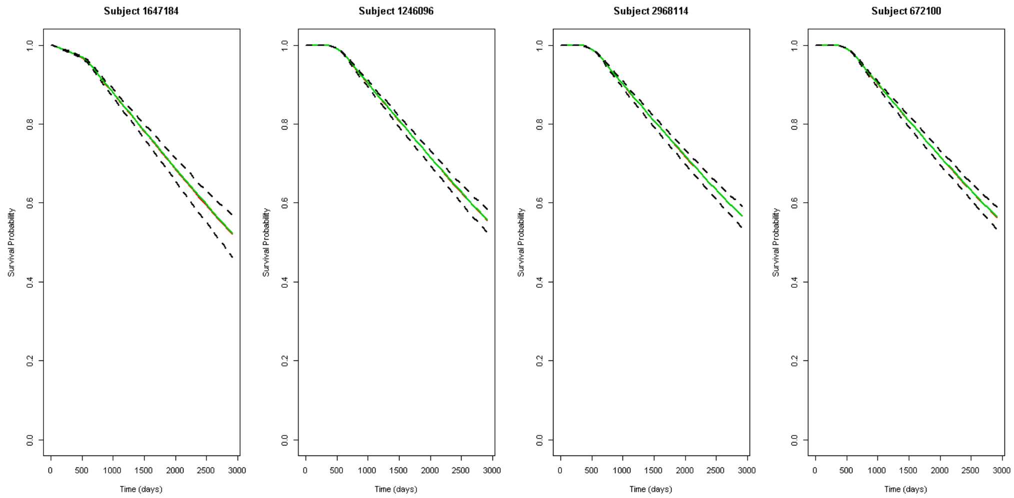


Figure 4.47: Survival probability plots for male patients aged 72 years, that did not go in IHC, one hospitalization and a Charlson index of 1. From the left panel patients have a PDC of 0.077, 0.255, 0.726 and 0.863.

KEY_ANAGRAFE	PDC	Time	Mean	Median	Lower	Upper
1647184	0.077	1 year	0.9778	0.9779	0.9755	0.9799
1246096	0.255	1 year	1.0000	1.0000	1.0000	1.0000
2968114	0.726	1 year	0.9983	0.9983	0.9982	0.9985
672100	0.863	1 year	1.0000	1.0000	1.0000	1.0000

Table 4.35: Mean and median values of patients' survival probabilities one year after the end of the follow up with the respective PDC and 95% confidence intervals.

KEY_ANAGRAFE	PDC	Time	Mean	Median	Lower	Upper
1647184	0.077	3 years	0.8556	0.8567	0.8424	0.8670
1246096	0.255	3 years	0.8803	0.8806	0.8690	0.8876
2968114	0.726	3 years	0.8785	0.8788	0.8672	0.8858
672100	0.863	3 years	0.8795	0.8799	0.8684	0.8867

Table 4.36: Mean and median values of patients' survival probabilities three years after the end of the follow up with the respective PDC and 95% confidence intervals.

KEY_ANAGRAFE	PDC	Time	Mean	Median	Lower	Upper
1647184	0.077	5 years	0.7115	0.7128	0.6849	0.7351
1246096	0.255	5 years	0.7403	0.7405	0.7205	0.7550
2968114	0.726	5 years	0.7418	0.7424	0.7232	0.7561
672100	0.863	5 years	0.7414	0.7418	0.7224	0.7561

Table 4.37: Mean and median values of patients' survival probabilities five years after the end of the follow up with the respective PDC and 95% confidence intervals.

Joint Modelling with data computed using "one tablet a day" approach

Using ACE Inhibitors final datasets computed through "one tablet a day" approach, we reached to the same conclusions. In fact, all the covariates are significant at confidence level $\alpha = 5\%$. In particular, the p-value of parameter `Assoct` (corresponding to α in Equation (3.16)) is equal to 0.0028, providing strong evidence for this covariate to be significant.
Therefore, we observe that in "one-tablet-a-day" setting the dichotomized variable for adherence was not statistically significant (as mentioned in the previous Section), while the time-dependent variable is.

Given these analyses, we arrive at the same conclusion of LR case: modelling the drug assumption process as time-varying covariates in a joint model setting is a richer inferential instrument than a simple Cox's model. In fact, it is an interpretative and forecasting tool for exploring the effects of pharmacological treatments on survival. For example, it allows us to confirm some pharmacoepidemiological intuition as the fact that medication nonadherence is commonly associated with adverse health conditions [18] in a more suitable way. Moreover, such a model enables a tailored prediction for different patients profiles.

Discussion and Conclusions

This thesis is a methodological work for analysing the effects of adherence to drug prescription on survival in HF patients of Lombardy and Friuli Venezia Giulia Regions. The first part of this work concerned the individuation of a time-dependent covariate for adherence to treatment which would result more realistic that a simple binary variable. We identified this feature in the curve of cumulative days covered by drug assumption. In the second part, we applied the Joint Modelling technique in order to investigate how patients' time-to-event outcome are influenced by longitudinal data, given by the pharmacological treatment curves, and we compared the obtained results with Cox's model outputs fitted using the binary variable for adherence.
Both in LR and in FVGR, we discovered that modelling the drug assumption process as a time-dependent covariate in a model setting is a richer instrument that a simple Cox's model. However, some improvements may be included into the longitudinal part of JM in order to provide a more proper modelling of the functional covariate. In particular, linear mixed effects part of JM given by Equation (3.17), which models the longitudinal process, could be modified using a more general and realistic tool, such as a counting process or a generalized linear model. This would surely imply many issues concerning the likelihood derivation and the numerical optimization. Nevertheless, once fixed the problem, it could be a powerful instrument of analysis.

Our methodological work opens the way for many further developments, both in the fields of statistical methods and pharmacoepidemiology.
In the first area, there are different statistical approaches to deal with time-dependent covariates. A first possibility is to fit a Cox's model in which a Gini correlation coefficient representing the angle of the curve is inserted.
Moreover, a second option consists in the application of functional data analysis (FDA) techniques, which could better exploit the great potential of our curves.
Finally, we have considered pharmacological treatments only as an internal time-varying covariates but they also might be perceived as external, if their values are prescribed at the beginning of the study. Therefore, a third development consists

in changing the approach in order to deal with exogenous time-dependent covariates, such as the use of functional principal component analysis (FPCA).

On the other hand, in the pharmacoepidemiological area, it could be interesting to further strengthen the Charlson index covariate, which was strongly significant in our analyses, introducing information related to other comorbidities through the pharmacological data.
Moreover, the use of DDD to analyse drug utilisation could be limiting: DDD is a dose that is rarely prescribed since it is an average of two or more commonly used doses. For this reason, it could be interesting to integrate our data with dosages prescribed by doctors, in order to obtain a more realistic analysis of coverage days.
Finally, in our work we have only considered adherence to monotherapies but patients often follow different pharmacological treatments. Therefore, a lot of work is needed in order to include simultaneously all the treatments in a not trivial way.

Appendix A

LR descriptive tables

ACE Inhibitors

Variable	Value	
No. patients		2,916
`gender`	*Male* (%)	1,681 (57.6%)
	Female (%)	1,235 (42.4%)
`age_in`	mean (sd)	72.17 (11.44)
`labelOUT`	*Dead* (%)	718 (24.6%)
	Censored (%)	2,189 (75.1%)
	Lost (%)	9 (0.3%)
`death`	0 (%)	2,198 (75.4%)
	1 (%)	718 (24.6%)
`timeOUT`	mean (sd)	1,543.45 (613.46)
`tot_hosp`	mean (sd)	2.35 (1.61)
`comorbidity`	mean (sd)	2.04 (1.07)
`tot_procedures`	mean (sd)	0.13 (0.36)
`PDC`	mean (sd)	0.72 (0.28)
`ADERENTE`	0 (%)	1,358 (46.6%)
	1 (%)	1,558 (53.4%)
`PDC_CLA`	1 (%)	286 (9.8%)
	2 (%)	422 (14.4%)
	3 (%)	506 (17.4%)
	4 (%)	1,702(58.4%)
`curvaMG`	0 (%)	152 (5.2%)
	1 (%)	2,764 (94.8%)

Table A.1: Summary of LR ACE Inhibitors final dataset.

Angiotensin Receptor Blockers

Variable	Value	
No. patients		1,473
gender	*Male* (%)	775 (52.6%)
	Female (%)	698 (47.4%)
age_in	mean (sd)	73.01 (10.20)
labelOUT	*Dead* (%)	356 (24.2%)
	Censored (%)	1,115 (75.7%)
	Lost (%)	2 (0.1%)
death	0 (%)	1,117 (75.8%)
	1 (%)	356 (24.2%)
timeOUT	mean (sd)	1,521.05 (609.92)
tot_hosp	mean (sd)	2.32 (1.63)
comorbidity	mean (sd)	2.20 (1.10)
tot_procedures	mean (sd)	0.10 (0.30)
PDC	mean (sd)	0.64 (0.29)
ADERENTE	0 (%)	880 (59.8%)
	1 (%)	593 (40.2%)
PDC_CLA	1 (%)	202 (13.7%)
	2 (%)	286 (19.4%)
	3 (%)	288 (19.6%)
	4 (%)	697 (47.3%)
curvaMG	0 (%)	201 (13.6%)
	1 (%)	1,272 (86.4%)

Table A.2: Summary of LR Angiotensin Receptor Blockers final dataset.

Beta Blocking agents

Variable	Value	
No. patients		2,890
`gender`	*Male* (%)	1,675 (58%)
	Female (%)	1,215 (42%)
`age_in`	mean (sd)	71.04 (11.35)
`labelOUT`	*Dead* (%)	633 (21.9%)
	Censored (%)	2,251 (77.9%)
	Lost (%)	6 (0.2%)
`death`	0 (%)	2,257 (78.1%)
	1 (%)	633 (21.9%)
`timeOUT`	mean (sd)	1,516.60 (615.30)
`tot_hosp`	mean (sd)	2.40 (1.64)
`comorbidity`	mean (sd)	2.01 (1.09)
`tot_procedures`	mean (sd)	0.14 (0.37)
`PDC`	mean (sd)	0.38 (0.23)
`ADERENTE`	0 (%)	2,676 (92.6%)
	1 (%)	87 (5.7%)
`PDC_CLA`	1 (%)	1,051 (36.4%)
	2 (%)	1,044 (36.1%)
	3 (%)	514 (17.8%)
	4 (%)	281 (9.7%)
`curvaMG`	0 (%)	5 (0.2%)
	1 (%)	2,885 (99.8%)

Table A.3: Summary of LR Beta Blocking agents final dataset.

Anti Aldosterone agents

Variable	Value	
No. patients		2,006
gender	*Male* (%)	1,040 (51.8%)
	Female (%)	966 (48.2%)
age_in	mean (sd)	73.61 (10.60)
labelOUT	*Dead* (%)	545 (27.2%)
	Censored (%)	1,459 (72.7%)
	Lost (%)	2 (0.1%)
death	0 (%)	1,461 (72.7%)
	1 (%)	545 (27.2%)
timeOUT	mean (sd)	1,473.91 (623.35)
tot_hosp	mean (sd)	2.45 (1.67)
comorbidity	mean (sd)	2.06 (1.05)
tot_procedures	mean (sd)	0.13 (0.35)
PDC	mean (sd)	0.38 (0.22)
adherent	0 (%)	1,900 (94.7%)
	1 (%)	106 (5.3%)
PDC_CLA	1 (%)	666 (33.2%)
	2 (%)	778 (38.8%)
	3 (%)	424 (21.1%)
	4 (%)	138 (6.9%)
curvaMG	0 (%)	996 (49.7%)
	1 (%)	1,010 (50.3%)

Table A.4: Summary of LR Anti Aldosterone agents final dataset.

Diuretics

Variable	Value	
No. patients		3,399
gender	*Male* (%)	1,821 (53.6%)
	Female (%)	1,578 (46.4%)
age_in	mean (sd)	73.82 (10.60)
labelOUT	*Dead* (%)	955 (28.1%)
	Censored (%)	2,438 (71.7%)
	Lost (%)	6 (0.2%)
death	0 (%)	2,444 (71.9%)
	1 (%)	955 (28.1%)
timeOUT	mean (sd)	1,477.81 (613.68)
tot_hosp	mean (sd)	2.40 (1.65)
comorbidity	mean (sd)	2.16 (1.13)
tot_procedures	mean (sd)	0.11 (0.34)
PDC	mean (sd)	0.61 (0.28)
adherent	0 (%)	2,279 (67%)
	1 (%)	1,120 (33%)
PDC_CLA	1 (%)	473 (13.9%)
	2 (%)	695 (20.4%)
	3 (%)	948 (27.9%)
	4 (%)	1,238 (37.8%)
curvaMG	0 (%)	0 (0%)
	1 (%)	3,399 (100%)

Table A.5: Summary of LR Diuretics final dataset.

Appendix B

Evaluation of LR curves

A first analysis of this work consists in the evaluation of some curves of cumulative days covered by drug assumption, introduced in Section 2.1.5, and of some curves of assumed dose of drug over time, introduced in Section 2.1.6.

Due to the fact that the curves are monotone and non-decreasing, we use as selection criterion the value of the curves at 365-th days and we keep the minimum, the maximum, the median and the mean (or the least integer greater than or equal to their values), both for adherent and non adherent patients. Since the considerations are the same for all the pharmacological classes, for brevity we report this analysis only for ACE Inhibitors. A summary of the patients' characteristics of ACE final dataset is reported in Appendix A, Table A.1.

B.1 Curves of cumulative days covered by drug assumption

First of all, we split patients into adherent and non adherent ones. Their distributions of final values of curves of cumulative days covered by drug assumption is shown in Figure B.1 and values obtained using selection criterion is reported in Table B.1.

	Minimum	Maximum	Median	Mean
Adherent	292	365	348	342
Not adherent	1	291	176	172

Table B.1: Selected values of the curves of cumulative days covered by drug assumption at 365-th days.

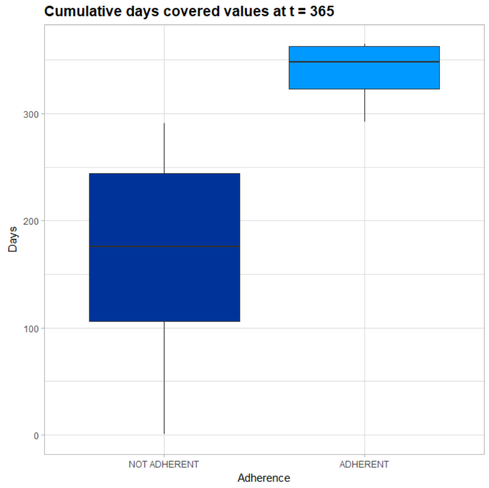


Figure B.1: Boxplots of the values of the curves of cumulative days at $t = 365$ days stratified by non adherent (blue) and adherent (light blue) patients.

Given that the Definition (1.5.1) of adherence comes from patient's total days of coverage, it is not surprising that the values of adherent patients are higher than those of non adherent patients, as confirmed by Wilcoxon test (p-value is $< 2 \cdot 10^{-16}$). Starting from these values, we select eight patients whose characteristics are reported in Tables B.2 and B.3.

The curves of selected patients are reported in Figure B.2 and we can observe that:

1. The curves of adherent patients (left panel) have higher final slopes because being adherent corresponds to have longer periods of coverage.
2. The curves of non adherent patients (right panel) have longer and more numerous plateaux that correspond to periods in which they do not assume the drugs.
3. All the growth zones of the curves have the same slope, that is 1, because when they increase they do it one day at a time.
4. The curve of the female non adherent patient with the minimum value (`COD_REG` = 15239614) has a coverage period of only one day (the last). In Table B.4 we report her pharmacological history and we can observe that she has only one ACE event and that, even if `qt_pharma` = 42, the observation period goes from `data_rif_ev` = "2007-01-16" to "2008-01-15". Therefore, that prescription only covers the last day of the 365 days considered.

Adherent	Minimum	Maximum	Median	Mean
COD_REG	10246668	10006065	10403409	13291616
lablelOUT	DEAD	CENSORED	CENSORED	CENSORED
timeOUT	1,388	2,316	9,68	1,616
age_in	70	68	85	81
gender	M	M	M	F
tot_hosp	6	2	2	1
comorbidity	1	2	4	2
tot_procedures	0	0	0	0
PDC	0.8	1	0.9534	0.937

Table B.2: Characteristics of adherent selected patients for the curves of cumulative days covered by drug assumption.

Non-adherent	Minimum	Maximum	Median	Mean
COD_REG	15239614	10147557	10745825	16733768
lablelOUT	CENSORED	DEAD	CENSORED	CENSORED
timeOUT	2,176	2,185	1,709	2,289
age_in	66	42	82	48
gender	F	F	F	F
tot_hosp	1	3	3	1
comorbidity	1	2	3	2
tot_procedures	1	0	0	0
PDC	0.003	0.797	0.482	0.471

Table B.3: Characteristics of non adherent selected patients for the curves of cumulative days covered by drug assumption.

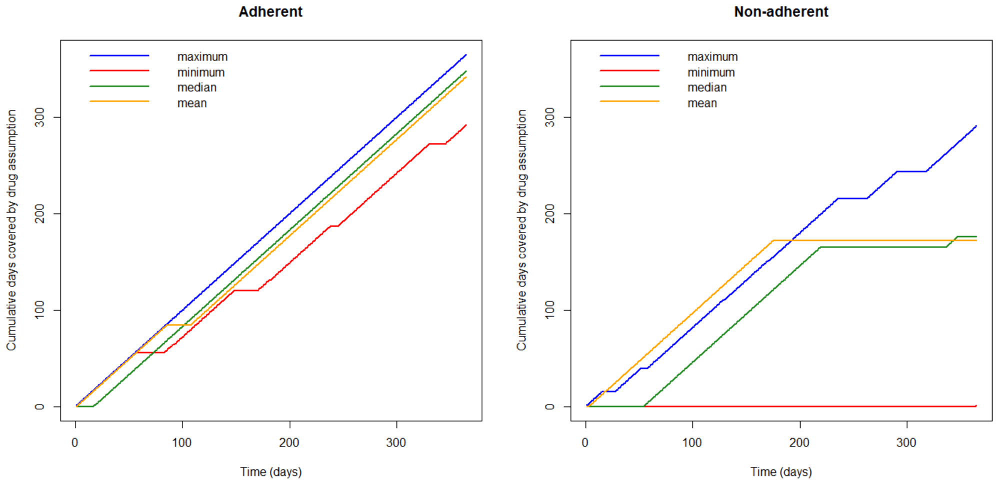


Figure B.2: Curves of cumulative days covered by drug assumption of selected patients, divided into adherent (left panel) and non adherent (right panel).

	COD_REG	data_rif_ev	data_studio_out	labelOUT	data_prest	hosp	pharm	dataADM	LOS	classe_pharma	ATC	qt_pharma	DDD	COMBO
1	15239614	2007-01-16	2012-12-31	CENSORED	2007-01-16	1		2007-01-01	15					
2	15239614	2007-01-16	2012-12-31	CENSORED	2007-01-16		1			DIU	C03CA01	38	40.00	0
3	15239614	2007-01-16	2012-12-31	CENSORED	2007-01-16		2			AA	C03DA01	16	75.00	0
4	15239614	2007-01-16	2012-12-31	CENSORED	2007-01-16		3			BB	C07AG02	5	37.50	0
5	15239614	2007-01-16	2012-12-31	CENSORED	2007-01-16		4			ARB	C09CA03	28	80.00	0
6	15239614	2007-01-16	2012-12-31	CENSORED	2007-01-29		5			AA	C03DA01	27	75.00	0
7	15239614	2007-01-16	2012-12-31	CENSORED	2007-01-29		6			BB	C07AG02	14	37.50	0
8	15239614	2007-01-16	2012-12-31	CENSORED	2007-01-29		7			ARB	C09CA03	84	80.00	0
9	15239614	2007-01-16	2012-12-31	CENSORED	2007-03-30		8			AA	C03DA01	27	75.00	0
10	15239614	2007-01-16	2012-12-31	CENSORED	2007-03-30		9			BB	C07AG02	14	37.50	0
11	15239614	2007-01-16	2012-12-31	CENSORED	2007-03-30		10			ARB	C09CA03	84	80.00	0
12	15239614	2007-01-16	2012-12-31	CENSORED	2007-07-20		11			BB	C07AG02	9	37.50	0
13	15239614	2007-01-16	2012-12-31	CENSORED	2007-07-20		12			ARB	C09CA03	56	80.00	0
14	15239614	2007-01-16	2012-12-31	CENSORED	2007-09-11		13			AA	C03DA01	16	75.00	0
15	15239614	2007-01-16	2012-12-31	CENSORED	2007-09-11		14			ARB	C09CA03	56	80.00	0
16	15239614	2007-01-16	2012-12-31	CENSORED	2007-09-12		15			AA	C03DA01	5	75.00	0
17	15239614	2007-01-16	2012-12-31	CENSORED	2007-09-12		16			BB	C07AG02	9	37.50	0
18	15239614	2007-01-16	2012-12-31	CENSORED	2007-10-10		17			DIU	C03CA01	38	40.00	0
19	15239614	2007-01-16	2012-12-31	CENSORED	2007-10-10		18			BB	C07AG02	9	37.50	0
20	15239614	2007-01-16	2012-12-31	CENSORED	2007-10-10		19			ARB	C09CA03	56	80.00	0
21	15239614	2007-01-16	2012-12-31	CENSORED	2007-11-07		20			AA	C03DA01	27	75.00	0
22	15239614	2007-01-16	2012-12-31	CENSORED	2007-12-20		21			DIU	C03CA01	38	40.00	0
23	15239614	2007-01-16	2012-12-31	CENSORED	2007-12-20		22			BB	C07AG02	9	37.50	0
24	15239614	2007-01-16	2012-12-31	CENSORED	2007-12-22		23			ARB	C09CA03	56	80.00	0
25	15239614	2007-01-16	2012-12-31	CENSORED	2008-01-15		24			ACE	C09BA02	42		1

Table B.4: Events data of patient 15239614.

B.2 Curves of assumed dose

As done in the previous Section, we split patients into adherent and non adherent ones. Their distributions of final values of curves of assumed dose over time is shown in Figure B.3 and values obtained using selection criterion is reported in Table B.1. Starting from these values, we select eight patients whose characteristics are reported in Tables B.6 and B.7.

	Minimum	Maximum	Median	Mean
Adherent	730	18,250	910	1,792.5
Not adherent	15	14,200	695	1,300

Table B.5: Selected values of the curves of assumed dose at 365-th days.

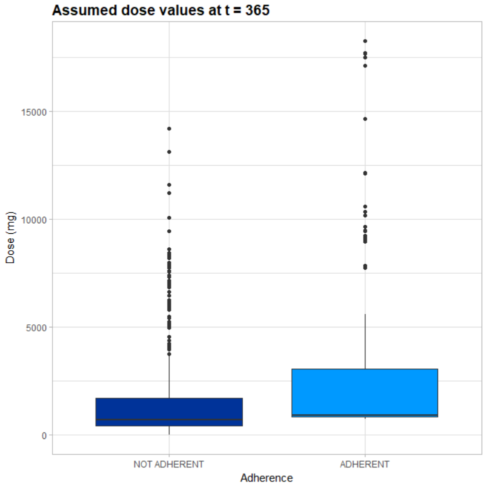


Figure B.3: Boxplots of the values of the curves of assumed dose at 365-th days divided into non adherent (blue) and adherent (light blue) patients.

Adherent	**Minimum**	**Maximum**	**Median**	**Mean**
COD_REG	12616938	17642869	10046187	16808145
lablelOUT	CENSORED	CENSORED	DEAD	DEAD
timeOUT	1,897	1,270	1,703	2,022
age_in	55	82	68	71
gender	M	M	M	M
tot_hosp	2	1	2	5
comorbidity	1	3	4	2
tot_procedures	0	1	0	1
PDC	0.8	1	0.997	0.838

Table B.6: Characteristics of adherent selected patients for the curves of assumed dose.

Non-dherent	**Minimum**	**Maximum**	**Median**	**Mean**
COD_REG	18482478	15677519	14444128	14639520
lablelOUT	CENSORED	DEAD	CENSORED	CENSORED
timeOUT	1,525	2,265	2,472	1,039
age_in	77	69	77	68
gender	M	M	M	M
tot_hosp	1	5	2	1
comorbidity	2	2	3	1
tot_procedures	1	0	0	0
PDC	0.016	0.778	0.762	0.553

Table B.7: Characteristics of non adherent selected patients for the curves of assumed dose.

The curves of selected patients are reported in Figures B.4 and B.5, where lefts panels show the plots for adherent patients and right panels those for non adherent patients. We can observe that:

1. There is no clear difference between the values of the two groups. This depends on the fact that, even if the adherent patients have a higher value of coverage days, the doses assumed range between $2mg$ and $150mg$, as shown in Table B.8. Therefore, measured values strongly depend on DDDs.

2. The curves of non adherent patients have longer and more numerous plateaux, corresponding to the periods in which they do not assume the drugs.

3. The slopes of the growth parts are different. They vary both between several curves and within the same curve depending on the assumed dose. Indeed each time the slope is given by:

$$\text{slope} = \frac{\text{daily dose in } mg}{1 \text{ day}}$$

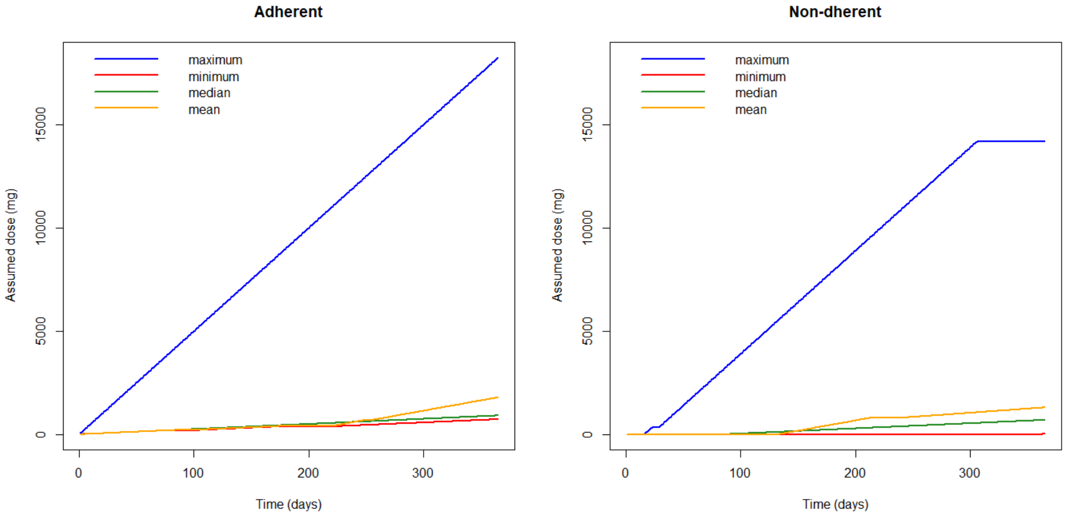


Figure B.4: Curves of assumed dose over time of of selected patients, divided into adherent (left panel) and non adherent (right panel).

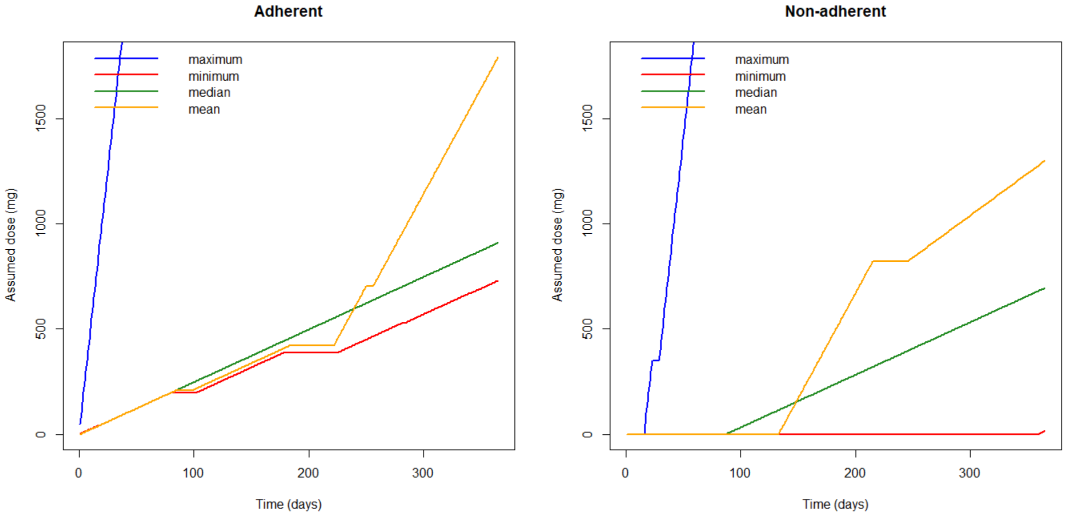


Figure B.5: Zoom of curves of assumed dose over time of selected patients, divided into adherent (left panel) and non adherent (right panel).

ATC	classe_pharma	COMBO	DDD (mg)
C09AA01	ACE	0	50.00
C09AA02	ACE	0	10.00
C09AA03	ACE	0	10.00
C09AA04	ACE	0	4.00
C09AA05	ACE	0	2.50
C09AA06	ACE	0	15.00
C09AA07	ACE	0	7.50
C09AA08	ACE	0	2.50
C09AA09	ACE	0	15.00
C09AA10	ACE	0	2.00
C09AA12	ACE	0	30.00
C09AA15	ACE	0	30.00
C09BA01	ACE	1	NA
C09BA02	ACE	1	NA
C09BA03	ACE	1	NA
C09BA04	ACE	1	NA
C09BA05	ACE	1	NA
C09BA06	ACE	1	NA
C09BA07	ACE	1	NA
C09BA08	ACE	1	NA
C09BA09	ACE	1	NA
C09BA12	ACE	1	NA
C09BA13	ACE	1	NA
C09BA15	ACE	1	NA
C09BB02	ACE	1	NA
C09BB04	ACE	1	NA
C09BB05	ACE	1	NA
C09BB07	ACE	1	NA
C09XA02	ACE	0	150.00

Table B.8: List of ACE Inhibitors ATC codes of our dataset and their DDDs.

Appendix C

Patients not in final FVG cohort

During preprocessing and selection of FVG cohort in Section 2.2.2, we select only patients with at least one event after the reference date. In this way we remove 121 patients that present pharmacological events only before the reference date and only the first one hospitalization. They are 81 females and 40 males with a mean age of 85.1 years old, as reported in Table C.1. Moreover, at the end of the study, 55 patients are dead and the other 66 are censored. Their follow up times range from 368 to 2,745 days (about seven and a half years) and a mean of 1,312 days (about three and a half years), as shown in Table C.2. Finally, Table C.3 reports the values of patients' Charlson comorbidity indices.

Age at the first hospitalization					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
56	80	87	85.1	91	101

Table C.1: Summary of patients' age.

Follow up time [days]					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
368	749	1,143	1,312	1,891	2,745

Table C.2: Summary of follow up time [days].

Charlson comorbidity index									
Index	0	1	2	3	4	5	6	7	8
Pts	17	33	36	18	9	1	2	3	2

Table C.3: Table of patients' Charlson comorbidity index.

Appendix D

Charlson comorbidity index

The Charlson comorbidity index [11] is calculated using hospital diagnosis based on ICD-9CM that occurred within five years before the first admission and integrated with laboratory data and diagnosis recorded at the first admission. In particular, for the diagnosis of diabetes mellitus we integrated information about glycosylated haemoglobin at admission and the recorded diagnosis of diabetes mellitus in the previous 5 years. Similarly, to assess the presence of a chronic kidney disease, we integrated the creatinine value at admission to compute the estimated glomerular filtration rate (eGFR) $<$ 60 ml/min (with the CKD-EPI formula) with the reported diagnosis of a chronic kidney disease in the previous 5 years [12].

Comorbidities considered are:

- myiocardial infarction
- congestive heart failure
- peripheric vascular disease
- cerebrovascular disease
- dementia
- chronic pulmonary disease
- rheumatic disease
- peptic ulcer disease
- mild liver disease
- diabetes with or without compliance

- hemiplagia
- renal disease
- tumour
- AIDS/HIV

The Charlson comorbidity index is computed as the sum of: myiocardial infarction, congestive heart failure, peripheric vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, rheumatic disease, peptic ulcer disease, mild liver disease, diabetes without compliance, twice the diabetes with compliance, twice the hemiplagia, twice the renal disease, twice any malignity, three times severe liver disease, six times the solid tumour and six times the sids.

The *R* computation of Charlson index for a patient is the following:

```
charlson_index = mi_glob + cong_glob + periph_glob +
                 cereb_glob + dem_glob + pulm_glob +
                 reum_glob + ulcera_glob + liver_glob +
                 diab_glob + 2*diabcompl_glob +
                 2*hem_glob + 2*renal_glob + 2*any_glob +
                 3*mod_sev_glob + 6*met_solid_glob +
                 6*aids_glob
```

where *disease_glob* indicates the presence of *disease* in anamnesis or at the first hospitalization.

Appendix E

FVGR descriptive tables

ACE Inhibitors

Variable	Value	
No. patients		8,481
`gender`	*Male* (%)	4,259 (50.2%)
	Female (%)	4,222 (49.8%)
`age_in`	mean (sd)	79.15 (9.82)
`labelOUT`	*Dead* (%)	4,092 (48.2%)
	Censored (%)	4,389 (51.8%)
`death`	0 (%)	4,389 (51.8%)
	1 (%)	4,092 (48.2%)
`timeOUT`	mean (sd)	1,309.25 (692.72)
`tot_hosp`	mean (sd)	2.15 (1.44)
`IHC`	0 (%)	7,375 (87%)
	1 (%)	1,106 (13%)
`CHARLSON`	mean (sd)	2.06 (1.76)
`tot_procedures`	mean (sd)	0.41 (0.51)

Table E.1: Summary of FVGR ACE Inhibitors final dataset.

Variable	Value	Mixed approach	One tablet a day
`PDC`	mean (sd)	0.52 (0.27)	0.47 (0.24)
`ADERENTE`	0 (%)	6,998 (82.5%)	7,740 (91.3%)
	1 (%)	1,483 (17.5%)	741 (8,7%)
`PDC_CLA`	1 (%)	1,713 (20.2%)	1,864 (22%)
	2 (%)	2,288 (27%)	2,724 (32.1%)
	3 (%)	2,563 (30.2%)	2,785 (32.8%)
	4 (%)	1,917 (22.6%)	1,108 (13.1%)

Table E.2: Summary of adherence variables of FVGR ACE Inhibitors final dataset dived by approaches.

Angiotensin Receptor Blockers

Variable	Value	
No. patients		4,004
gender	*Male* (%)	1,730 (43.2%)
	Female (%)	2,274 (56.8%)
age_in	mean (sd)	79.36 (9.28)
labelOUT	*Dead* (%)	1,822 (45.5%)
	Censored (%)	2,182 (54.5%)
death	0 (%)	1,822 (45.5%)
	1 (%)	2,182 (54.5%)
timeOUT	mean (sd)	1,320.49 (683.07)
tot_hosp	mean (sd)	2.14 (1.38)
IHC	0 (%)	3,529 (88.1%)
	1 (%)	475 (11.9%)
CHARLSON	mean (sd)	2.12 (1.77)
tot_procedures	mean (sd)	0.41 (0.51)

Table E.3: Summary of FVGR Angiotensin Receptor Blockers final dataset.

Variable	Value	Mixed approach	One tablet a day
PDC	mean (sd)	0.43 (0.27)	0.40 (0.24)
ADERENTE	0 (%)	3,599 (89.9%)	3,826 (95.6%)
	1 (%)	405 (10.1%)	178 (4.4%)
PDC_CLA	1 (%)	1,177 (29.4%)	1,226 (30.6%)
	2 (%)	1,173 (29.3%)	1,288 (32.2%)
	3 (%)	1,087 (27.1%)	1,170 (29.2%)
	4 (%)	567 (14.2%)	320 (8%)

Table E.4: Summary of adherence variables of FVGR Angiotensin Receptor Blockers final dataset dived by approaches.

Beta Blocking agents

Variable	Value	
No. patients		9,341
gender	*Male* (%)	4,510 (48.3%)
	Female (%)	4831 (51.7%)
age_in	mean (sd)	78.59 (9.82)
labelOUT	*Dead* (%)	4,064 (43.5%)
	Censored (%)	5,288 (56.5%)
death	0 (%)	5,288 (56.5%)
	1 (%)	4,064 (43.5%)
timeOUT	mean (sd)	1,288.39 (681.65)
tot_hosp	mean (sd)	2.11 ()
IHC	0 (%)	8,289 (88.7%)
	1 (%)	1,052 (11.3%)
CHARLSON	mean (sd)	2.03 (1.76)
tot_procedures	mean (sd)	0.43 (0.52)

Table E.5: Summary of FVGR Beta Blocking agents final dataset.

Variable	Value	Mixed approach	One tablet a day
PDC	mean (sd)	0.38 (0.22)	0.51 (0.23)
ADERENTE	0 (%)	8,900 (95.3%)	8,314 (89%)
	1 (%)	441 (4.7%)	1,027 (11%)
PDC_CLA	1 (%)	3,256 (34.9%)	1,447 (15.5%)
	2 (%)	3,283 (35.1%)	2,817 (30.2%)
	3 (%)	2,135 (22.9%)	3,504 (37.5%)
	4 (%)	667 (7.1%)	1,573 (16.8%)

Table E.6: Summary of adherence variables of FVGR Beta Blocking agents final dataset dived by approaches.

Anti Aldosterone agents

Variable	Value	
No. patients		6,025
gender	*Male* (%)	2,812 (46.7%)
	Female (%)	3,213 (53.3%)
age_in	mean (sd)	79.28 (9.80)
labelOUT	*Dead* (%)	3,009 (49.9%)
	Censored (%)	3,016 (50.1%)
death	0 (%)	3,016 (50.1%)
	1 (%)	3,009 (49.9%)
timeOUT	mean (sd)	1,264.19 (686.75)
tot_hosp	mean (sd)	2.23 (1.47)
IHC	0 (%)	5,202 (86.3%)
	1 (%)	823 (13.7%)
CHARLSON	mean (sd)	2.01 (1.74)
tot_procedures	mean (sd)	0.44 (0.52)

Table E.7: Summary of FVGR Anti Aldosterone agents final dataset.

Variable	Value	Mixed approach	One tablet a day
PDC	mean (sd)	0.37 (0.21)	0.36 (0.21)
ADERENTE	0 (%)	5,868 (97.4%)	5,870 (97.4%)
	1 (%)	157 (2.6%)	155 (2.6%)
PDC_CLA	1 (%)	2,084 (34.6%)	2,097 (34.8%)
	2 (%)	2,296 (38.1%)	2,296 (38.1%)
	3 (%)	1,364 (22.6%)	1,359 (22.6%)
	4 (%)	281 (4.7%)	273 (4.5%)

Table E.8: Summary of adherence variables of FVGR Anti Aldosterone agents final dataset dived by approaches.

Diuretics

Variable	Value	
No. patients		12,387
gender	*Male* (%)	5,820 (47%)
	Female (%)	6,567 (53%)
age_in	mean (sd)	80.12 (9.42)
labelOUT	*Dead* (%)	6,162 (49.7%)
	Censored (%)	6,225 (50.2%)
death	0 (%)	6,225 (50.2%)
	1 (%)	6,162 (49.7%)
timeOUT	mean (sd)	1,260.45 (677.68)
tot_hosp	mean (sd)	2.08 (1.41)
IHC	0 (%)	10,812 (87.3%)
	1 (%)	1,575 (12.7%)
CHARLSON	mean (sd)	2.10 (1.80)
tot_procedures	mean (sd)	0.42 (0.51)

Table E.9: Summary of FVGR Diuretics final dataset.

Variable	Value	Mixed approach	One tablet a day
PDC	mean (sd)	0.44 (0.23)	0.57 (0.25)
ADERENTE	0 (%)	11,318 (91.4%)	9,628 (77.7%)
	1 (%)	1,069 (8.6%)	2,759 (22.3%)
PDC_CLA	1 (%)	2,885 (23.3%)	1,755 (14.2%)
	2 (%)	4,686 (37.8%)	3,229 (26.1%)
	3 (%)	3,294 (26.6%)	3,874 (31.3%)
	4 (%)	1,522 (12.3%)	3,529 (28.5)

Table E.10: Summary of adherence variables of FVGR Diuretics final dataset dived by approaches.

Appendix F

R Code

In this Appendix we report the main R scripts used to develop our analyses.

F.1 Dataset preprocessing

In this first part, we focus on R scripts concerning the codes of FVGR data rearrangements, as described in details in Section 2.2. In particular, we focus on the second part of the procedure supposing that we have already selected our FVGR cohort of 13,619 patients corresponding to 218,843 events, as explained in Section 2.2.2, and added the auxiliary variables, introduced in 2.2.3, and the Charlson comorbidity index (see Appendix D for details).
At this point, the covariates in our dataset are the following:

```
> load("data_marta_friuli_04.Rdata")
> dim(data_friuli_04)
[1]   218843   36
> names(data_friuli_04)
[1]   "KEY_ANAGRAFE"   "tipo"             "gender"           "ana_data_decesso"
[5]   "fa"             "crtd"             "crt"              "coro"
[9]   "stato"          "data_inizio_tot"  "data_fine_tot"    "ATC"
[13]  "AIC"            "N_PEZZI"          "farma"            "hosp"
[17]  "rsa_index"      "data_rif_ev"      "labelOUT"         "data_studio_out"
[21]  "age_in"         "timeOUT"          "classe_pharma"    "pre_rif"
[25]  "only_post_do"   "pharm"            "death"            "LOS"
[29]  "QTA_MG_CPR"     "QTA_MG_IFAR"      "QTA_MG_CPR_CAR"   "N_CPR"
[33]  "COMBO"          "DDD"              "IHC"              "CHARLSON"
```

which are all described in Section 2.2, except for the following three auxiliary indices:

- `rsa_index` = index of IHC service

- pre_rif = binary variable which marks if the patient has pharmacological events before the reference date

- only_post_do = binary variable which marks if the patient has only hospitalization events after the reference date

Duration of prescriptions

The following codes allow to compute the duration of each pharmacological prescription, as described in Section 2.2.4. The first box concerns the computation using *mixed approach*, while the second one is for *"one tablet a day" approach*.

```
###############################################
# DURATION OF PRESCRIPTIONS - Mixed Approach #
###############################################
rm(list = ls())
library(data.table)

load("data_marta_friuli_04.Rdata")
data=data_friuli_04

## 1 - Compute DOSE
data[farma==T & !(is.na(DDD)), DOSE:=DDD]
data[farma==T & is.na(DDD) & !(is.na(QTA_MG_IFAR)), DOSE:=QTA_MG_IFAR]
data[farma==T & is.na(DDD) & is.na(QTA_MG_IFAR) & !(is.na(QTA_MG_CPR)),
     DOSE:=QTA_MG_CPR]

## 2 - Compute QTA_BOX
data[farma==TRUE & is.na(QTA_MG_IFAR), QTA_BOX:=N_CPR*QTA_MG_CPR]
data[farma==TRUE & !(is.na(QTA_MG_IFAR)), QTA_BOX:=N_CPR*QTA_MG_IFAR]

## 3 - Compute qt_pharma_BOX
data[farma==TRUE & COMBO==0 & !(is.na(DOSE)), qt_pharma_BOX:=floor(QTA_BOX/DOSE)]
data[farma==TRUE & (COMBO==1 | is.na(DOSE)), qt_pharma_BOX:=floor(N_CPR)]

## 4 - Compute qt_pharma
data[farma==TRUE, qt_pharma:=qt_pharma_BOX*N_PEZZI]

## 5 - Recompute variables for unrealistic cases
# qt_pharma_BOX>100
data[farma==T & qt_pharma_BOX>100 & !(is.na(QTA_MG_IFAR)), DOSE:=QTA_MG_IFAR ]
data[farma==T & qt_pharma_BOX>100 & is.na(QTA_MG_IFAR) & !(is.na(QTA_MG_CPR)),
    DOSE:=QTA_MG_CPR ]
data[farma==T & qt_pharma_BOX>100, qt_pharma_BOX:=N_CPR ]

# qt_pharma_BOX<7
data[farma==T & qt_pharma_BOX<7 & !(is.na(QTA_MG_IFAR)), DOSE:=QTA_MG_IFAR]
data[farma==T & qt_pharma_BOX<7 & is.na(QTA_MG_IFAR) & !(is.na(QTA_MG_CPR)),
    DOSE:=QTA_MG_CPR ]
data[farma==T & qt_pharma_BOX<7, qt_pharma_BOX := N_CPR ]

# Recompute qt_pharma
data[farma==TRUE, qt_pharma:=qt_pharma_BOX*N_PEZZI]

## 6 - Compute  data_fine_tot for each pharmacological event
data[farma==T, data_fine_tot:=(data_inizio_tot + qt_pharma - 1)]

```

```
## 7 - Reorder dataset
setcolorder(data, c(
 # global variables
 "KEY_ANAGRAFE","tipo","gender","age_in","data_rif_ev","death",
 "ana_data_decesso","data_studio_out","labelOUT","timeOUT",
 # events
 "data_inizio_tot","data_fine_tot", "pre_rif", "only_post_do",
 # hosp block
 "stato","hosp","rsa_index","RSA","LOS",
 #pharma block
 "farma","pharm", "classe_pharma","ATC","DDD","COMBO","AIC","N_PEZZI",
 "QTA_MG_CPR", "QTA_MG_IFAR", "QTA_MG_CPR_CAR", "N_CPR","DOSE", "QTA_BOX",
 "qt_pharma_BOX","qt_pharma",
 # comorbidity and procedures block
 "CHARLSON","fa", "crtd", "crt", "coro")
)


## FVG complete data with mixed approach
data_FVG_final=data

rm(list=setdiff(ls(), c("data_FVG_final")))
save.image("data_FVG_final.Rdata")
```

```
############################################################
# DURATION OF PRESCRIPTIONS - "One tablet a day" Approach #
############################################################
rm(list = ls())
library(data.table)

load("data_marta_friuli_04.Rdata")
data=data_friuli_04

## 1 - Compute DOSE
data[farma==T & !(is.na(QTA_MG_IFAR)), DOSE:=QTA_MG_IFAR]
data[farma==T & is.na(QTA_MG_IFAR) & !(is.na(QTA_MG_CPR)), DOSE:=QTA_MG_CPR]

## 2 - Compute QTA_BOX
data[farma==TRUE, QTA_BOX:=N_CPR*DOSE]

## 3 - Compute qt_pharma_BOX
data[farma==TRUE, qt_pharma_BOX:=N_CPR]

## 4 - Compute qt_pharma
data[farma==TRUE, qt_pharma:=qt_pharma_BOX*N_PEZZI]

## 5 - Compute data_fine_tot for each pharmacological event
data[farma==T, data_fine_tot:=(data_inizio_tot + qt_pharma - 1)]

## 6 - Reorder dataset
setcolorder(data, c(
 # global variables
 "KEY_ANAGRAFE","tipo","gender","age_in","data_rif_ev","death",
 "ana_data_decesso","data_studio_out","labelOUT","timeOUT",
 # events
 "data_inizio_tot","data_fine_tot", "pre_rif", "only_post_do",
 # hosp block
 "stato","hosp","rsa_index","RSA","LOS",
 #pharma block
 "farma","pharm", "classe_pharma","ATC","DDD","COMBO","AIC","N_PEZZI",
 "QTA_MG_CPR", "QTA_MG_IFAR", "QTA_MG_CPR_CAR", "N_CPR","DOSE", "QTA_BOX",
 "qt_pharma_BOX","qt_pharma",
 # comorbidity and procedures block
 "CHARLSON","fa", "crtd", "crt", "coro")
)


## FVG complete data with "one tablet a day" approach
data_FVG_onlyCPR_final=data

rm(list=setdiff(ls(), c("data_FVG_onlyCPR_final")))
save.image("data_FVG_onlyCPR_final.Rdata")
```

Preselection of final retained variables

The following code is used for a preselection of the final retained variables (see Table 2.31) that represent patient's characteristics which do not depend on the pharmacological class, such as gender, age or Charlson index.

```
########################################
# FVG Patients with Retained Variables #
########################################
rm(list = ls())
library(data.table)

load("data_marta_friuli_04.Rdata")
data=data_friuli_04

## 1 - Select variable to retain
selection=data[hosp==1]
selection=selection[, .(KEY_ANAGRAFE,data_rif_ev,data_studio_out,labelOUT,timeOUT,
     death,tipo,gender,age_in,RSA,CHARLSON,fa,crtd,crt,coro)]

## 2 - Compute tot_procedures
selection[, tot_procedures:=fa+crtd+crt+coro]
selection[, fa:=NULL]
selection[, crtd:=NULL]
selection[, crt:=NULL]
selection[, coro:=NULL]

## 3 - Compute tot_hosp
all_patients=unique(data$KEY_ANAGRAFE)
for(i in 1:length(all_patients)){
 paziente=all_patients[i]
 hosp_max=data[data$KEY_ANAGRAFE==paziente & (data$stato==0 | data$stato==1),hosp]
 selection[KEY_ANAGRAFE==paziente, tot_hosp := max(hosp_max)]
}

## 4 - Convert to factor
selection[, tipo:=as.factor(tipo)]
selection[, labelOUT:=as.factor(labelOUT)]
selection[, gender:=as.factor(gender)]
selection[, IHC:=as.factor(IHC)]

## FVG Patients with Retained Variables
FVG_patients=selection

rm(list=setdiff(ls(), c("FVG_patients")))
save.image("FVG_patients.Rdata")
```

Once executed the code above, we can split the dataset into the five pharmacological classes. The code below performs this computation in the case of *ACE Inhibitors*. During this procedure we must pay particular attention to exclude those patients without ACE events during the observation period: this happens for patients with ACE prescriptions only before the reference date and only the first one hospitalization. The selection of the correct pharmacological cohort of patients is fundamental to proceed to the computation of time-varying curves: in fact, it ensures to avoid any missing value problem.

```r
#-------------------------------------------------#
# ACE Inhibitors Patients with Retained Variables #
#-------------------------------------------------#
rm(list = ls())
library(data.table)

load("FVG_patients.Rdata")
load("data_marta_friuli_04.Rdata")
data=data_friuli_04

classe = "ACEINIBITORI"   # Change for another pharmacological class

# Patients with pharmacological events before the reference date
pre=unique(data[classe_pharma==classe & pre_rif==1]$KEY_ANAGRAFE)

# Patients with pharmacological events before the reference date and at least one
    hospitalization/IHC after the first discharge for HF
pre_with_hosp=unique(data[KEY_ANAGRAFE %in% pre & (hosp>1 | RSA==1)]$KEY_ANAGRAFE)

# Patients with pharmacological events after the reference date
post=unique(data[classe_pharma==classe & pre_rif==0]$KEY_ANAGRAFE)

codici=union(post, pre_with_hosp)
# In this way we do not consider patients with ACE events Only before the
    reference date and only the first one hospitalization
length(codici) #8481

## ACE Inhibitors Patients with Retained Variables
acei=FVG_patients[KEY_ANAGRAFE %in% codici]

rm(list=setdiff(ls(),c("acei")))
save.image("FVG_in_acei.Rdata")
```

Time-varying curves and final datasets

The following code concerns the computation of the adherence variables described in 2.2.5, of the curves of cumulative days covered by drug assumption, as explained in Sections 2.2.6, and of the curves of assumed dose, as explained in 2.2.7.
The code is set to

- pharmacological class of *ACE Inhibitors*
- duration of prescriptions computed with *mixed approach*

```r
#########################################################
# ACE Inhibitors Time-Varying curves and Final dataset #
#########################################################
rm(list = ls())
library(data.table)

load("data_FVG_final.Rdata")
data=data_FVG_final

classe="ACEINIBITORI"
load("FVG_in_acei.Rdata")
temp_data = acei

## 0 - List of ACE Inhibitors correct patients
codici=unique(temp_data$KEY_ANAGRAFE)

## 1 - Auxiliary variables for time-varying curves computation
data[, data_inizio:=data_inizio_tot] # Start date
data[, data_fine:=data_fine_tot] # End date
data[farma==F, quant:=LOS] # Duration
data[farma==T, quant:=qt_pharma] # Duration

## 2 - Check on missing values
data.na = data[is.na(quant)]
data = data[!is.na(quant)]

## 3 - Set ADERENTE and ADERENZA to 0
temp_data[, ADERENTE:=0] # Binary variable for adherent pts
temp_data[, ADERENZA:=0] # Coverage days

## 4 - Auxiliary matrix for curves of days
DAYS=matrix(0, length(unique(codici)), 366, byrow=F)
DAYS[, 1]=unique(codici)
rownames(DAYS)=seq(1:length(unique(codici)))
nomi=NULL
for(i in 1:365){ nomi=c(nomi, paste("day_", i, sep="")) }
colnames(DAYS)=c("KEY_ANAGRAFE", nomi)

## 5 - Auxiliary matrix for curves of dose
MG=matrix(0, length(unique(codici)), 366, byrow=F)
MG[, 1]=unique(codici)
rownames(MG)=seq(1:length(unique(codici)))
nomi=NULL
for(i in 1:365){ nomi=c(nomi, paste("dose_", i, sep="")) }
colnames(MG)=c("KEY_ANAGRAFE", nomi)

## 6 - For loop on list of ACE patients
for (i in 1:length(codici)){
  paz_corrente=codici[i] # Current patient
  data_rif=data[KEY_ANAGRAFE==paz_corrente, unique(data_rif_ev)] # Reference date
  data_stop=(data_rif+365) # End of observation period

  # Vectors of events start dates, end dates, durations and doses
  vet_inizio=data[KEY_ANAGRAFE==paz_corrente & (classe_pharma==classe | farma==F),
       data_inizio]
  vet_fine=data[KEY_ANAGRAFE==paz_corrente & (classe_pharma==classe | farma==F),
      data_fine]
  vet_quant=data[KEY_ANAGRAFE==paz_corrente & (classe_pharma==classe | farma==F),
      quant]
  vet_dose=data[KEY_ANAGRAFE==paz_corrente & (classe_pharma==classe | farma==F),
      DOSE]
```

```r

# Median of doses
mediana = median(vet_dose, na.rm=T)

# Type of events
vet_stato=data[KEY_ANAGRAFE==paz_corrente & (classe_pharma==classe | farma==F), stato]
vet_hosp=which(!(is.na(vet_stato)))
vet_pharma=which(is.na(vet_stato))

# Set doses in hospitalizaions and IHC
for(j in 1:length(vet_hosp)){
 ok=which(vet_pharma<vet_hosp[j])
 if(length(ok)>=1){
  vet_dose[vet_hosp[j]]=vet_dose[vet_pharma[ok[length(ok)]]]
 }else{
  vet_dose[vet_hosp[j]]=mediana
 }
}

# Delete the first hosp and events before the reference date
pre_rif=data[KEY_ANAGRAFE==paz_corrente & (classe_pharma==classe | farma==F), pre_rif]
keep=which(pre_rif==0)
vet_inizio=vet_inizio[keep]
vet_fine=vet_fine[keep]
vet_quant=vet_quant[keep]
vet_dose=vet_dose[keep]

# Recompute the durations of events considering only distinct days
vet_inizio_start=vet_inizio
vet_fine_start=vet_fine
if(length(vet_inizio)>1){ # if I have more than one event
 for (j in 2:length(vet_inizio)){
  # if the beginning of the next event is before the end of the previous one
  if (vet_inizio[j]<vet_fine[j-1]){
   vet_inizio[j] = vet_fine[j-1]+1 # postpone the start of the next event
   # if the beginning of the next event is after its end
   if(vet_inizio[j]>vet_fine[j]){
    vet_fine[j]=vet_inizio[j] # postpone the end
   }
  }
 }
}

# Recompute vectors of events start dates and end dates
for (h in 1:length(vet_inizio)) {
 vet_inizio[h]=min(vet_inizio[h],data_stop-1) # start date before the end of fup
 vet_fine[h]=min(vet_fine[h],data_stop-1) # end date before the end of fup
 vet_inizio[h]=max(vet_inizio[h],data_rif) # start date after the reference date
 vet_fine[h]=max(vet_fine[h],data_rif) # end date after the reference date
}

# Difference in days between start dates and reference date
index_inizio= difftime(vet_inizio, data_rif, units="days")
# Difference in days between end dates and reference date (plus 1 because the end date represents the last day covered by the event)
index_fine=difftime(vet_fine, data_rif, units="days") + 1

# Vectors of daily coverage and daily dose
day=rep(0,365) # days covered by drug assumption: 1 or 0
mg=rep(0,365) # daily assumed dose: dose (mg) or 0
```

```
  last=0
  for(l in 1:length(vet_inizio)){
   inizio=index_inizio[l]+1
   fine=index_fine[l]
    if(inizio<=365 & inizio<=fine){
     for(m in inizio:fine){
      day[m]=1
      mg[m]=vet_dose[l]
     }
    }
  }

  # Vectors of time-varying curves for the current patient
  # cumsum(day) is the curve cumulative days covered by drug assumption
  DAYS[i,2:366]=cumsum(day)
  # cumsum(mg) is the curve of cumulative assumed dose
  MG[i,2:366]=cumsum(mg)

  # Coverage days for current patient
  temp_data[KEY_ANAGRAFE==paz_corrente, ADERENZA:=sum(day)]
}



## 7 - PDC: Proportion of Days Covered
temp_data[, PDC:=ADERENZA/365]

## 8 - PDC_CLA: Adherence levels
temp_data[ADERENZA>=0 & PDC<0.25, PDC_CLA:=1]
temp_data[ADERENZA>0 & PDC<0.50 & PDC>=0.25, PDC_CLA:=2]
temp_data[ADERENZA>0 & PDC<0.75 & PDC>=0.5, PDC_CLA:=3]
temp_data[ADERENZA>0 & PDC>=0.75, PDC_CLA:=4]
temp_data[, PDC_CLA:=as.factor(PDC_CLA)]

## 9 - ADERENTE: dherent and non adherent patients
temp_data[PDC>=0.8, ADERENTE:=1]
temp_data[, ADERENTE:=as.factor(ADERENTE)]

## 10 - ACE INHIBITORS FINAL DATASET
acei_tot=as.data.table(cbind(temp_data,DAYS[,2:366],MG[,2:366]))

rm(list=setdiff(ls(),c("acei_tot")))
save.image("FVG_ACE.Rdata") # save.image("FVG_ACE_onlyCPR.Rdata")
```

F.2 Applications

In the following Sections we report the codes used for the applications on *ACE Inhibitors final dataset with mixed approach* computed in the previous Section.

In particular, we present present codes for Cox's Proportional Hazard models and for Joint Models. The background theory of these models are described in Sections 3.1.5 and 3.2, respectively, and the results given by their applications are reported in Sections 4.2.3 and 4.2.4, respectively.

Cox PH models

```r
##############
# COX MODELS #
##############
rm(list = ls())
library(data.table)
library(survival)
library(survminer)

load("FVG_ACE.Rdata")
data = acei_tot

# Cox's model
cox_mod = coxph(Surv(timeOUT & death) ~  age_in + gender + tot_hosp + CHARLSON +
                                         IHC + ADERENTE, data=data)
summary(cox_mod)

# KM of survival stratified by adherent and non andherent patients
new_data = with(data,
                data.frame(age_in = rep(mean(age_in, na.rm = TRUE), 2),
                           gender = c("F","F"),
                           tot_hosp = c(2,2),
                           CHARLSON = c(2,2),
                           IHC = c(1,1),
                           ADERENTE = c("0","1")))
new_data$gender=factor(new_data$gender, levels=c("M","F"))

fit=survfit(cox_mod, newdata=new_data)
x11()
ggsurvplot(fit, data=new_data, conf.int=TRUE,
           legend.labs=c("ADERENTE=0","ADERENTE=1"), ggtheme=theme_minimal(),
           break.time.by=365, xlab="Time (days)", xlim=c(365,2555), censor=F)
```

Joint models

To use `jointModel()` function of `JM` package [26], our data need to be in two precise formats.

For the longitudinal part of JM, we need a dataset with more rows for each patient, one for each different value of his/her curve of cumulative days covered by drug assumption with the respective time of observation. Therefore, for each patient, each row is different from the others since observation times and curve values change each time. Our longitudinal dataset for JM, that we called `matrix`, contains:

- `KEY_ANAGRAFE`, `death`, `timeOUT`, `age_in`, `gender`, `tot_hosp`, `CHARLSON`, `IHC`
- `obstime` = observation time
- `cum_days` = observed value of cumulative days covered by drug assumption

For the survival part of JM, we need a dataset, that we called `matrix.id`, with only one row for each patient containing:

- `KEY_ANAGRAFE`, `death`, `timeOUT`, `age_in`, `gender`, `tot_hosp`, `CHARLSON`, `IHC`
- `obstime` = first observation time (always 1)
- `day_1` = curve of cumulative days covered by drug assumption at $t = 1$

The computation of these matrices is done through the following code:

```
###############
# JM matrices #
###############
rm(list = ls())
library(data.table)

load("FVG_ACE.Rdata")
data=acei_tot

## 0 - List of ACE patients
pazienti=unique(data$KEY_ANAGRAFE)

## 1 - Matrix for the longitudinal part (matrix)
matrix=NULL
for(i in 1:length(pazienti)){
 current=pazienti[i]
 # Select variables
 row=c(as.double(t(data[KEY_ANAGRAFE==current, .(KEY_ANAGRAFE,death,timeOUT,
 age_in,tot_hosp,CHARLSON,IHC)])), data[KEY_ANAGRAFE==current]$gender)
 # Compute obstime and days_unique
 days_all=as.vector(t(data[KEY_ANAGRAFE==current, 19:383]))
 days_unique=unique(days_all) # vector of cum_days values
 obstime=NULL
 for(j in 1:length(days_unique)){
  obstime=c(obstime &min(which(days_all==days_unique[j])))
 }
 nR=length(obstime)
 mat_paz=matrix(as.vector(row), nrow=nR, ncol=8, byrow=T)
 mat_paz=cbind(mat_paz, obstime, days_unique)
 matrix=rbind(matrix, mat_paz)
}

rownames(matrix)=seq(1,dim(matrix)[1])
colnames(matrix)=c("KEY_ANAGRAFE","death","timeOUT","age_in","tot_hosp",
                   "CHARLSON","IHC","gender","obstime","cum_days")
matrix=as.data.frame(matrix) # N.B. matrix deve essere un data.frame
matrix$gender[which(matrix$gender==1)]="F"
matrix$gender[which(matrix$gender==2)]="M"

## 2 - Matrix for the the survival part (matrix.id)
matrix.id=data[KEY_ANAGRAFE %in% pazienti, .(KEY_ANAGRAFE,death,timeOUT,age_in,
                                             tot_hosp,CHARLSON,gender,IHC,day_1)]
matrix.id[, obstime:=1]

## 3 - Reorder
setcolorder(matrix,c("KEY_ANAGRAFE","death","timeOUT","age_in","gender",
                     "tot_hosp","CHARLSON","IHC","obstime","cum_days"))
```

```
setcolorder(matrix.id,c("KEY_ANAGRAFE","death","timeOUT","age_in","gender",
                        "tot_hosp","CHARLSON","IHC","obstime","day_1"))

## 4 - Save matrix and matrix.id
rm(list=setdiff( ls(), c("matrix","matrix.id")))
save.image("JM_ACE_days.Rdata")
```

The obtained matrix for the longitudinal part (`matrix`) is:

```
> load("JM_ACE_days.Rdata")   # matrix, matrix.id
> dim(matrix)
[1]   1602415   10
> as.data.table(matrix)
KEY_ANAGRAFE   death   timeOUT   age_in   gender   tot_hosp   CHARLSON   IHC   obstime   cum_days
         689       1      2394       78        F          1          3     0         1          0
         689       1      2394       78        F          1          3     0        25          1
         689       1      2394       78        F          1          3     0        26          2
         689       1      2394       78        F          1          3     0        27          3
         689       1      2394       78        F          1          3     0        28          4
          —
     3450450       1       577       89        F          1          0     0       354        189
     3450450       1       577       89        F          1          0     0       355        190
     3450450       1       577       89        F          1          0     0       356        191
     3450450       1       577       89        F          1          0     0       357        192
     3450450       1       577       89        F          1          0     0       365        193
```

The obtained matrix for the survival part (`matrix.id`) is:

```
> load("JM_ACE_days.Rdata")   # matrix, matrix.id
> dim(matrix.id)
[1]   8481   10
> matrix.id
KEY_ANAGRAFE   death   timeOUT   age_in   gender   tot_hosp   CHARLSON   IHC   obstime   day_1
         689       1      2394       78        F          1          3     0         1       0
         950       1      1230       85        M          2          3     1         1       0
        1014       1      1309       86        M          3          1     0         1       0
        2285       0      2874       77        M          1          5     0         1       0
        2736       0      2458       80        F          1          0     0         1       0
          —
     3447805       0       932       91        F          1          1     1         1       1
     3447928       1      1542       97        F          2          0     0         1       0
     3448135       0      1857       74        M          3          1     0         1       1
     3448477       1       916       95        M          2          1     0         1       0
     3450450       1       577       89        F          1          0     0         1       0
```

Once the matrices have been computed, we are able to fit the model, using the following code:

```
################
# JOINT MODELS #
################
rm(list = ls())
library(data.table)
library(JM)
library(lattice)
library(survival)
```

```r

load("JM_ACE_days.Rdata") # matrix, matrix.id
head(matrix)
head(matrix.id)

# Longitudinal process
ctrl=lmeControl(opt="optim")
fitLME=lme((cum_days^(1/3) ~ obstime, random=~obstime|KEY_ANAGRAFE, control=ctrl,
    data=matrix)
summary(fitLME)

# Time-to-event process
fitSURV=coxph(Surv(timeOUT &death) ~ age_in + gender + tot_hosp + CHARLSON + IHC,
    data=matrix.id, x=TRUE)
summary(fitSURV)

# Joint model
fit.JM=jointModel(fitLME, fitSURV, timeVar="obstime", method="piecewise-PH-GH")
summary(fit.JM)

rm(list=setdiff(ls() & c("fitSURV","fitLME","fit.JM")))
save.image("piecewise3_JM_ace_days.Rdata")
```

In the call to coxph() we need to specify the argument x=TRUE in order for the design matrix of the Cox model to be included in the returned object, as mentioned in [26].

The code below produces the the plots of residuals:

```r
#-----------------#
# Residuals Plots #
#-----------------#
rm(list = ls())
library(data.table)
library(JM)
library(lattice)
library(survival)

load("JM_ACE_days.Rdata") # matrix, matrix.id
load("piecewise3_JM_ace_days.Rdata") # fitSURV, fitLME, fit.JM


# plotResid
plotResid = function (x, y, ...) {
 plot(x,y,...)
 lines(lowess(x,y),col="red",lwd=2)
 abline(h=0,lty=3,col="grey",lwd=2)
}


## 1 - Cox-Snell Residual
x11()
resCST=residuals(fit.JM, process="Event", type="CoxSnell")
sfit=survfit(Surv(resCST, death) ~ 1, data = matrix.id)
plot(sfit, mark.time=FALSE, conf.int=TRUE, lty=c(1,2,2),
     xlab="Cox-Snell Residuals", ylab="Survival Probability",
     main="Survival Function of Cox-Snell Residuals")
curve(exp(-x), from=0, add=TRUE, col="red", lwd=2)

```

```
## 2 - Subject-Specific and Marginal Residuals
x11()
par(mfrow=c(1,2))

resSubY=residuals(fit.JM, process="Longitudinal", type="stand-Subject")
fitSubY=fitted(fit.JM, process="Longitudinal", type="Subject")
plotResid(fitSubY, resSubY, xlab="Fitted Values", ylab="Residuals",
          main="Subject-Specific Residuals vs Fitted Values")

resMargY=residuals(fit.JM, process = "Longitudinal", type="stand-Marginal")
fitMargY=fitted(fit.JM, process = "Longitudinal", type = "Marginal")
plotResid(fitMargY, resMargY, xlab = "Fitted Values", ylab = "Residuals",
          main="Marginal Residuals vs Fitted Values")
```

The code below computes the expected survival for a group of selected patients (line 16) and it produces the respective plots:

```
 1 #-------------------#
 2 # Expected Survival #
 3 #-------------------#
 4 rm(list = ls())
 5 library(data.table)
 6 library(JM)
 7 library(lattice)
 8 library(survival)
 9 
10 load("JM_ACE_days.Rdata") # matrix, matrix.id
11 load("piecewise3_JM_ace_days.Rdata") # fitSURV, fitLME, fit.JM
12 load("FVG_ACE.Rdata") # acei_tot
13 data=acei_tot
14 
15 ## 1 - Select patients
16 paz_selected=c(1647184, 1246096, 2968114, 672100)
17 ND=matrix[matrix$KEY_ANAGRAFE %in% paz_selected, ]
18 ND$gender=factor(ND$gender, levels=c("M","F"))
19 
20 ## 2 - Characteristics of selected patients
21 data[KEY_ANAGRAFE %in% paz_selected, .(KEY_ANAGRAFE,timeOUT,death,gender,age_in,
       IHC,CHARLSON,tot_hosp,PDC,day_365)]
22 
23 ## 3 - Compute expected survival for selected patients
24 set.seed(123)
25 predSurv=survfitJM(fit.JM, newdata=ND, idVar="KEY_ANAGRAFE", last.time="obstime")
26 predSurv
27 
28 ## 4 - Plots
29 x11()
30 par(mfrow=c(2,2))
31 plot(predSurv, which=as.character(paz_selected[1]), conf.int=TRUE,
32      ylab="Survival Probability", xlab="Time (days)", lwd=2)
33 plot(predSurv, which=as.character(paz_selected[2]), conf.int=TRUE,
34      ylab="Survival Probability", xlab="Time (days)", lwd=2)
35 plot(predSurv, which=as.character(paz_selected[3]), conf.int=TRUE,
36      ylab="Survival Probability", xlab="Time (days)", lwd=2)
37 plot(predSurv, which=as.character(paz_selected[4]), conf.int=TRUE,
38      ylab="Survival Probability", xlab="Time (days)", lwd=2)
```

Bibliography

[1] AALEN O, *Nonparametric estimation of partial transition probabilities in multiple decrement models.*, The Annals of Statistics, Vol. 6, No. 3 (1978), pp. 534-545.

[2] AALEN O, BORGAN O, GJESSING HK, *Survival and Event history Analysis*, Springer Science & Business Media, (2008).

[3] ANDERSEN PK, GILL RD, *Cox's regression model for counting processes: a large sample study*, The Annals of Statistics, Vol. 10, No. 4 (Dec, 1982), pp. 110-1120.

[4] ANDRADE SE, KAHLER KH, FRECH F, CHAN FA, *Methods for evaluation of medication adherence and persistence using automated databases*, Pharmacoepidemiology and Drug Safety, Vol. 15 (2006), pp. 565–574.

[5] CHRISTENSEN R, JOHNSON W, BRANSCUM A, HANSON TE, *Bayesian Ideas and Data Analysis. An Introduction for Scientists and Statisticians*, Taylor and Francis Group, (2011).

[6] CORRAO G, GHIRARDI A, IBRAHIMA B, MERLINO L, MAGGIONI AP, *Short- and long-term mortality and hospital readmissions among patients with new hospitalization for heart failure: A population-based investigation from Italy*, International Journal of Cardiology, Vol. 181 (2015), pp. 81–87

[7] CORRAO G, MANCIA G *Generating Evidence From Computerized Healthcare Utilization Databases*, Hypertension, Vol.65 (2015), pp. 490-498.

[8] CORRAO G, PARODI A, NICOTRA F, ZAMBON A, MERLINO L, CESANA G, MANCIA G, *Better compliance to antihypertensive medications reduces cardiovascular risk*, Journal of Hypertension, Vol. 29, No. 3 (2011), pp. 610–618.

[9] COX DR, *Regression Models and Life-Tables*, Journal of the Royal Statistical Society, Vol. 34, No. 2 (1972), pp. 187-220.

[10] DICKSTEIN K et al., *ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM)*, European Heart Journal, Vol. 29, No. 19 (2008), pp. 2388- 2442.

[11] GAGNE JJ, GLYNN RJ, AVORN J, LEVIN R, SCHNEEWEISS S, *A combined comorbidity score predicted mortality in elderly patients better than existing scores.*, Journal of clinical epidemiology, Vol. 64, No. 7 (2011), pp. 749-759.

[12] GASPERONI F, IEVA F, BARBATI G, SCAGNETTO A, IORIO A, SINAGRA G, DI LENARDA A, *Multi-state modelling of heart failure care path: A population-based investigation from Italy*, PLoS ONE, Vol. 12 (Jun, 2017), pp. 1-15.

[13] HUNT SA et al., *2009 Focused update incorporated into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines: developed in collaboration with the International Society for Heart and Lung Transplantation*, Journal of the American College of Cardiology, Vol. 53, No. 15 (2009), pp. 1-90.

[14] IEVA F, JACKSON CH, SHARPLES LD, *Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology*, Statistical Methods in Medical Research, (2015).

[15] IEVA F, PAGANONI AM, PIGOLI D, VITELLI V, *Multivariate functional clustering for the morphological analysis of electrocardiograph curves*, Journal of the Royal Statistical Society, Applied Statistics, Vol. 62 (2013), Part 3, pp. 401–418.

[16] KALBFLEISCH JD and PRENTICE RL, *The statistical analysis of failure time data*, John Wiley & Sons, Vol. 360 (2011).

[17] KAPLAN EL and MEIER P, *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association, Vol. 53, No. 282 (1958), pp. 457-481.

[18] KARVE S, CLEVE MA, HELM M, HUDSON TJ, WEST DS, MARTIN BC, *Prospective Validation of Eight Different Adherence Measures for Use with Administrative Claims Data among Patients with Schizophreniavhe*, Value In Health, Vol. 12, No. 6 (2009).

[19] KLEINBAUM DG, KLEIN M, *Survival Analysis: A Self-Learning Text*, Statistics for Biology and Health, Springer Science & Business Media, (2005).

[20] LEVY D et al., *Long-Term Trends in the Incidence of and Survival with Heart Failure*, The New England Journal of Medicine, Vol. 347, No. 18 (2002), pp. 1397-1402.

[21] MAGGIONI AP, SPANDONARO F, *Lo scompenso cardiaco acuto in italia*, Giornale italiano di cardiologia, Vol. 15, No. 2 (Suppl. 2) (2014), pp. 3S-4S.

[22] MAZZALI C et al., *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 Service Research, Vol. 16, No. 234 (2016).

[23] PAZOS-LOPEZ P et al., *The causes, consequences, and treatment of left or right heart failure*, Vascular Health and Risk Management, Vol. 7 (2011), pp. 237-254.

[24] R CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, (2016).
`https://www.R-project.org/`

[25] RIZOPOULOS D, *Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data*, Biometrics, Vol. 67 (Sep. 2011), pp. 819-829.

[26] RIZOPOULOS D, *JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data*, Journal of Statistical Software, Vol. 35, No. 9 (2010), pp. 1–33.
`http://www.jstatsoft.org/v35/i09/`

[27] RIZOPOULOS D, *Package 'JM'*, (June, 2017).
`https://cran.r-project.org/web/packages/JM/JM.pdf`

[28] RIZOPOULOS D, VERBEKE G, LESAFFRE E, *Fully Exponential Laplace Approximations for the Joint Modelling of Survival and Longitudinal Data*, Journal of the Royal Statistical Society B, No. 71 (2009), pp. 637–654.

[29] ROGER VL et al., *Trends in heart failure incidence and survival in a community-based population*, The Journal of the American Medical Association, Vol. 292, No. 3 (2004), pp. 344-350.

[30] STROM B, *Textbook of Pharmacoepidemiology*, West Sussex, England: John Wiley and Sons, (2006).

[31] THERNEAU TM, *A Package for Survival Analysis in S*, version 2.38, (2015). `https://CRAN.R-project.org/package=survival`

[32] THERNEAU TM, GRAMBSH PM, *Modeling Survival Data: Extending the Cox Model*, Statistics for Biology and Health, Springer Science & Business Media, New York (2000). ISBN 0-387-98784-3.

[33] THERNEAU TM, LUMLEY T, *Package 'survival'*, (Apr, 2017). `http://cran.irsn.fr/web/packages/survival/survival.pdf`

[34] WORLD HEALTH ORGANIZATION, *Introduction to Drug Utilization Research*, WHO Library Cataloguing-in-Publication Data, (2003).

Ringraziamenti

So di non essere solita ad esternare le mie sensazioni ma questa volta mi sembra doveroso ringraziare le persone che mi sono state accanto durante questo lungo percorso.

Vorrei innanzitutto ringraziare la professoressa Francesca Ieva e la dottoressa Francesca Gasperoni per avermi seguita passo a passo in questo lavoro così stimolante, facendomi entrare nel mondo della statistica biomedica. Le ringrazio per tutto il supporto che mi hanno dato in questi mesi, per tutti i consigli, le email e per la loro incessante disponibilità. Le ringrazio soprattutto per aver creduto in me e nelle mie capacità, permettendomi di condurre il lavoro in prima linea ed accettando con entusiasmo le mie idee. Grazie alla loro guida ho avuto la possibilità di imparare moltissimo, arricchendo ulteriormente il bagaglio costruito in questi cinque anni.

Ringrazio il dottor Federico Rea per la disponibilità a chiarire ogni mio dubbio, per tutto il materiale che ci ha fornito e per i suoi utilissimi consigli.

Ringrazio dal profondo del cuore tutta la mia famiglia, che ha saputo supportarmi ma, soprattutto, sopportarmi anche quando lo studio matto e disperatissimo per gli esami mi faceva sembrare pazza. Ringrazio mia mamma Veronica per tutto quello che ha sempre fatto per me, le parole non basterebbero ad esprimere tutta la mia gratitudine. Ringrazio i miei fratelli, Roberto, Riccardo e Tommaso, che ogni giorno mi mancano da morire, perché il loro affetto è sempre stata la mia forza più grande. Spero di essere stata un buon esempio per loro. Ringrazio i miei nonni, Ermanno e Stefania, i miei primi sostenitori, per tutto l'amore che mi hanno dato e per essersi presi l'impegno di portarmi ogni giorno in stazione affinché io arrivassi in tempo al Poli. Ringrazio mio nonno Luigi e Rosanna, per essersi sempre dimostrati entusiasti del percorso che ho intrapreso e della persona che sono diventata. Ringrazio mio zio Carlo, perché so che col cuore mi è sempre stato vicino, sin da quando mi portava in giro dentro ad una coperta. Ringrazio mio zio Claudio e Romina, per avermi sempre sostenuta, soprattutto nei momenti più difficili, e per avermi sempre detto che devo essere orgogliosa di chi sono. E ringrazio i miei cugini, Emma, Nicolò, Anna e Mattia, per i sorrisi che mi hanno sempre donato. Ringrazio tutti voi per essere stati la mia forza psicologica in questo percorso e perché, nonostante la lontananza, non mi avete mai fatta sentire sola.

Ringrazio tutti i miei amici del Poli, che con me hanno condiviso le gioie ed i dolori di questi anni. Dalle lezioni agli esami, dai progetti ad ARF, dalle partite a lupus alla briscola a chiamata, dai gossip al fantacalcio, dalla palestra agli aperitivi ed a qualsiasi altra cosa. Un particolare grazie va ad Anna, Margherita, Giordano ed Alberto.

Ringrazio la mia compagna di stanza, Giada, e la nostra terza coinquilina, Michela, per aver rallegrato le mie giornate in CDS con la loro compagnia, anche nelle più buie sessioni d'esame.

Ringrazio la mia migliore amica Rachele, per esserci ancora.

Ringrazio mia zia Carlotta, che meritava un paragrafo a parte. La ringrazio per tutta la vita vissuta insieme, l'una al fianco dell'altra. La ringrazio per aver avuto l'enorme pazienza di sopportare un carattere difficile e testardo come il mio. La ringrazio perché ha sempre un pensiero per me e per l'entusiasmo che dimostra nel sentirmi raccontare dei miei esami e progetti, anche quando per lei sono difficilmente comprensibili. Ma soprattutto la ringrazio perché sa emozionarsi per i miei successi, dimostrandomi sempre l'affetto incondizionato che ci unisce.

Infine ringrazio la persona che in questi cinque lunghi anni ha condiviso con me ogni attimo. Ringrazio Riccardo, per avermi dato la forza di non mollare mai, anche quando tutto sembrava terribilmente difficile. Lo ringrazio dal cuore per essere riuscito a stare al mio fianco, proteggendomi con il suo affetto e mettendomi sempre al primo posto. Lo ringrazio per la persona meravigliosa che è sia con me sia con la mia famiglia. Ma soprattutto lo ringrazio per tutte le gioie che insieme abbiamo vissuto e per aver sempre creduto con orgoglio nelle mie capacità, più di quanto ci credessi io stessa. Se ho raggiunto questo traguardo oggi è anche grazie al coraggio che mi ha saputo trasmettere, dimostrandomi che l'amore è un continuo crescere e migliorarsi insieme.

Grazie a tutti, davvero, di cuore.
Marta