Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Corso di Laurea Magistrale in Ingegneria Matematica

MODELS FOR PREDICTING READMISSIONS IN HEART FAILURE PATIENTS: A COMPARISON BETWEEN LOMBARDIA AND ENGLAND

Relatore:
Prof.ssa Annamaria Paganoni

Co-Relatori:
Prof. Alex Bottle
Prof.ssa Francesca Ieva

Tesi di Laurea Magistrale di:
Chiara Maria Ventura
Matr. 804479

Anno Accademico 2013 - 2014

Contents

1 Introduction **1**

2 Background and Motivations **3**
2.1 The Dataset from Lombardia 3
2.2 The Dataset from England 5
2.3 A Common Preliminary Analysis 5

3 Chosen Methods: Theory and Implementation **14**
3.1 Logistic Regression 15
3.1.1 Theoretical structure 15
3.1.2 Adaptation to the problem 16
3.2 Hurdle and Zero-Inflated Models 20
3.2.1 Theoretical structure 20
3.2.2 Adaptation to the problem 21
3.3 Multi-State Models 24
3.3.1 Theoretical structure 24
3.3.2 Adaptation to the problem 25

4 Analysis of the Results **29**
4.1 Results of Logistic Regression 29
4.2 Results of Hurdle and Zero-Inflated Models 38
4.2.1 Models overview 38
4.2.2 Outcomes of Chosen Counting Models 40
4.3 Results of Multi-State Models 47
4.3.1 Three states model 47
4.3.2 Multi-State Model with detailed admissions 54

5 Conclusive Remarks **59**

6 Code **62**

List of Figures

2.1 Histogram of age at first admission in Lombardia (above) and in England (below) dataset 7
2.2 In Hospital mortality rate and 95% Confidence Intervals of English patients. All admissions (above), only emergency admissions (below). 8
2.3 In Hospital mortality rate and 95% Confidence Intervals of Lombardia patients. 9
2.4 Distribution of frequency of *renal* (above), *pulmonarydz* (middle) and *arrhythmia* (below) disease along age of first admission. Lombardia dataset. 12
2.5 Distribution of frequency of *renal* (above) and *pulmonarydz* (middle) and *alcohol* (below) disease along age of first admission. England dataset. 13

3.1 Dataset for Logistic Regression: first readmission and second readmission. 17
3.2 Readmission Rates within 30 days along *Age* and *LOS*. Lombardia (above) and England (below) dataset. Limits of Y axes: from 0 to 0.4. 19
3.3 Dataset for Hurdle and Zero-Inflated Models before (above) and after (below) the shrinkage. 22
3.4 Multistate Model with three states. 26
3.5 Dataset for Multi-State Models before (above) and after (below) the adjustment. 27
3.6 Multi-State Model with specification of admissions/discharge. . . 27

4.1 Odds Ratio and confidence intervals for the procedures of all Logistic regression models. Lombardia (above) and England (below) dataset. First readmission: red and blue. Second readmission: orange and light blue. 33
4.2 Odds Ratio and confidence intervals for the comorbidities of all models. Lombardia (above) and England (below) dataset. First readmission: red and blue. Second readmission: orange and light blue. 34
4.3 Odds Ratio and confidence intervals for the comorbidities of all models. Lombardia (above) and England (below) dataset. First readmission: red and blue. Second readmission: orange and light blue. 36

4.4 Odds Ratio and confidence intervals for Logistic Regression with all rows. Lombardia (orange) and England (light blue) dataset. All covariates. 37
4.5 Histogram of readmissions per patient within a year since the first admission. Lombardia and England dataset. 39
4.6 Confidence intervals for Zero-Hurdle (Zero counting part in Hurdle model) exponential coefficients (blue: England, red: Lombardia), and Zero-Inflation (Zero inflation part in Zero-Inflated model) coefficients (light blue: England, orange: Lombardia). . . 43
4.7 Confidence intervals for comorbidities exponential coefficients in counting part of Hurdle model (blue: England, red: Lombardia), and of Zero-Inflation model (light blue: England, orange: Lombardia). 44
4.8 Trend of probability of readmission in Heart Failure patients. Lombardia dataset (red) and England dataset(blue). From 0 to 4 years (left) and from 0 to 2 months (right). 49
4.9 Trend of probability of death inside the hospital(left) and discharge (right) in Heart Failure patients. Lombardia dataset (red) and England dataset(blue). From 0 to 4 years (above) and from 0 to 2 months (below). 50

List of Tables

2.1 Percentage or Mean (SD) for original dataset 6
2.2 Comorbidities percentage referred to patients not dead inside the hospital during the first admission. Lombardia dataset (34,146 patients), England dataset (249,156 patients) 10
2.3 Procedures percentage referred to patients discharged alive from the first admission. Lombardia dataset (34,146 patients), England dataset (249,156 patients) 11

3.1 General description of Logistic regression 14
3.2 General description of Hurdle and Zero-Inflated models 15
3.3 General description of Multi-State models 15
3.4 Dimension of the dataset and percentage of readmission within 30 days in Logistic regression. Lombardia and England dataset. . 17

4.1 Odds ratio from Logistic regression with first readmission. Lombardia and England dataset. 31
4.2 Odds ratio from Logistic regression with second readmission. Lombardia and England dataset. 32
4.3 Estimate of goodness of fit for all counting models (-2×LogLikelihood and zeros predicted). Lombardia and England dataset. 40
4.4 Zero coefficients for Hurdle model (binomial with logit link). Lombardia and England dataset. 45
4.5 Counting coefficients for Hurdle model (Negative Binomial). Lombardia and England dataset. 46
4.6 Probability of each state being next, conditional to the change of state. Multi-State model with three states. Lombardia and England dataset. 48
4.7 Mean Sojourn Time and Total Length of Stay. Multi-State model with three states. Lombardia and England dataset. 48
4.8 Exponential hazard rate: Alive Inside the Hospital to Alive Outside the Hospital (live discharge). 52
4.9 Exponential hazard rate: Alive Inside the Hospital to Death (death inside the hospital). 52
4.10 Exponential hazard rate: Alive Outside the Hospital to Alive Inside the Hospital (readmission). 53
4.11 Probability of each state being next, conditional to the change of state. Multi-State model with detailed admissions. Lombardia and England dataset. 55

4.12 Mean Sojourn Time of Multi-State model with detailed admissions. Lombardia and England dataset. 55
4.13 Exponential hazard rate of ICD and CABG. Lombardia and England dataset. 56
4.14 Exponential hazard rate of PTCA and SHOCK. Lombardia and England dataset. 57
4.15 Exponential hazard rate of *renal*, *arrhythmia*, *pulmonary disease* and *hypertension*. Lombardia dataset. 57
4.16 Exponential hazard rate of *renal*, *arrhythmia*, *pulmonary disease* and *hypertension*. England dataset. 58

Abstract

This work deals with the problem of predicting readmissions in Heart Failure patients. The aim is to highlight which factors may be helpful in the prediction process. The use of different statistical methods allows to gain a multiple perspective. However, the real innovation of this work lies in the comparison of two different countries. The available data, indeed, come from administrative databases of Lombardia, an Italian region, and of England, a country of United Kingdom. Each dataset contains in detail demographics, administrative and clinical information, which constitute the history of the patient and, furthermore, the covariates. Using the free available software **R**, we have implemented three different models: a Logistic Regression for both first and second readmission within 30 days; Hurdle and Zero-Inflated models for the total number of readmissions per patient within a year and, finally, two Multi-State models, that shape the process of admission-discharge-death. For each model chosen, we have adjusted in the right way the structure of the dataset. From our analysis interesting results come out: few common comorbidities increase the probability of readmission for Heart Failure in both the dataset (*renal*, *arrhythmia*, *pulmonarydz* and *hypertension*), while others have a different impact depending on the country (*compdiabetes* and *pulmcirc* are influential only in Lombardia, while *alcohol*, *tumor*, *dementia* and *pvd* only in England). The procedures are relevant in different way in both the countries, and they also are influenced by the timing in which considering a readmission. About the implementation, we can assert that the use of different models allows a good balance between computational costs and completeness in describing the scenario, highlighting various features. Logistic regression and Hurdle/Zero-inflated models, indeed, are easy to implement but they don't have a wide perspective of the clinical history of the patient. Multi-State models catch in the best way the process and the impact of covariates on transitions, but present computational limitations.

KEYWORDS: Heart Failure, Predicting Readmissions, Hurdle and Zero-Inflated models, Multi-State models, Comparison between Countries.

Sommario

Questo lavoro tratta il problema di predire le riospedalizzazioni di pazienti affetti da Scompenso Cardiaco. Lo scopo e' cercare di evidenziare quali possano essere i fattori che aiutino nel processo di previsione. L'utilizzo di diversi metodi statistici permette di affrontare il problema da piu' punti di vista. Tuttavia, la vera innovazione portata da questo lavoro risiede nella comparazione di due differenti stati. I dati disponibili, infatti, provengono da database amministrativi della Lombardia, una regione italiana, e dell'Inghilterra, una nazione del Regno Unito. Ogni dataset contiene dettagliate informazioni di tipo demografico, amministrativo e clinico, le quali costituiscono la storia clinica del paziente e, inoltre, le covariate di interesse. Grazie all'uso del software **R**, sono stati implementati tre modelli differenti: un regressione logistica sia per la prima che per la seconda riospedalizzazione entro 30 giorni; i modelli Hurdle e Zero-Inflated, che tengono conto del numero totale di riammissioni in un anno; infine, i modelli Multi-Stato, che modellano l'intero processo di ammissione-dimissione-morte. La struttura del dataset e' stata di volta in volta adattata al modello scelto. Dalle analisi eseguite, emergono interessanti risultati: alcune patologie accrescono la probabilita' di riammissione per Scompenso Cardiaco in entrambi i dataset (*renal*, *arrhythmia*, *pulmonarydz* and *hypertension*), mentre altre diversificano il loro impatto a seconda del paese (*compdiabetes* e *pulmcirc* sono rilevanti solo in Lombardia, mentre *alcohol*, *tumor*, *dementia* e *pvd* solo in Inghilterra). Le procedure mediche, pur differentemente, in entrambi i paesi hanno un'importante impatto, che cambia anche a seconda dell'orizzonte temporale considerato. Per quanto riguarda l'aspetto implementativo, possiamo asserire che l'uso di diversi modelli consente un buon bilanciamento tra i costi computazionali e la completezza nella descrizione dello scenario. La regressione logistica e i modelli di conteggio, ad esempio, sono facilmente implementabili, ma non consentono una visione dettagliata della storia clinica del paziente. I modelli Multi-Stato, invece, sono ottimi nel catturare l'intero processo e l'impatto delle covariate sulle singole transizioni, ma presentano forti limitazioni computazionali, non consentendo di inserire tutte le covariate di interesse come predittori.

Chapter 1

Introduction

Heart Failure (HF) is a chronic disease that occurs when the heart fails to pump sufficiently to maintain blood flow at the right pressure for human needs. It may be caused by many conditions that lead damage to the heart muscle: coronary artery disease, high blood pressure, heart muscle weakness, heart rhythm disturbance, damage with heart's valves or a combination of all these.
Nowadays, Heart Failure is one of most common disease in our society, due to many causes, for example ageing of population. To understand the relevance of this disease, we just point out that Heart Failure is one of the most important cause of hospitalisation in people over 65 (2014).
When dealing with patients affected by chronic disease (like Heart Failure), the matter of predicting readmissions is a real challenge for hospitals, mainly for two related reasons. The first one is concerned with the high costs of hospitalization, so, discovering the reasons of readmission may lead to improve hospital care and, consequently, save money. Much more important could be the second reason: to find which features in patients determine a higher incidence of readmission, in order to improve the therapies and to target interventions. This is twice as useful, as for it takes benefits to patients and to hospitals as well.
Evaluating hospital readmissions and linked quantities for any kind of chronic disease is one of the aims of the statistical research, thanks to the large amount of data collected by hospitals. Several approaches have been applied to different chronic pathologies (see, for example, Bartolomeo et al. (2008), Bottle et al. (2014), Castaeda and Gerrits (2010), Postmus et al. (2012) or Ieva et al. (2015)), because the underlying process is similar. That gives an idea of the interest lifted up by this issue.
This thesis work focuses on the problem of predicting readmissions of patients affected by a specific chronic disease: Heart Failure. We are going to apply different statistical methods to model the process of readmissions of patients affected by this common pathology. However, an important innovation is introduced in this work, that differentiates it from all the previous analysis. The new perspective given in this work is in fact the comparison between two dataset collected by different countries. All the chosen methods will be symmetrically applied to patients coming from Lombardia, an Italian region, and from England, within the United Kingdom.
This two-step analysis is stimulating for many reasons. If we consider the "longitudinal" investigation along different models, we're gaining a multiple perspec-

tive of predicting readmissions. This is useful, because the standard approach in facing this problem is to use logistic regression to predict a single readmission. This approach, however, is incomplete, because patients with chronic diseases can have multiple admissions and the length of stay is not considered in logistic regression. Therefore we want to try other modelling approaches.
Even better, the comparison between different dataset is a tool to understand strength and weakness of the correspondent Health Systems in facing Heart Failure readmissions and to investigate the difference between the two populations as well.
So, the aims of this work are the following: (i) to find out what covariates are good predictors for readmissions of Heart Failure patients, (ii) to compare these outcomes between Lombardia and England, even detecting the basic differences between the populations, (iii) to understand the strength and the weakness of the models adopted, depending on the response of interest.
The structure of the work provides an initial overview of the dataset (Chapter 2), in order to interpret the features of both populations. In detail, we will focus on anthropological and clinical (procedures and comorbidities) quantities of interest, which are our predictors. A second step (Chapter 3) is a systematic review of the statistical methods that will be applied and the adjustment of the dataset to the models, depending on the response of interest. Our attention has been centred on three different models: Logistic regression, Counting models (Hurdle and Zero-Inflated models) and Multi-State models. These tools are described from the theory to the implementation, delineating the necessary passages to reach the suitable structure of the dataset as input for the models. The main core of the work, however, will be the comparison of the results between Lombardia and England dataset (Chapter 4): the different weight of covariates on readmission process and similarities/dissimilarities in the process of admission/discharge will be investigated. These results are further deepened in the conclusive chapter of the work (Chapter 5).
All the statistical models (as well as plots and other useful tools) have been implemented by using **R** software R Core Team (2014), a useful open source statistical software.

Chapter 2

Background and Motivations

In this chapter we present the datasets used for the analysis. They come from two different contexts: the first one collects data from Lombardia, a region in the northern part of Italy, the second one collects data from England, the biggest country of the United Kingdom.
In this case study, we are handling a specific source of data, that now we explain. Our analysis, indeed, take advantage of data coming from administrative databases. Administrative data refers to information collected primarily for administrative (and not research) purposes. They play a central role in the evaluation of health-care systems, due to their diffusion and low cost of information. Although they are collected for administrative purposes, this kind of data are increasingly approved by clinical epidemiologists, and they have already been used in several studies about HF readmissions yet (see, for example, Philbin and DiSalvo (1999)).
At the beginning, we inspect the peculiarities of each dataset in their original structure (Section 2.1 and Section 2.2). We will therefore give a first empirical analysis that helps giving a general overview and a comparison of both the populations (Section 2.3).

2.1 The Dataset from Lombardia

The first dataset used for this works comes from Lombardia, a region counting 9,955 million citizens.
Lombardia Health System is one of the most efficient in Italy, and its accuracy allowed a collection of a complete dataset, that is going now to be illustrated.
The dataset in its original form is part of a bigger dataset of all the patients admitted in a hospital of Lombardia in a period that runs from 1^{st} January 2006 to 31^{st} December 2012, for a total follow-up period of six years. This dataset has been divided in different groups depending on different codification of disease. Among these groups, we have analysed the patients coming from the first group, which are the most likely patients admitted for Heart Failure.
The first structure counts $70,236$ observations of 53 significant variables. Each

row is an admission into hospital for reasons related to heart-conditions, especially Congestive Heart Failure (99.89 % of the admissions), and an *ID* variable keeps track of the identity of the patient, in order to know the total number of patients $(37,565)$, and in order to record the variables of interest.
The important covariates recorded are related to anthropological, administrative and clinical information. The demographic and outcome covariates are: *sex*, *age* of admission, paediatric indicator *ped_ind* (if at the beginning the patient is younger than 18 years old), indicator of death before the end of the follow-up period (*death_ind*) and indicator of death during the current hospitalization (*death_intraH_ind*).
The administrative covariates are related both to the admission schedules and to the identification of the hospital. In the first case we obtain: the dates of admission, discharge, exit from the follow-up, the period between the beginning (or the end) of the admission and the previous (or the following) admission or death. In the second case there are two different covariates: the first is an indicator variable that indicates if the hospital is situated in Lombardia, the second is a label that identifies the hospital. The number of hospitals recorded in the dataset is 505, 202 of which are located in Lombardia and 303 in adjacent counties. It is interesting to note that the percentage of admissions in a hospital not located in Lombardia is 2.1% meaning that, whereas the number of recorded hospitals not located in Lombardia is high, the number of admissions in this hospital is very low and not meaningful. Moreover, the number of hospitals that have been overall visited less than 20 times for Heart Failure is 340. This is a high value and corresponds to the small percentage of hospitals not situated in Lombardia that constitute the small percentage of admissions outside Lombardia.
Clinical information is both related to medical procedures and to comorbidities. We introduce them now, highlighting in brackets their name as covariates in our models, coming from the relative ICD9 code.
A heart failure patient can be submitted to the following procedures : Coronary Artery Bypass Graft ($CABG$), Percutaneous Transluminal Coronary Angioplasty ($PTCA$), Implantable Cardioverter Defibrillator (ICD); he can also be recovered in Intensive Therapy (*ti*) or partly in Rehabilitation (*riab*), and he can also be submitted to intervention in Heart Surgery (*cardiochir*).
Moreover, a series of comorbidities are checked in each admission: metastasis (*metastatic*), congestive heart failure (chf), dementia (*dementia*), renal pathology (*renal*), weight loss (*wtloss*), hemiplegia (*hemiplegia*), alcoholism (*alcohol*), tumour (*tumor*), arrhythmia ($arrhythmia$), pulmonary disease (*pulmonarydz*), coagulopathy (*coagulopathy*), diabetes (*compdiabetes*), anaemia (*anemia*), electrolytes in blood (*electrolytes*), liver disease (*liver*), peripheral vascular disease (*pvd*), psychosis (*psychosis*), pulmonary circulation disease(*pulmcirc*), HIV or aids (*hivaids*) and hypertension (*hypertension*). Once a comorbidity appears in the clinical history of the patient, it stays until the end of the follow-up period. The reason of this convention lies in the fact that all these comorbidities are chronic and difficult to vanish.
All the values above are recorded at each admission, and as giving a picture of the clinical history of the patient once recovered from Heart Failure.
All these significant informations have been arranged depending on the type of model implemented, as described in the following chapter. In this section we want to give some information on the dataset with the aid of graphs and plots,

so as to describe the population that we are going to examine.

2.2 The Dataset from England

The dataset from UK is part of a bigger dataset that collects informations from the NHS (National Health Service), specifically in England rather than the whole United Kingdom.
One of the first differences is related to the dimensions of the available data, because the geographical areas we are considering have different width. However, the problem that we are facing is the same, just on two different scales and, fortunately, this is not a hurdle for our analysis. The original structure of the dataset is close to the Lombardia one, with some exception related to some covariates.
We have $1,410,215$ observations, that means 20 times the Italian rows, and 48 covariates for each row. The total number of patients is $263,775$ (7 times the number from Lombardia). These two informations are meaningful, because they point out that the number of admission, on average, is higher for patients from England than Lombardia. These data are collected from April 2006 to end July 2011. So, the follow-up period is of 5 years.
There are some differences with the Italian dataset, for example there are not paediatric (easily inferable by age), Intensive Therapy and Rehabilitation indicators, while there are three more administrative covariates: the first one is $EMERG$, a flag indicating whether an admission is of emergency or not. This covariate is fundamental as for it permits to distinguish between a planned admission and a not scheduled one. This information is substantial for our research because, of course, there is no interest in predicting a readmission that we already know. When using the models described above, we want to predict the admissions that are not scheduled or planned, thus the emergency one.
The other covariates not present in Lombardia dataset are: Cardiac Resynchronization Therapy (*crt*), a specific treatment for heart failure, and Biventricular Pacemaker *pacing not crt*, which is another medical care. In order to have a homogeneous dataset, we have not considered these covariates (as it was samely carried out with Italian covariates not present in England data).
In the English dataset, not all comorbidities are supposed to be chronic. For instance, anaemia could be recorded in some admissions and then it could disappear as reappearing again later; in that case, the comorbidities have been treated in the same way of Lombardia case: once it enters in the clinical history of the patient, it remains until the end.
In the English dataset, the number of hospitals is lower than in the Italian one: the recorded number of hospitals is 348, and 78 of them have been visited less than 20 times (so we can consider them as small hospitals).

2.3 A Common Preliminary Analysis

Before entering into the specific models that we have implemented, we are now briefly giving an idea about both datasets. Specifically, we give some preliminary descriptive analysis of quantities that are not strictly predictors but that can help to understand the populations.

We start with a summarizing table presenting the percentage referred to the anthropological features, the procedures and the comorbidities. We have kept the original structure of the dataset, so most of the quantities are referred to all rows, without manipulations done to suit the data to our future models.
Table 2.1 only presents some quantities related to anthropological characteristics. We can see that the percentages related to the sex are quite similar, while the English population is younger on average. In Figure 2.1 we can see the distri-

	Lombardia dataset	**England dataset**
Anthropological cov		
Age (years)	77.29 (±11.19)	76.37 (±12.15)
Sex	47.67% Males	50.61% Males
Administrative cov		
Number of admissions	2.09(±1.84)	6.13 (±9.75)
Death Indicator	58.23%	51.01%
Intra Hospital Death Indicator	9.37% (of the whole population)	32.76% (of the whole population)
	16.09% (of the dead patients)	64.22% (of the dead patients)

Table 2.1: Percentage or Mean (SD) for original dataset

bution of the age at first admission in both cases. We can notice similarities and dissimilarities too. The growing side (from early age to 84 years) is similar for both the curves, while the decreasing side is quite different. Although it is steep in both cases, the decreasing side behaves differently: in the England dataset it is smooth, in the Lombardia dataset we have a local maximum around 90 years. A first reason could be the different amount of data. The second reason lies in the not scheduled admissions of Lombardia dataset. Moreover, a small but important part of patients are infant in both the dataset.

Another meaningful information is the In-Hospital Mortality rate. The In-Hospital mortality rate is the number of patients dead during an admission divided for the number of current admissions. We have grouped the death inside the hospital from the 10^{th} admission in both cases, due to the sparsity of data for higher admissions.
In addition, in the case of the English data we have plotted the In-Hospital mortality rate in two cases: using all admissions without distinguishing the emergency one, and, secondary, using only emergency admission. It is evident (Figure 2.2) that, on average, the rate increases in the second case. The trend, however, is similar: increasing until the 5^{th} admission, and then monotonic decreasing is registered. That is curious, because we are expecting that when the admissions become higher, the probability of survival should be lower (hence, the rate of death inside the hospital should increase). But this is not the trend in both cases (also emergency ones), probably related to the fact that the most severely ill patients die early (with few admissions).

The case of the Italian dataset, as previously seen with the age at first admission, has a different behaviour (Figure 2.3). That could be related to the lower quantity of data. The trend is positively linear until the 4^{th} admission, presenting then an increasing (yet more stormy) behaviour until the 9^{th} admission, and, finally, it steeply decreases.
In both cases, the width of confidence intervals increases as much as the number

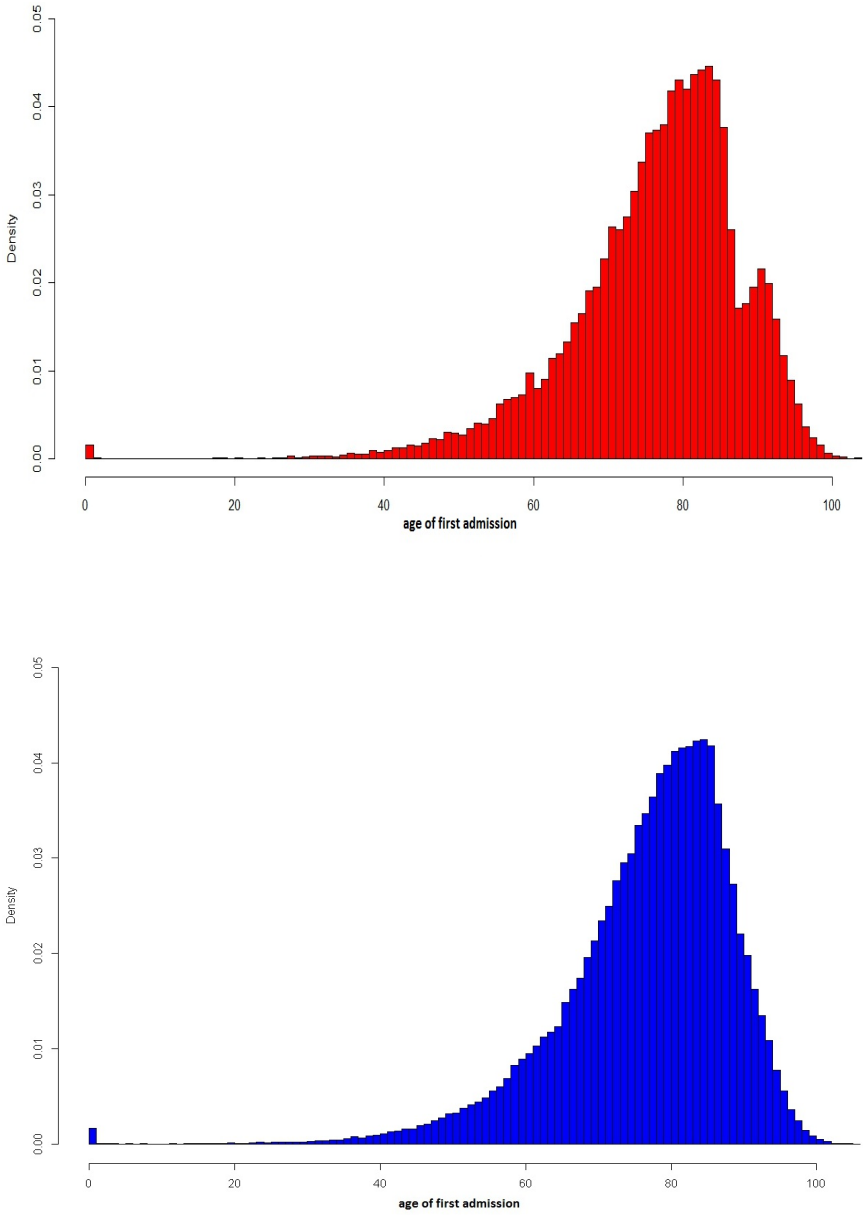

Figure 2.1: Histogram of age at first admission in Lombardia (above) and in England (below) dataset

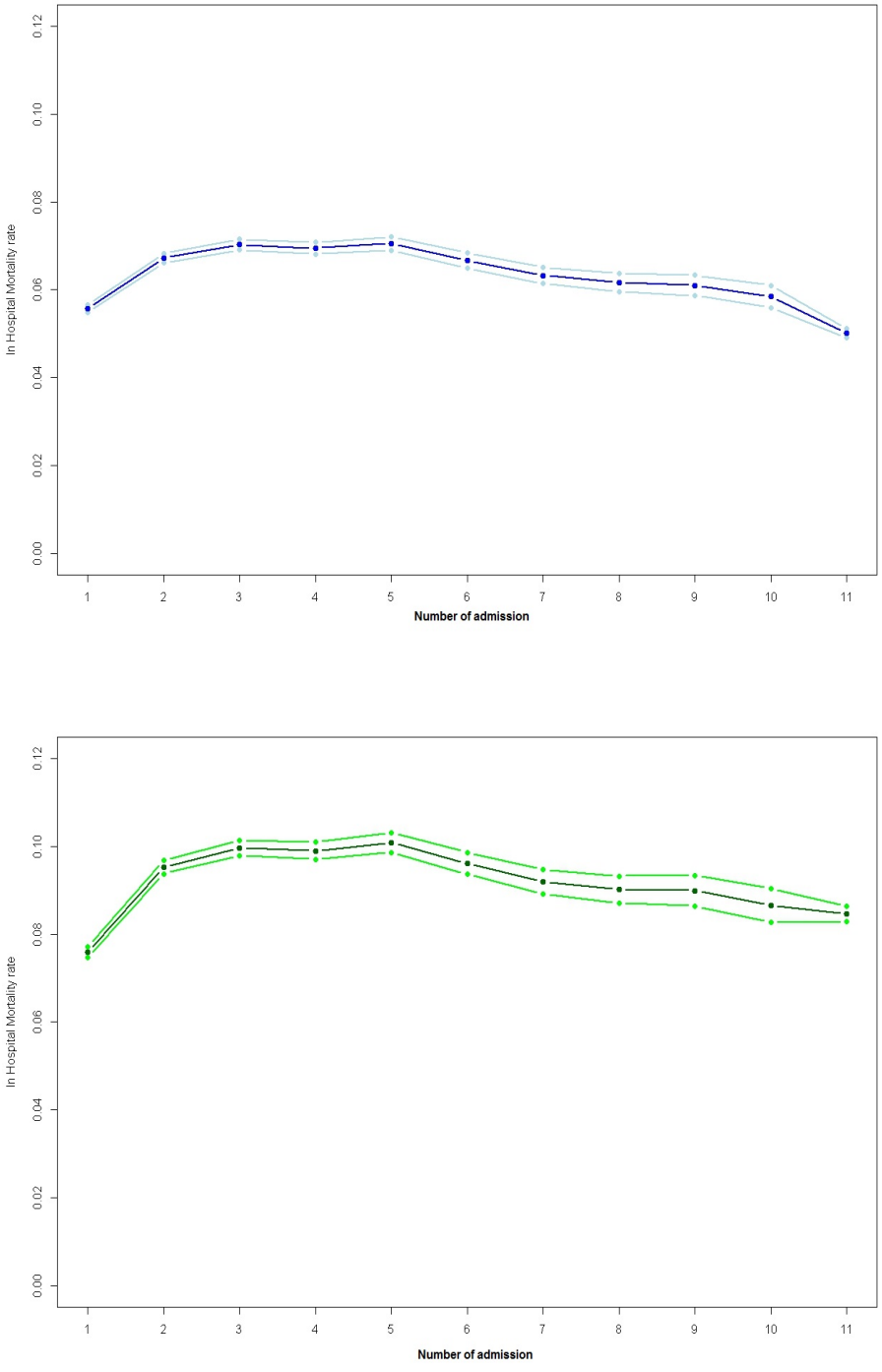

Figure 2.2: In Hospital mortality rate and 95% Confidence Intervals of English patients. All admissions (above), only emergency admissions (below).

of admission becomes higher. That is what we expected, due to the decreasing number of data available for high readmissions. Furthermore, it highlights the difference between England and Lombardia dataset, because the the width of Lombardia Confidence Intervals is higher than England data, which means a less quantity of available data.

After having analysed the features related to administrative informations, we

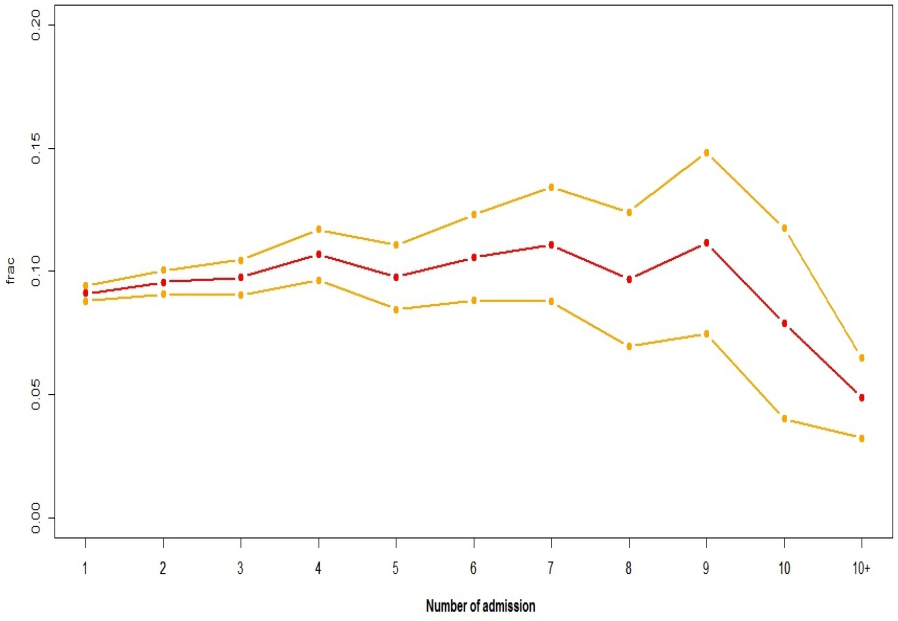

Figure 2.3: In Hospital mortality rate and 95% Confidence Intervals of Lombardia patients.

focus on the properties related to clinical knowledge. So, we watch over the percentage related to comorbidities and procedures. That is because it is interesting and helpful to understand the most frequent pathologies affecting Heart Failure patients and the medical procedures carried out.

Our perspective, now, is much more focused on the problems of readmissions. This is the reason being for the following tables and figures to refer to patients that did not die inside the hospital during their first admission. The informations brought by those patients, indeed, are not useful in a predictive perspective.

Table 2.2 shows the percentage of comorbidities referred to both populations. As we can see, the observed percentages are different, but they share a similar trend: high (low) values for comorbidities for the Lombardia dataset correspond to high (low) values for the England dataset. That is important, as it tells us that the diseases affecting patients are similar in both the countries, whereas sometimes the effect is higher (lower) in a case rather than the other.

The highest percentage are referred to the following disease: *renal*, *arrhythmia*, *pulmonarydz* and *hypertension*. As expected, some of them are strictly related to Heart Failure (*arrhythmia* and *hypertension*), being diseases that strain the regular function of heart, facilitating its damage. The others (*renal* and *pulmonarydz*) are not directly related but still determinant as well. Of course, we can imagine that these covariates are going to impact our results.
Other comorbidities present a different behaviour depending on the dataset. For example, the percentage of *wtloss*, *alcohol*, *electrolytes* and *hypertension* are higher in England dataset than in Lombardia, otherwise *compdiabetes* and *arrhythmia* are higher in Italian dataset. The reason can be found in the different cultural background, but may also be due to coding quality.
Similar is (less than 1%) the behaviour of *metastatic*, *dementia*, *hemiplegia*, *anemia*, *psychosis* and *pulmcirc*. These covariates, moreover, have a low impact on the whole dataset and we are expecting that they don't affect our results substantially.

This is not the only empirical analysis that we can do about comorbidities.

	Lombardia dataset		**England dataset**	
	Percentage	Number of Patients	Percentage	Number of Patients
Metastatic	1.72 %	589	2.3 %	5,742
Dementia	5.08 %	1,738	6.45 %	16,093
Renal	21.81 %	7,453	32.33 %	80,572
Wtloss	0.45 %	154	2.82 %	7,033
Hemiplegia	2.58 %	882	2.35 %	5,866
Alcohol	0.26 %	89	3.53 %	8,802
Tumor	7.31 %	2,498	6.22 %	15,514
Arrhythmia	49.36%	16,857	59.72%	148,817
Pulmonarydz	24.90%	8,505	29.55%	73,624
Coagulopathy	0.71%	245	1.41 %	3,522
Compdiabetes	7.90%	2,699	5.87%	14,618
Anemia	8,88%	3,034	9.61%	23,950
Electrolytes	4.22%	1,444	16.38%	40,815
Liver	4.84%	1,653	2.74%	6,830
Pvd	12.31%	4,205	12.63%	31,479
Psychosis	0.69%	238	0.69%	1,744
Pulmcirc	4.82%	1,649	7.36%	18,352
Hypertension	44.91%	15,335	69.74%	173,759

Table 2.2: Comorbidities percentage referred to patients not dead inside the hospital during the first admission. Lombardia dataset (34,146 patients), England dataset (249,156 patients)

Figure 2.4 and Figure 2.5, for example, show the distribution of different disease along the age of first admission. These graphs are simple yet important as well. *age*, indeed, is a the first available information and knowing the distribution of disease along age of admission could be a handy tool to understand the probability of readmission.
In Lombardia patients the comorbidity *arrhtymia* is one of the most frequent disease at first admission, despite the age, and it often affects the 50% of the population as well as *hypertension* (not reported). The comorbidities *pulmonarydz* and *renal* are less frequent, but they share a similar behaviour: a peak in the early ages. That is probably because of different causes of the Heart Failure in young people compared with elder ones, but it also may be due to the small

number of young patients affected by Heart Failure, which can lead to a misclassification of the results.
English patients (Figure 2.5) behave in the same way as Lombardia ones. The comorbidities that affect more the patients are *arrhtymia* and *hypertension* (not reported), with a percentage around the 60%. As in Italian dataset, the most relevant comorbidities also have a peak in the early ages. But the most interesting disease is represented by *alcohol*, which is much more frequent in youngest people and it is not a merely local peak, but a maximum. This kind of phenomenon is not caught in the Italian data. The percentage referred to procedures are important as well. As we can see in Table 2.3, there is a difference between the percentage of procedure in the two dataset. Indeed, Lombardia patients are, on average, more exposed to medical procedures than English one. In $PTCA$ procedure, especially, the gap between the two datasets is much more evident than in the other procedures. Analysing the impact of the procedures may be interesting, because they may represent possible and easily available predictors (a medical record is sufficient and precise).

	Lombardia dataset		**England dataset**	
	Percentage	Number of Patients	Percentage	Number of Patients
ICD	4.82 %	1,647	2.32%	5,784
CABG	4.11 %	1,405	1.39%	3,486
PTCA	8.16 %	2,788	2.58%	6,439
SHOCK	2.78 %	951	0.31%	786

Table 2.3: Procedures percentage referred to patients discharged alive from the first admission. Lombardia dataset (34,146 patients), England dataset (249,156 patients)

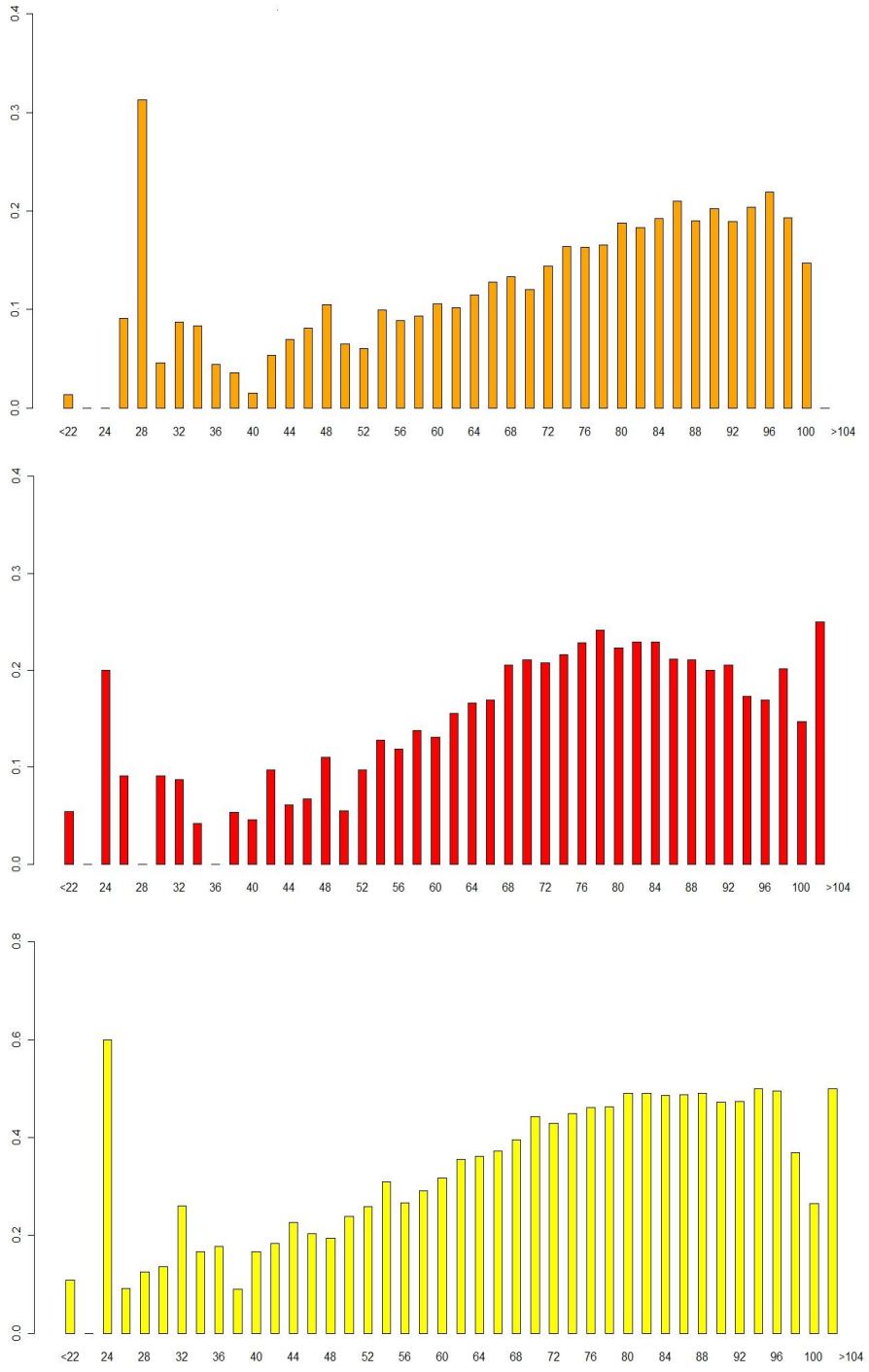

Figure 2.4: Distribution of frequency of *renal* (above), *pulmonarydz* (middle) and *arrhythmia* (below) disease along age of first admission. Lombardia dataset.

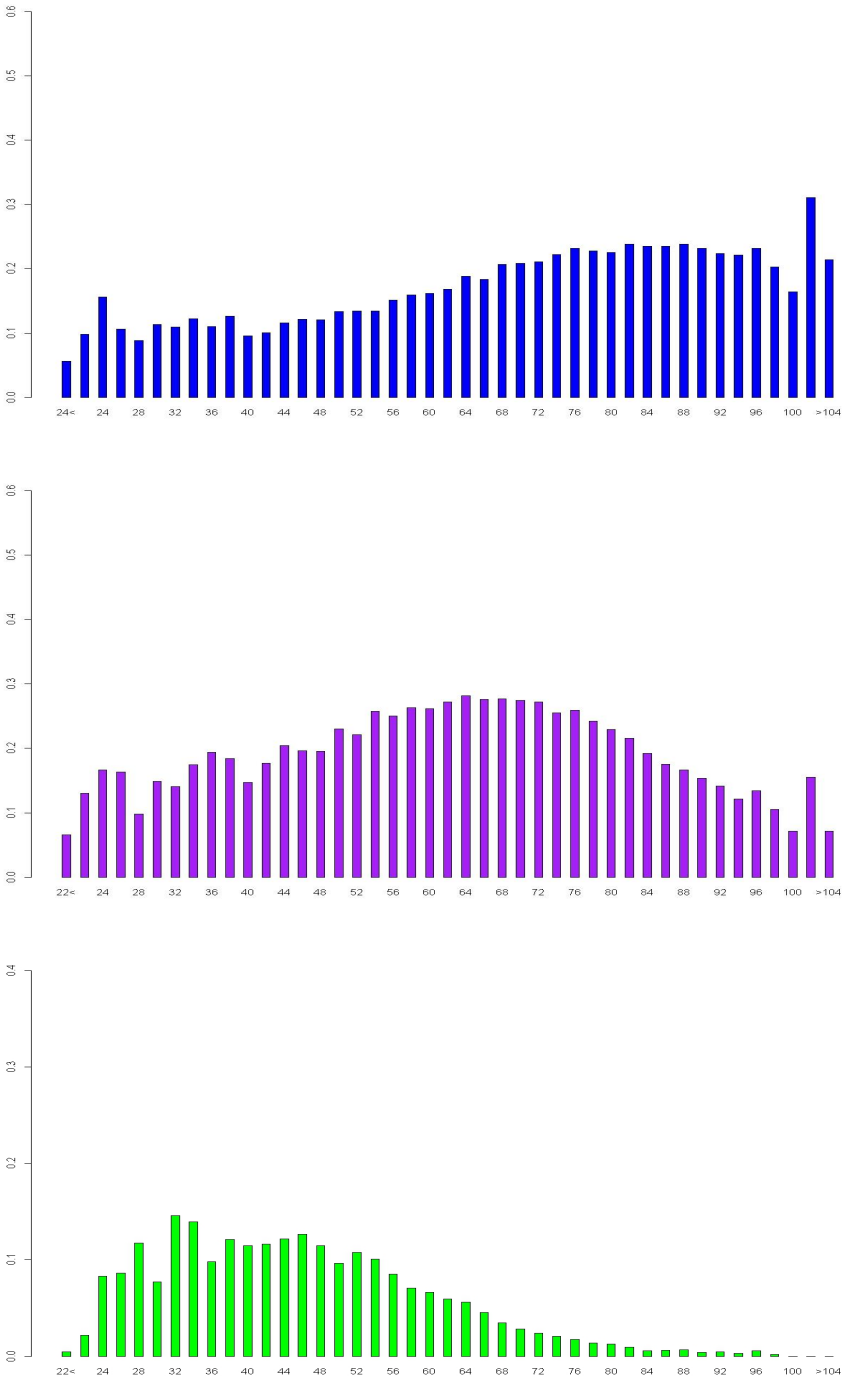

Figure 2.5: Distribution of frequency of *renal* (above) and *pulmonarydz* (middle) and *alcohol* (below) disease along age of first admission. England dataset.

Chapter 3

Chosen Methods: Theory and Implementation

The aims of our research, as mentioned in the introduction, are to compare different statistical methods and, thanks to this, to compare two different datasets. In particular, we have chosen three models: Logistic regression, Counting models (Hurdle and Zero-Inflated models) and Multi-State Models. They go from a simple (Logistic regression) to a complex one (Multi-State model) in terms of theoretical features, outcomes as well as implementation. In Table 3.1, Table 3.2 and Table 3.3, we have given a general synthetic idea about what each model provides and about the statistical bases underlying the implementation.

Since each model requires different inputs and provides different outputs, we have to match the problem of readmissions of Heart Failure patients to the models chosen in the right way.
Thus, this chapter explains why a model has been chosen, how it is related to the response of interest (Heart Failure readmissions) and how it has been implemented. For each method, the first step provided is a theoretical overview, that gives a generic explanation of the statistical model. Then, we move to the comprehension of the problem and the consequent choice of the response variables and of the covariates. The last step will describe the implementation of the model by using **R** software and the arrangement of the dataset to a suitable structure for R-functions, in order to lose the least quantity of information in the adjustment. This latter work has taken long time to run, due to high dimensions and calculations.
A common work has been done before the arrangements needed in each model. In logistic regression and counting models, indeed, we have not considered all

	Logistic Regression
Parameter of interest	p
Regression Relationship	$\log\frac{p}{1-p}=\beta_0+\beta_1 x_1$
Response variable	Readmission within 30 days
Output	OR/coefficients for each covariates
Goodness of fit	Residuals, AIC, LogLikelyhood

Table 3.1: General description of Logistic regression

	Count Models (Hurdle/ZeroInflated Models)
Parameter of interest	p(probability of readmission), λ or θ(Poisson or NegBin parameters)
Regression Relationship	$\log(\mu_i)=x_i\beta_i+\log(1-f_0(0;z_i))-\log(1-f_{count}(0;z_i))$ $\mu_i=\pi_i\cdot 0+(1-\pi_i)\cdot\exp(x_i\beta_i)$
Response variable	Total number of readmissions
Output	OR/coefficients for each covariates
Goodness of fit	Residuals,LogLikelyhood

Table 3.2: General description of Hurdle and Zero-Inflated models

	Multi-State Models
Parameter of interest	$q_{rs}(u)$ hazard rates
Equation	$q_{rs}(u,\boldsymbol{z}(t))=q_{rs}^{(0)}\exp(\beta_{rs}\mathbf{z}(t))$
Response variable	Number of patients in each state-transition
Output	Hazard rates valuation, transition probabilities,expected survival mean sojourn time in each state
Goodness of fit	LogLikelyhood

Table 3.3: General description of Multi-State models

the patients dead inside the hospital during the first admission, excluding them from the entire analysis (14.693 patients from England, 3.420 patients from Lombardia dataset). If this work is simple when dealing with Italian dataset (there is no distinction between emergency admission and planned admission), it becomes different when dealing with the UK dataset, because we have considered each situation differently depending on the model adopted.
Interesting results come out of the implementation of these models, such as an understanding upon how they are linked together. Indeed, it will be appreciable to observe that every model is somehow an evolution of the previous one. This means that we can use the results gained from the simplest model just to go straight to our objective when using more complex ones, in order to save time and attempts. Moreover, all the models inspect similar or different parameters in a different way, going more and more into depth towards a complete interpretation.

3.1 Logistic Regression

3.1.1 Theoretical structure

Logistic regression, a particular case of general linear model, is the simplest model used in this work and, after all, is one of the easiest models to implement. At the basis of logistic regression there is the purpose of finding the parameter p of a population sampled from Binomial distribution, which usually describes a process of binary result, in which the probability of successful event is p and the probability of failure is *1-p*. Of course, the primary way to find this parameter of interest is the Maximum Likelihood Method, in which p is approximated to the total number of success on total number of attempts. Consequently, when a series of covariates is given, the main interest is to find in which way the

variables are influential on the process.
The regression method is based on the *logit* regression model described below:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} \tag{3.1}$$

where p_i is the probability of success for the subject i, β_j (j: $1 \cdots$ N) are the coefficients indicating how the covariates x_{ji} of the subject i are influential on the *logit* ratio. The aim of Logistic regression is in fact to find out the leverage β_j of each covariate x_j on the underlying probability of success. The p, instead, is easily found thanks to the Maximum Likelihood Method, as mentioned above.

3.1.2 Adaptation to the problem

As mentioned above, Logistic regression is useful when we want to predict the probability of success in a Binomial process. In our case, the event of interest is the readmission in the hospital within a fixed period.
Through R software, logistic regression is easy to implement thanks to `glm()` function, which needs the following input: a vector of binary variables (1 if the event recorded is a success, 0 otherwise) and the corresponding list of covariates related to the event. As output, we obtain several interesting quantities: the estimate of the parameter of interest, the list of coefficients estimates and their significance, the Odds Ratio for each coefficient and its Confidence Interval, the goodness of fit (through LogLikelihood, AIC and Residual Standard Error). In that way, we can obtain important information about our population coming from a Binomial distribution: the estimate of the probability of success p, the impact and the consistency of each factor on the response.
In our case, the response variable Y_{all} is the indicator that takes value 1 when a patient has been readmitted within 30 days from the previous discharge, 0 otherwise. The choice of 30 days as time limit for readmission is typical and verified by literature (for example, see Bottle et al. (2014)).
In the case of Italian dataset, we have considered all re-admissions available in the dataset, while in the UK dataset we have considered as valid the emergency readmission within 30 days, because in both cases we are focusing on unplanned readmissions.
In both implementations, the predictors are constituted by the anthropological, administrative and clinical covariates described above (the ones in common).
The purpose of logistic regression, in that case, is finding out the probability of readmission and the meaningful covariates that may influence its increase or decrease, thanks to the Odds Ratio or the estimate of coefficients.
We have implemented two different logistic regression: the first includes only the first readmission, while the second one considers only the second readmission. We have chosen this strategy because the covariates relations might change during the progression of the patient's Heart Failure. We therefore want to investigate which factors may be related to the progression of the disease and, consequently, to multiple readmissions. In Table 3.4 a summary about the dimensions of the problem is given (depending on the different regression and on the different dataset).
These different approaches have of course multiple effects: the dataset dimension lessens, the coefficients and the OR estimates change. This is related to

	Lombardia dataset		**England dataset**	
	N. of patients	N. and % of readm within 30 days	N. of patients	N. and % of readm within 30 days
First readm.	34,146	2,205 (6.46%)	249,077	31,164 (12.51%)
Second readm.	15,412	1,186 (7,69%)	223,774	28.339 (12.66%)

Table 3.4: Dimension of the dataset and percentage of readmission within 30 days in Logistic regression. Lombardia and England dataset.

the phenomenon of readmissions: multiple admissions are in fact less frequent. Furthermore, the distinction between first and second readmission may highlight the difference between the consequences of a good/bad primary care from the problems rose once the disease has become chronic.
Of course, in the case of Logistic regressions we have reorganized the dataset to make it suitable for each implementation. For each admission, we have calculated the number of days until the next readmission, and we have create the Y_{all} indicator variable (1 if the following admission is within 30 days, 0 otherwise). To each row we have associated the related covariates. Our aim is to find the role of covariates at the moment of discharge that may help to predict a early readmission.
In the Figure 3.1 we show how the dataset has been transformed. Before il-

Y_all	ID	age	sex	adm_number	year_discharge	dateADM	dateDISCHARGE	dateDEATH	dateOUT	DEATH_ind
0	1	79	0	1	2006	2006-07-14	2006-07-25	2010-06-18	2010-06-18	1
0	2	76	1	1	2006	2006-06-28	2006-07-17	2010-08-13	2010-08-13	1
0	3	72	1	1	2006	2006-03-15	2006-03-21	NA	2012-12-31	0
0	4	72	0	1	2006	2006-02-10	2006-02-19	2006-03-28	2006-03-28	1
0	5	78	0	1	2006	2006-03-10	2006-03-27	2010-10-15	2010-10-15	1

Y_all	ID	age	sex	adm_number	year_discharge	dateADM	dateDISCHARGE	dateDEATH	dateOUT	DEATH_ind
0	1	80	0	2	2006	2006-09-01	2006-09-15	2010-06-18	2010-06-18	1
0	5	78	0	2	2006	2006-07-03	2006-07-11	2010-10-15	2010-10-15	1

Figure 3.1: Dataset for Logistic Regression: first readmission and second readmission.

lustrating the results, we have to show two important graphics useful for the model we are going to implement. We observe that most of our covariates are indicators. So, when getting the linear coefficients, the interpretation of the values becomes very simple. In fact, the presence/absence of the covariate gives directly a positive/negative contribution. When dealing with continuous covariates, on the other hand, this interpretation is weaker. We should verify the empirical trend of the readmission rate along the continuous covariates (it is great if the behaviour is linear).
The only continuous variables that we are going to use are: *age* and *LOS* (Length of Stay). In Figure 3.2 we can see the enlarged trend of the Readmission rate in both dataset. In the case of Lombardia we have scattered tails (in early age and long LOS), while in the UK dataset this behaviour is less emphasized, but already visible. The reason could be connected to the lower quantity

of data, especially in the extreme values. Despite this, we can imagine a linear behaviour, much more evident when considering all the range from 0 to 1. In this latter case, the linear trend is much more evident: quite constant for the Lombardia dataset, decreasing for the England dataset. Now we can be ensured that these two covariates may be predictors in the Logistic Regression model.

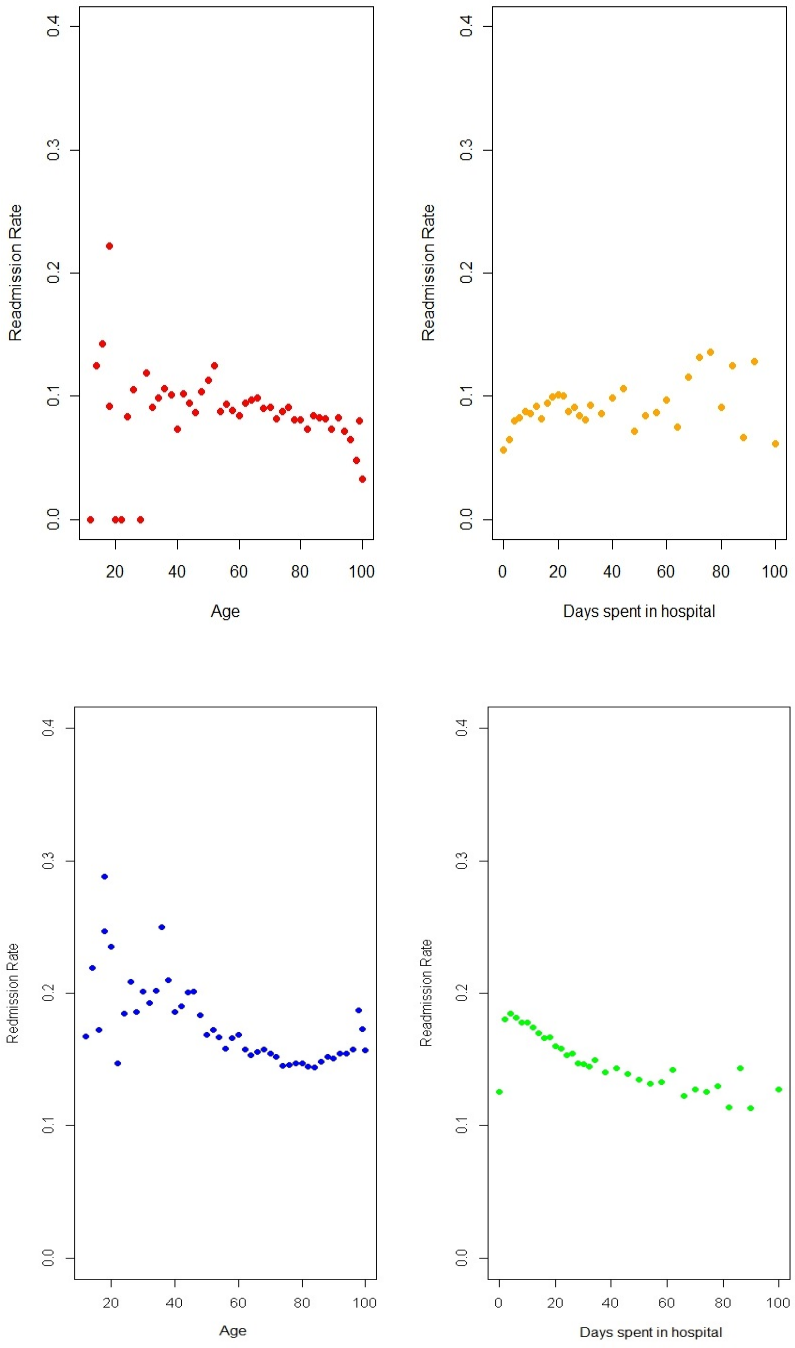

Figure 3.2: Readmission Rates within 30 days along *Age* and *LOS*. Lombardia (above) and England (below) dataset. Limits of Y axes: from 0 to 0.4.

3.2 Hurdle and Zero-Inflated Models

Logistic regression forecasts success probability in a population sampled from a Binomial distribution. Sometimes, instead, it could be interesting to analyse a population that deals with counting response, such as the total number of events in a determinate period of time. In the problem of readmissions of Heart Failure patients, for example, it is useful to enlarge the perspective on the process and to find out the factors that may influence multiple readmissions. Furthermore, counting all readmissions is a better estimate than a single readmission of the burden to both the patient and the economy. This is the reason being for us why we need to introduce counting models to inference on readmissions of Heart Failure patients. The second step of this work, indeed, is to determine the factors that bring patients to be repeatedly readmitted for Heart Failure.
Nevertheless, the usual models of regression for counting data suppose a Poisson distribution or, in the case of heavy tails provoked by higher dispersion, a Negative-Binomial distribution. However, in particular circumstances, counting data can present a high quantity of zeros that a traditional counting distribution can't fit with high precision. In problems concerning health and clinical data, as in this case, this is a frequent issue, but fortunately there are several ways to solve it (see, for example, Hu et al. (2011), Atkins et al. (March 2013) and Buu et al. (2012)). Hurdle and Zero-Inflated models try to deal with this question, as they differentiate the part associated with the zeros from the counting part. This allows to have a not-misrepresented outcome.
Before specifying the main characteristics of the model, we can hazard an hypothesis on the interpretation of Hurdle and Zero-Inflated models: they are a natural evolution of the Logistic Regression of a counting process, especially for the first readmission. We can say so, because Hurdle and Zero-Inflated include a Binomial Process (counting part is like a successful event versus zero) but they also give details on the counting part, specifying the different integer values.
In the subsections below, we will give the theoretical structure of Hurdle and Zero-Inflated models and the consequent adaptation to the problem of readmissions.

3.2.1 Theoretical structure

Hurdle and Zero-Inflated models deal with counting data with an ex excessive number of zeros. However, they try to solve this problem differently. To clarify the differences, we will now explain the main features, the equation and the regression method to implement them.
Hurdle models are two-component models: a truncated count component, such as Poisson, Geometric or Negative Binomial, is employed for positive counts, whereas a hurdle component models zero vs. larger counts, for which a binomial model is usually employed.
The Likelihood can be interpreted in this way: a count data model f_{count}(y;x,β), which is left truncated at y=1, and a zero-hurdle model f_{zero}(y;z,γ), which is right censored at y=1:

$$f_{Hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{zero}(0; z, \gamma) \\ (1 - f_{zero}(0; z, \gamma)) \cdot f_{count}(y; x, \beta)/(1 - f_{count}(0; x, \beta) \end{cases} \tag{3.2}$$

All the parameters of the model (β,γ and even θ if a Negative Binomial distribution is used for counting part) are estimated by Maximum Likelyhood. The implementation uses the mean of regression relationship as below:

$$log(\mu_i) = x_i \top \beta + log(1 - f_0(0; z_i)) - log(1 - f_{count}(0; z_i)) \tag{3.3}$$

Hurdle models have the opportunity to distinguish the covariates of the Hurdle component (z_i) from the ones of the counting component (x_i). Of course, they can also be the same.
Zero-Inflated models use a different approach to model the supposed distribution. They are two-component mixture models combining a point mass at zero with a count distribution such as Poisson, Geometric or Negative Binomial. Thus, there are two sources of zeros: zeros may come from both the point mass and from the count component. For modelling the unobserved state (zero vs. count), a binary model is used, potentially containing regressors.
The Likelihood can be written in the following way:

$$f_{ZeroInflated} = f_{zero}(0; z, \gamma) \cdot I_0(y) + (1 - f_{zero}(0; z, \gamma)) \cdot f_{count}(y; x, \beta) \tag{3.4}$$

where the probability of observing a zero from $f_{count}(0;x,\beta)$ is inflated by $f_{zero}(0;z,\gamma)$.
The corresponding regression equation for the mean as follows:

$$\mu_i = \pi_i \cdot 0 + (1 - \pi_i) \cdot exp(x_i\beta) \tag{3.5}$$

where π_i corresponds to the estimate for $f_{zero}(0;z,\gamma)$.
In Zero-Inflated model, the covariates used for the estimation of the parameters of interest (β,γ and θ in the case of Negative-Binomial distribution for counting part) could be different for the Zero-Inflation part and for the counting part.

3.2.2 Adaptation to the problem

Logistic regression forecasts the probability of readmission within a fixed period (30 days in our context), and discovers the most important covariates that influence it. For hospital, as mentioned above, it could also be useful to find out which covariates may influence the total number of readmission, in order to supervise, provide for and plan a better coordination between hospital and primary care. For these reasons, we have considered this different class of models, being it suitable for this purpose and helping to compare the results between Lombardia and England datasets.
In order to give to **R** function the right input, we have adapted the dataset for pursuing in the right way our purpose. In that case, indeed, each row must represent the whole clinical history of the patient: the response variable Y is the total number of readmission and the relative covariates are a summary of the covariates recorded in each admission.
The real challenge to build this new structure is losing the minimum of information when shrinking the dataset. A substantial precaution that we have used has been considering only one year of follow-up since the patient enters the study. The reason for this to be useful is that we don't penalize the results from patients arrived at the end of the study. Of course, we have to differentiate the

ID	age	sex	adm_number	DEATH_ind	DEATH_intraH_ind	ped	ti	ICD	CABG	PTCA	SHOCK
1	79	0	1	1	0	0	0	0	0	0	0
1	80	0	2	1	0	0	0	0	0	0	0
2	76	1	1	1	0	0	0	0	0	0	0
3	72	1	1	0	0	0	0	0	0	0	0
4	72	0	1	1	0	0	1	0	0	0	0
5	78	0	1	1	0	0	0	0	0	0	0
5	78	0	2	1	0	0	0	0	0	0	0
5	79	0	3	1	0	0	0	0	0	0	0

ID_pat	n_adm	ind_sex	age_first	ind_DEATH_intraH	ped_ind	INTRAH_days	n_ti	n_cardiochir	n_ICD	n_CABG	n_PTCA
1	2	0	79	0	0	25	0	0	0	0	0
2	1	1	76	0	0	19	0	0	0	0	0
3	1	1	72	0	0	6	0	0	0	0	0
4	1	0	72	0	0	9	1	0	0	0	0
5	3	0	78	0	0	38	0	0	0	0	0

Figure 3.3: Dataset for Hurdle and Zero-Inflated Models before (above) and after (below) the shrinkage.

two datasets: in the case of the one of Lombardia, we have considered all the admissions within a year since the first admission; in the English one, on the other hand, we have considered the emergency readmissions within a year since the first emergency admission. Moreover, we don't extend our time-window for a too long period, so we can stay close to the Logistic regression.
As sketched above, we are now going to explain how we have summarized the covariates of interest in order to lose the minimum of information. Some covariates are easy to summarize: for example, sex is a constant variable, age has been summarized in two different way, taking into account the age both of the first admission and of the last admission (a mean of the age was senseless from a clinical point of view). In each admission a set of medical procedures may be done and recorded; in that case we have summarized this indicator variable with the total number of taken procedures (one for each type). Comorbidities has been treated in a different way; in fact, once a comorbidity appear in the clinical history of the patient, it is reasonable that it stays since that moment. So, in that case, we have summarized comorbidities with an indicator variable: 1 if the comorbidity is present almost once in the clinical history of the patient, 0 otherwise. Others covariates, like Length Of Stay (inside or outside the hospital) in each admission have been summarized with the Total Length Of Stay along the follow-up of the patient.
After the shrinkage, the length of the dataset has changed: 34,146 rows for Lombardia dataset, 238,482 for the England dataset. We can see an extract in Figure 3.3.

After the shrinkage, we can give to **R** the right input to complete our analysis. **R** software, indeed, provides two specific function to implement Hurdle and Zero-Inflated models (`hurdle()` and `zeroinfl()`) in the **pscl** package (for further details, see Zeileis et al. (2008)). Both functions take as input the counting variable of response, the corresponding covariates, the family of the counting part (*Poisson*,*Negative Binomial* or *Geometric*) and gives as output the estimates of the regression coefficients, their significance, the estimates of the parameters for the counting distribution and, of course, the values to evalu-

ate the goodness of fit (Residuals, LogLikelihood). For further details and code, see Chapter 5.

3.3 Multi-State Models

Logistic regression and Hurdle and Zero-Inflated Models can forecast the probability of readmission, the main factors that can influence it and, also, can tell about the total number of readmissions. But a limitation of these model is the non-distinguishing the timing of different readmissions. Consequently, they don't distinguish the influence of each factor in a specific transition, because they give an overall estimate. However the timing of transitions is a very precious information for an hospital.
Fortunately, a specific class of model like Multi-State Models can help to reach this aim. Multi-State models are helpful to gain a wider view of the process of readmissions of Heart failure patients. They much more complicated than the previous ones but, on the other hand, much more complete.
Multi-state models are statistical tools, useful in describing a stochastic process, in which a subject at any time occupies one of few possible states. These models are frequent in medicine, especially in chronic diseases where the states can describe the patients' conditions (healthy, ill or dead) and where it is also possible to observe the event time between different states.
This class of models is widely treated in literature; Hougaard (1999) has described six special cases as standard structures: the mortality model (states: dead or alive), the competing risks model (for multiple cause of death), the disability model (for irreversible disease), the bivariate model (for bivariate failure times), recurrent events (suitable for describing reproductive life history of a woman) and, lastly, the alternating model (for reversible disease). All these models can be suitable for modelling different clinical scenarios, including our problem related to readmissions of patients affected by chronic disease.
Furthermore, they can help much more in discovering the differences between Lombardia and England datasets. We are gaining now, indeed, much more information about the process itself, about the role of comorbidities/procedures on the transitions. This are features that can highlight similarities/dissimilarities between the two countries considered.
As in the previous sections, in the following subsections we give an overview of the underlying theoretical structure and of the adaptation to our issue, comprehensive of the implementation by **R** software.

3.3.1 Theoretical structure

Multi-State models are suitable when analysing stochastic processes, where a subject can move among several state in subsequent times. In medicine, for example, the available data are often *panel* data, meaning that the change of state as well as the clinical information are collected in exact times. This data collection is suitable for being enquired by Multi-State models. Dealing with *panel* data brings consequence as follows: the state $S_i(t)$ of each individual (a readmission, for example) i=$1 \cdots m$ is only known at a finite series of times t=$(t_{i1}, \cdots t_{in_i})$ and not continuously in time. Another important feature of *panel* data is that they depend on Markov assumption that future evolution depends on the current state and time, but not on the whole history. Multi-State models are suitable for wider classes of stochastic process, not only for the need one of *panel* data. However, in this theoretical overview, we focus on this particular

scenario (*panel* data and Markov assumption).
First of all, we have to suppose that an individual can move in a set of discrete states 1,$\cdots$,R. We also consider the changes of state in continuous time ($S(t)$ is the state occupied at time t). The movement on the discrete state space is governed by *transition intensities* $q_{rs}(t,\boldsymbol{z(t)})$, where r and s belong to the state space 1,$\cdots$,R. The *transition intensities* may depend on time t and on a set of individual and time-dependent variables $\boldsymbol{z(t)}$ as well. Each intensity represents the instantaneous risk of moving from state r to state s ($r \neq s$):

$$q_{rs}(t, \boldsymbol{z(t)}) = \lim_{\delta t \to 0} P(S(t+\delta t) = s | S(t) = r)/\delta t.$$

The *transition intensities* form a $R \times R$ matrix Q whose rows sum to zero (that is, the diagonal entries are exactly q_{rr} = - $\sum_{s \neq r} q_{rs}$).
Once we know the intensities, we gain the *probability of transition* under Markov assumption, which are more handy to use:

$$P_{rs}(u,t) = Pr(S(t) = s | S(u) = r)$$

where $u \leq t$.
There is also another interesting extension of the classic Markov model, called semi-Markov process, where the *transition intensities* do not depend on the current time, but only on the duration in the current state. Of course, once the *transition intensities* are known then all the quantities of interest can be obtained: length of stay in a state, visiting frequencies and so on.
The last peculiarity of Multi-State models lies in the possibility of introducing the role of covariates in the *transition intensities*. The effect of explanatory variables on the rates of transition, indeed, is introduced by using a proportional intensities model. Consequently, the intensity matrix *Q(t,**z(t)**)* depends also on the covariate vector $\mathbf{z(t)}$. For each entry of *Q(t,**z(t)**)* , the transition intensity for patient i at observation time t becomes:

$$q_{rs}(z_{it}) = q_{rs}^{(0)} exp(\beta_{rs}^T z_{it})$$

3.3.2 Adaptation to the problem

Multi-State models are versatile and, on the other hand, simple to use. They can be implemented in different way, depending on the kind of research that we want to do and on the complexity. In this work, for example, we have tried to implement two different models: a basic model and a more complicated one. Now, we are going to describe the chosen models and the relative implementation.
The first fitted model is described in Figure 3.4. We consider only three possible states for the patient: alive inside the hospital (State 1), alive outside the hospital (State 2) and dead (State 3). In this case we are not distinguishing between different readmission, we just consider each admission in hospital and the relative discharge or death. The possible transitions are stated by the arrows in Figure 3.4.

All the adaptations, of course, are made in order to give the right input to **R**. Each admission has been considered an entrance in State 1, and the set of related covariates (comorbidities and medical procedures) has been recorded. If the patient is discharged alive, he passes in State 2 until the next readmission.

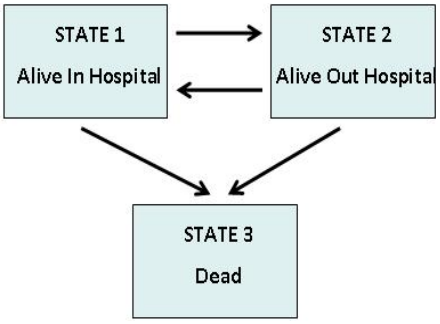

Figure 3.4: Multistate Model with three states.

While the patient is alive outside the hospital, the set of related covariates is set on the values of the previous admission. This is reasonable, because this operation gives continuity to the clinical history of the patient and because we can't forget that, once a set of comorbidities is recorded, we can't forget them in the history of the patient. From one of these two states, a patient can go to the absorbing state State 3, that represent the death.
Adapting the dataset for Multi-State models requests some necessary operation: the transitions lasting zero days have been adjusted adding a small quantity (like 0.5) that indicates that the length is less than one day; if length equal to zero is left, it can cause problems to `msm()`function. For the England dataset, moreover, we have considered only the emergency admissions. Differently from Hurdle and Zero-Inflated models, we have not imposed a limit to the follow-up period and we have included patients dead at first admission.
Of course, considering all patients and all possible transitions, the main consequence (for a computational evaluation) is the lengthening of the dataset (155,971 rows for Lombardia dataset and 1,659,717 rows for England dataset). In Figure 3.5, an extract of the dataset before and after the adjustment.

A spontaneous evolution of the first model (Figure 3.4) is another one in which the number of admission is specified and not generic (Figure 3.6). This is an interesting evolution, because in that case each transition between different admission is specified, and it is not considered a general one. Furthermore, this structure has been already tested, for example in the work done by Ieva et al. (2015).
In order to have a parallelism with the Logistic regression, the Multi-State model with more then three states will be structured differently. That is because we are specifically interested in the first three admissions and the relative transition from these states to death or to the discharge. A problem arises when dealing with patients admitted more than three times. To solve this complication, we have considered a fourth state that collects all the admission higher than the third. The transition from the fourth state to death as been included as well.

So, the final structure considers the following states: first, second third admission (*"1"*, *"2"* and *"3"*); first, second and third discharge (*"1a"*, *"2a"* and *"3a"*), further admissions (*"4"*) and, at last, "Death". The allowed transitions are shown in Figure 4.7. Of course, the transition between a discharge to a further readmission is equivalent to the visit in the state "Alive Outside the

ID	age	adm_number	dateADM	dateDISCHARGE	DEATH_ind	DEATH_intraH_ind	ICD	CABG	PTCA	SHOCK	metastatic	chf	dementia
1	79	1	2006-07-14	2006-07-25	1	0	0	0	0	0	0	1	0
1	80	2	2006-09-01	2006-09-15	1	0	0	0	0	0	0	1	0
2	76	1	2006-06-28	2006-07-17	1	0	0	0	0	0	0	1	1
3	72	1	2006-03-15	2006-03-21	0	0	0	0	0	0	0	1	0
4	72	1	2006-02-10	2006-02-19	1	0	0	0	0	0	0	1	0

ID_paz	state	days	ICD_msm	CABG_msm	PTCA_msm	SHOCK_msm	metastatic_msm	dementia_msm	renal_msm	wtloss_msm	hemiplegia_msm
1	1	0	0	0	0	0	0	0	0	0	0
1	2	11	0	0	0	0	0	0	0	0	0
1	1	49	0	0	0	0	0	0	0	0	0
1	2	63	0	0	0	0	0	0	0	0	0
1	3	1435	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0	0	0
2	2	19	0	0	0	0	0	1	0	0	0
2	3	1507	0	0	0	0	0	1	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0
3	2	6	0	0	0	0	0	0	0	0	0

Figure 3.5: Dataset for Multi-State Models before (above) and after (below) the adjustment.

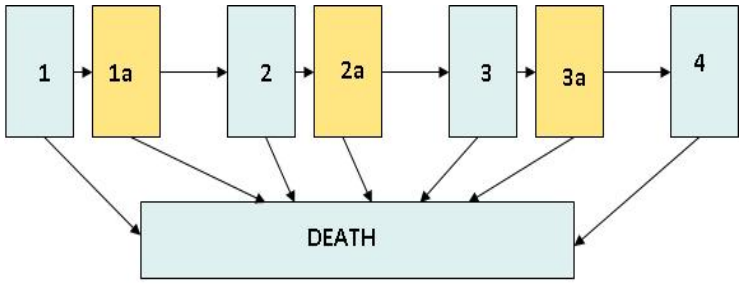

Figure 3.6: Multi-State Model with specification of admissions/discharge.

hospital".
Multi-State Models and the relative implementation by **R** have been widely treated in literature, often connected with Competing Risks (see, for example, Beyersmann et al. (2012)) or related to survival analysis (see, for example, Willekens (2014)). These books deal with Multi-State models and their implementation thanks to multiple packages (`survival()`, `mstate()` and others).
In this work, due to our data, we have used the **msm** package, well explained in useful manual edited Jackson (2014).
In **msm** package, we can analyse a set of panel data in order to enquire the underlying law of the Markov process. Data are supplied as a series of observations, grouped by patient. This should be a dataframe with variables indicating the observed state of the process and the time of the observation. If the data come from more than one individual, then a subject identification variable must also be supplied. It is really important that the observations from the same subject must be adjacent in the dataset, and observations must be ordered by time within subjects.

In Chapter 5, the code for implementation is supplied. Furthermore, **msm** package provides a set of useful functions to describe features of Markov process: intensity matrices, transition probability matrices, mean sojourn times, probability that each state is next, total length of stay, expected number of visits and others.

Chapter 4

Analysis of the Results

In this chapter, we present the results collected by all our analysis. In each subsection we will give a systematic comparison between Lombardia and England datasets, according to the chosen method. Moreover, a comparison between the selected method will be provided. We start from the simplest model, Logistic regression, and we conclude with Multi-State models, passing through Hurdle and Zero-Inflated models.

4.1 Results of Logistic Regression

The implementation of logistic regression is interesting when comparing crosswise different results. We can compare the output of two different models (first readmission, second readmission) on the same dataset. We can also compare the result of the same model on different datasets. All models have been implemented with the same covariates, because, otherwise, the comparison is hard and misrepresented.
Our aim is trying to investigate, in addition to the impact of each covariate, if there are comorbidities that mostly affect a specific readmission rather than a following/previous one.
In the tables below (Table 4.1 and Table 4.2), we give the compared results of logistic regressions implemented: first readmission and second readmission. As we have specified in Chapter 3.1.2, the dimensions of the dataset are different, but this is not the only reason that produces different results. Firstly, we watch out the relevance of the coefficients and then we compare their influence (positive/negative) thanks to the odds ratio.
We start analysing each country separately, that means that we give a look to Lombardia features and then to the England ones.
In the Lombardia dataset, the importance given to covariates decreases along the models. In first readmission, indeed, we find out 9 meaningful variables, while in second readmission only 5 covariates still remain. We expected this result because, of course, data available lessen from the first to the second model. We can observe that the covariate *renal* is relevant in all models as well as the *age*. There is an evident discrepancy between the leftover relevant comorbidities: in the first readmissions, *coagulopathy*, *compdiabetes*, *liver* and *pulmcirc* are significant. That is no longer valid for second readmission model, where,

apart form *renal* disease, the only good predictor for readmission seems to be *hemiplegia*. That behaviour may highlight the different impact of comorbidities along the progression of heart damaging. Procedures behave differently, because they have a much more decisive impact: $CABG$, $PTCA$ and $SHOCK$, indeed, may influence the readmission 30 days since the first discharge, while $SHOCK$ and ICD are more incisive in the second readmission.
In the England dataset the trend of the results is different, because we can notice that much more covariates are significant in both the regression models. In the first model, we record 18 relevant variables (the double of Lombardia situation), while in the second one we record 17 relevant covariates. This different behaviour could be due to the gap of the amount of data, but this is only an hypothesis. In the England dataset, some covariates are significant for both the models, while other change their impact. Among comorbidities, *renal*, *wtloss*, *hemiplegia*, *alcohol*, *tumor*, *arrhythmia*, *pulmonarydz*, *pvd* and *hypertension* still remain in both models, as well as *age*, LOS, ICD, $CABG$ and $PTCA$. Some of them are important only in first readmission model (*anemia* and *electrolytes*), others in second readmission model (*psychosis* and *pulmcirc*). In the England dataset, as well as in the Lombardia one, the only covariate never relevant is *hivaids*.
The quantity of relevant covariates underline a dissimilarity among between the two countries. Much more interesting, however, is the comparison of the weight of each covariate on the response of interest (the readmission within 30 days). That can clarify much more the similarities/dissimilarities between Lombardia and England.

	Lombardia dataset		England dataset	
	OR (CI)	**p-value**	**OR (CI)**	**p-value**
sex (male)	1.14(1.04-1.26)	0.004(**)	0.99(0.96-1.01)	0.435
age	0.99(0.98-0.99)	0.001(**)	0.99(0.99-0.99)	<0.001(***)
LOS	1.00(1.00-1.00)	<0.001(***)	1.00(1.00-1.00)	<0.001(***)
Procedures				
ICD	0.83(0.57-1.20)	0.326	0.52(0.39-0.68)	<0.001(***)
CABG	0.64(0.48-0.85)	0.002 (**)	1.84(1.40-2.48)	<0.001(***)
PTCA	1.16(0.98-1.38)	0.076 (.)	3.19(2.84-3.59)	<0.001(***)
SHOCK	1.41(0.96-2.08)	0.081 (.)	1.66(1.14-2.41)	0.007(**)
Comorbidities				
Metastatic	1.04(0.69-1.57)	0.843	1.17(0.98-1.40)	0.083(.)
Dementia	1.02(0.80-1.30)	0.878	1.72(1.60-1.84)	<0.001(***)
Renal	1.49(1.33-1.68)	<0.001 (***)	1.33(1.28-1.39)	<0.001(***)
Wtloss	0.66(0.24-1.82)	0.423	1.05(0.90-1.22)	0.540
Hemiplegia	1.14(0.86-1.52)	0.356	1.13(0.98-1.30)	0.087 (.)
Alcohol	0.57(0.18-1.82)	0.341	1.21(1.09-1.32)	0.003 (**)
Tumor	0.97(0.79-1.18)	0.736	1.16(1.05-1.29)	0.004 (**)
Arrhythmia	1.01(0.92-1.10)	0.867	1.15(1.12-1.18)	<0.001(***)
Pulmonarydz	1.07(0.95-1.20)	0.240	1.15(1.11-1.19)	<0.001(***)
Coagulopathy	1.58(0.92-2.72)	0.095 (.)	0.97(0.78-1.20)	0.778
Compdiabetes	1.37(1.16-1.62)	<0.001 (***)	1.01(0.92-1.11)	0.804
Anemia	0.87(0.71-1.061)	0.166	1.07(0.99-1.16)	0.081(.)
Electrolytes	0.96(0.71-1.29)	0.767	1.37(1.29-1.46)	<0.001(***)
Liver	0.77(0.59-0.98)	0.034 (*)	1.00(0.87-1.16)	0.952
Pvd	1.11(0.96-1.28)	0.147	1.12(1.05-1.19)	0.003(**)
Psychosis	1.17(0.66-2.07)	0.594	1.42(1.19-1.71)	0.556
Pulmcirc	1.60(1.25-2.05)	<0.001 (***)	0.97(0.88-1.07)	0.656
Hivaids	0.82(0.11-6.28)	0.850	0.63(0.08-4.87)	0.645
Hypertension	1.00(0.91-1.10)	0.990	1.05(1.03-1.08)	<0.001(***)

Table 4.1: Odds ratio from Logistic regression with first readmission. Lombardia and England dataset.

The main tools for this purpose are the Odds Ratio and the relative Confidence interval (on 95% level), because it gives an immediate idea of the impact of each covariate on the probability of readmission. We start analysing the effect of the procedures on the probability of readmission (see Figure 4.1). In the case of the England dataset, interventions like $CABG$ and $PTCA$, being invasive procedures, could be strongly associated to the readmission of a patient. We have to underline, especially, the consequence of $PTCA$ procedure at first admission. It highly increases, indeed, the probability of readmission. On the other hand, light procedure like ICD lessen the probability of readmission in all considered cases. This is interesting, because it highlights the benefits brought by this medical procedure.
In the Italian dataset all procedures behave differently, because they reduce or don't influence the probability of readmission. This is observable especially in the first readmission. In this case, indeed, only $CABG$ plays a determinant role, reducing the probability of readmission after the first recovery. This effect is different from the England one. In Italian dataset, moreover, the effect of procedures is much more evident in the second readmission: $SHOCK$ and ICD,

	Lombardia dataset		England dataset	
	OR (CI)	**p-value**	**OR (CI)**	**p-value**
sex (male)	1.13(0.99-1.28)	0.053 (.)	0.99(0.97-1.02)	0.733
age	0.99(0.98-1.00)	0.039 (*)	0.99(0.99-1.00)	<0.001(***)
LOS	1.00(0.99-1.01)	0.138	1.00(1.00-1.00)	<0.001(***)
Procedures				
ICD	0.65(0.45-0.92)	0.016(*)	0.64(0.52-0.80)	<0.001(***)
CABG	0.76(0.44-1.32)	0.329	2.03(1.71-2.41)	<0.001(***)
PTCA	0.97(0.70-1.34)	0.850	1.57(1.37-1.80)	<0.001(***)
SHOCK	0.35(0.18-0.676)	0.002(**)	1.10(0.70-1.73)	0.676
Comorbidities				
Metastatic	0.73(0.41-1.31)	0.291	1.29(1.13-1.47)	0.001 (**)
Dementia	0.77(0.49-1.08)	0.127	1.29(1.21-1.38)	<0.001(***)
Renal	1.31(1.15-1.50)	<0.001(***)	1.12(1.08-1.16)	<0.001 (***)
Wtloss	0.41(0.99-1.68)	0.216	1.10(0.99-1.24)	0.083(.)
Hemiplegia	0.40(0.23-0.69)	<0.001(***)	1.15(1.03-1.29)	0.012(*)
Alcohol	1.12(0.39-3.16)	0.832	1.18(1.08-1.29)	0.002(**)
Tumor	1.04(0.82-1.31)	0.756	1.14(1.05-1.23)	0.001(**)
Arrhythmia	0.97(0.86-1.10)	0.652	1.06(1.03-1.08)	<0.001(***)
Pulmonarydz	1.05(0.91-1.21)	0.496	1.18(1.14-1.21)	<0.001 (***)
Coagulopathy	0.82(0.38-1.76)	0.604	1.13(0.96-1.32)	0.140
Compdiabetes	1.05 (0.86-1.28)	0.654	1.03(0.96-1.11)	0.456
Anemia	0.94(0.77-1.16)	0.588	0.98(0.92-1.05)	0.582
Electrolytes	0.89(0.65-1.21)	0.463	1.03(0.98-1.08)	0.270
Liver	1.01(0.77-1.33)	0.931	0.99(0.87-1.11)	0.809
Pvd	1.09(0.92-1.29)	0.313	1.08(1.02-1.13)	0.0386(*)
Psychosis	0.76(0.33-1.73)	0.507	1.58(1.36-1.86)	<0.001(***)
Pulmcirc	1.04(0.81-1.35)	0.741	0.90(0.83-0.97)	0.062 (*)
Hypertension	1.04(0.92-1.17)	0.573	1.06(1.03-1.09)	<0.001(***)

Table 4.2: Odds ratio from Logistic regression with second readmission. Lombardia and England dataset.

indeed, are inclined to lower the probability of a second readmission. This is similar to the England dataset. The effect of the comorbidities is observed in Figure 4.2 and Figure 4.3.

First of all, we can see that the presence of a comorbidity increases or keeps constant the probability of readmission in quite all cases. This consideration lets us say that the models are coherent with the reality (in fact, it is senseless that the presence of a comorbidity decreases the probability of readmission). We now analyse the odds ratio in detail.

Starting from Figure 4.2 and 4.3, it is remarkable the effect that each covariates has on English results. The following comorbidities (*dementia*, *renal*, *alcohol*, *tumor*, *arrhythmia*, *pulmonarydz*, *pvd*, *psychosis* and *hypertension*) have the effect to increase the probability of readmission in both models. The comorbidities *dementia*, *tumor*, *alcohol* and *psychosis* have wider intervals than the others, because these diseases are less common and, consequently, the margins of errors increase. In the other cases, the estimates are very accurate. That is what we expected, because *arrhythmia* and *hypertension* are comorbidity directly related to Heart Failure, so they are good predictors for further readmissions. Furthermore, jointly to *renal* and *pulmonarydz*, they also are very common comorbidities (if compared with alcohol or psychosis, see Table 2.2). Most of the covariates that are not influential on the first readmission, keep constant their

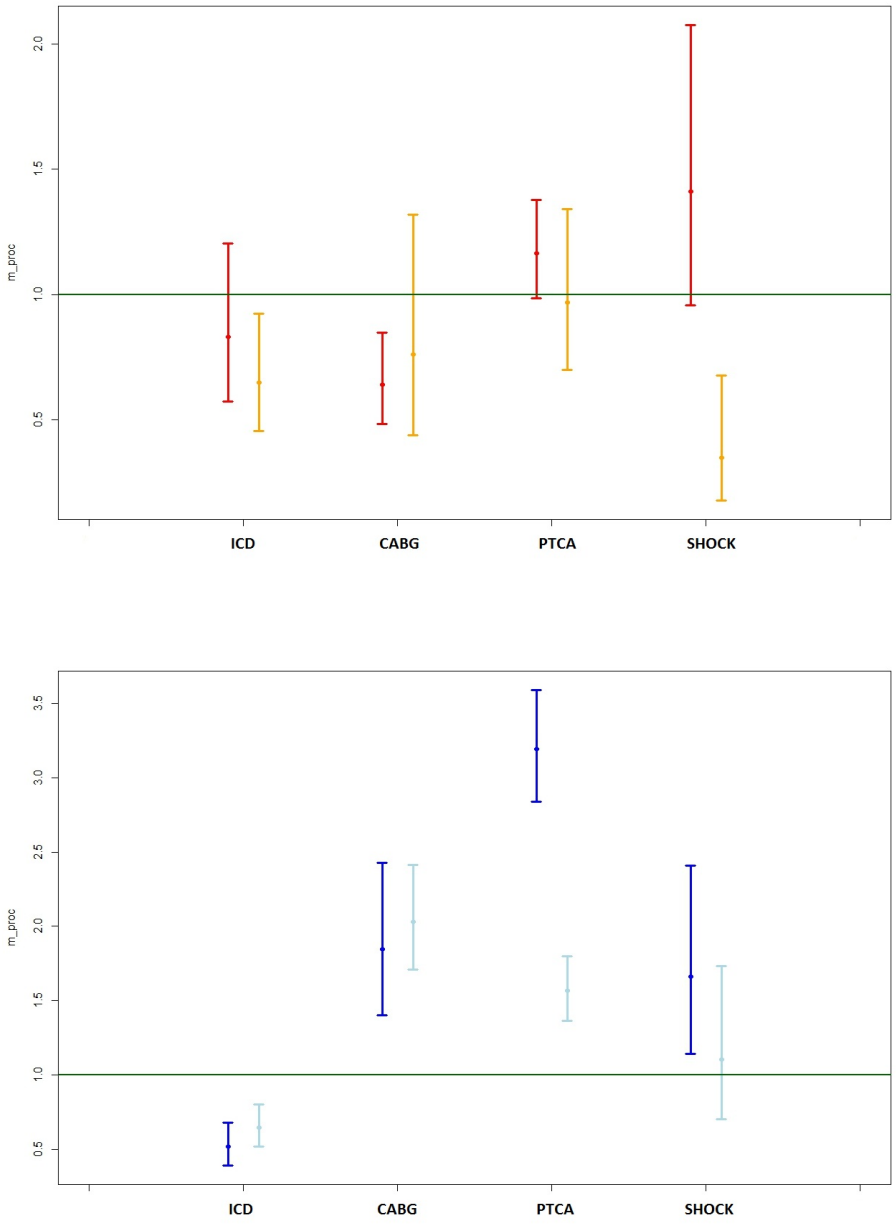

Figure 4.1: Odds Ratio and confidence intervals for the procedures of all Logistic regression models. Lombardia (above) and England (below) dataset. First readmission: red and blue. Second readmission: orange and light blue.

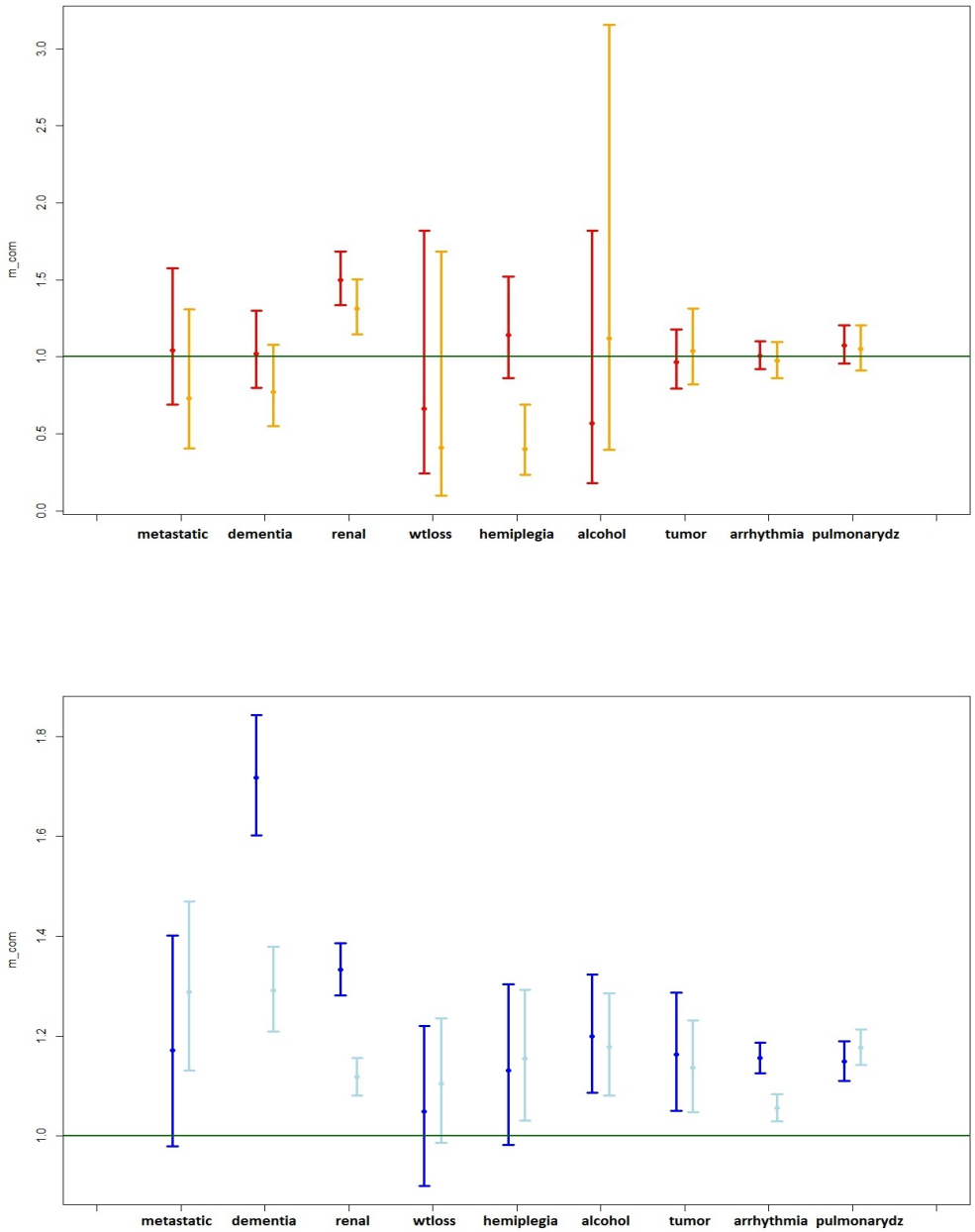

Figure 4.2: Odds Ratio and confidence intervals for the comorbidities of all models. Lombardia (above) and England (below) dataset. First readmission: red and blue. Second readmission: orange and light blue.

effect in the second readmission too. So, *wtloss*, *coagulopathy*, *compdiabetes*, *anemia* and *liver* are never determinant. On the other hand, we can see that some covariates are differently influential: *electrolutes*, for example, is a good predictor for a first readmission, while it becomes inconsequential in predicting a second re-hospitalization. Good predictors for a second readmission but not for a first one is *metastatic*.
The Italian dataset has a similar behaviour, but includes less significant variables. Of course, we have to take count of the lesser quantity of data. The *renal* comorbidity is the only covariate that keeps its positive influence in both model. Also in this case, the width of the confidence interval is small (if compared with other quantities). The behaviour of the leftover covariates is quite similar. The comorbidities *compdiabetes*, *pulmcirc* and *pvd* increase the probability of first readmission, but they become less important in the other model. On the other hand, *liver* seems to decrease the probability of first readmission, although it is close to the value 1. The last interesting behaviour is recorded by *hemiplegia*, which has a negative effect ion the probability of second readmission. Due to the low bearing, we have to consider it as predictor with attention.
In Figure 4.4 we can see directly the effect of the covariates in both the dataset. We have considered only the model with first readmission, because it will be useful also for further models implemented, especially for Hurdle and Zero-Inflated models. The first remarkable observation is the different width of confidence intervals; that is due to the different amount of data as well as a difference quality of data.
Apart from these kind of considerations, we now observe the compared effect of comorbidities on the probability of first readmission within 30 days. The effect of *renal* disease is relevant in both cases, but more incisive in Lombardia than in England, because it increases the risk of readmission about of the 50% while in the case of UK the effect of renal disease is around the 33%. That could be due to the less information brought by Lombardia data, mainly in those disease with low percentage. It is not unlikely that the *renal* disease includes the informations of these disease.
Looking at the other covariates, we note that, apart for *pvd*, the behaviour of the two dataset is quite discordant: those relevant comorbidities for the Lombardia dataset (*compdiabetes* or *pulmcirc*, which increase the probability of first readmission of 37% and 60% respectively) are not important for the England dataset, and vice versa. We can observe, indeed, that some comorbidities are relevant for the England dataset and absolutely not for the Lombardia dataset: *dementia* (impact of 60%), *arrhythmia* (impact of 16%), *pulmonarydz* (impact of 15%), *electrolytes* (impact of 37%) and *hypertension* (impact of 5%).
Basically, the application of these first two models allows to say that the factors that help in predicting a readmission within 30 days are quite different for the two countries. The only comorbidity in common between Lombardia and England, indeed, is *renal*, which we can consider as a strong predictor. All the other pathologies have a different impact in the short-term. That could be due to the time frame chosen, to the endemic differences between Lombardia and England and, above all, to the fact that in Lombardia there is just less power to detect small effects. To verify these assertions, we proceed in implementing a new model that give new perspective to the problem of predicting readmissions.

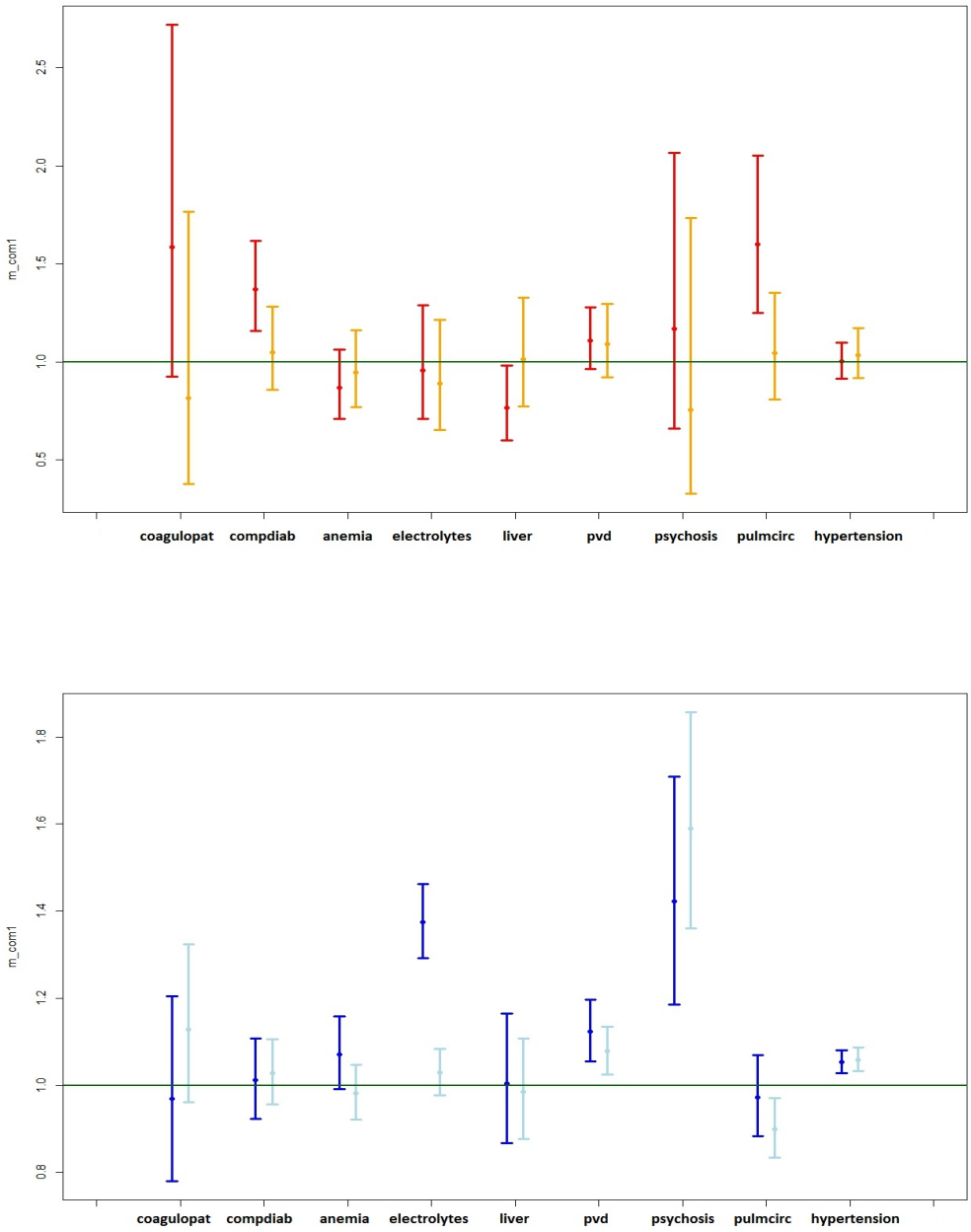

Figure 4.3: Odds Ratio and confidence intervals for the comorbidities of all models. Lombardia (above) and England (below) dataset. First readmission: red and blue. Second readmission: orange and light blue.

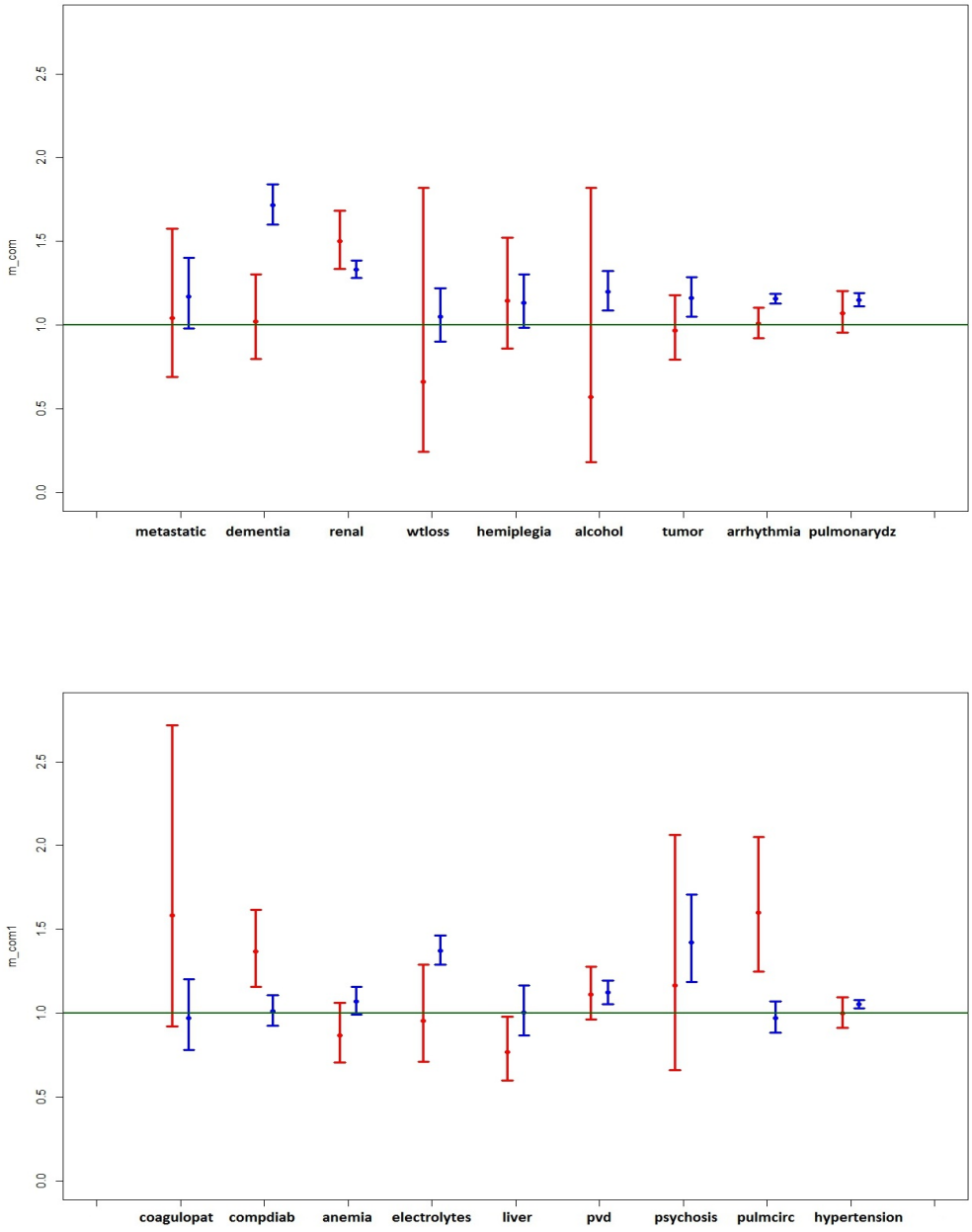

Figure 4.4: Odds Ratio and confidence intervals for Logistic Regression with all rows. Lombardia (orange) and England (light blue) dataset. All covariates.

4.2 Results of Hurdle and Zero-Inflated Models

Once we have examined the probability of first and second readmission, we have implemented a different model. That is because we want to widen our perspective and to discover which covariates may be influential in predicting multiple readmissions, and not just a specific one within a fixed short time.
As previous said, in this analysis the response is the total number of re-hospitalizations per patient within a year since the first admission. That allows to widen the time frame (from 30 days to one year), which is useful for longer-term predictions. This is a second step that complete our overview on the prediction of Heart Failure readmissions.
So, in the case of counting data, each row represents a patient and contains a list of summarized quantities of interest (comorbidities and procedures).

We have implemented six different models just to choose the one that fits better the data. So, before presenting the results related to comorbidities and procedures, we give an overview of the data distribution and an estimate of the significance of each model.
A first sight, of course, is given to the shape of readmissions. We can immediately notice (Figure 4.5) the large amount of zeros. This is the main reason that supports the use of Hurdle and Zero-Inflated models to compare the results from both the datasets. When dealing with counting data, in fact, the large amount of zeros could twist the counting process. So, Hurdle and Zero-Inflated models are a suitable way to remedy this problem and to have the right response.
Another different feature is the distribution of readmissions between Lombardia and England. In the first case we have a high percentage of zeros and a small quantity of readmissions (the highest number of readmission is 10); in the latter case the peak in zero is smaller, the number of readmissions increases in percentage and the tails become longer (the highest value is 25).

4.2.1 Models overview

We have implemented six different models in order to understand which is the best one to fit the data. We proceed in that way because we just want to have the best predictors for the number of readmissions within a year. After having chosen the right one, we proceed with the comparison of the results.
A summary of the goodness of fit in terms of LogLikelihood and zero predicted is given in Table 4.3. Those are the chosen tools to find the best model. As we can see, the Hurdle models are the best in predicting the number of zeros, in both dataset. The Poisson model drift towards underestimate of number of zeros, while Negative Binomial and Zero-Inflated models overestimate it. This is a first important feature that we have to consider.
The second one, of course, is the estimation of Maximum LogLikelihood. The outputs given by **R** are symmetrical. In each model, in fact, the choice of Negative Binomial distribution for the counting part leads to better results in terms of LogLikelihood, that means that the data are scattered or skewed. Meanwhile, the Zero-Inflated models give a best estimation of LogLikelihood, if we compare them with the relative Hurdle model.

Moreover, another difference lies in the computational costs: Hurdle models require less iterations in BFGS (BroydenFletcherGoldfarbShanno) optimization

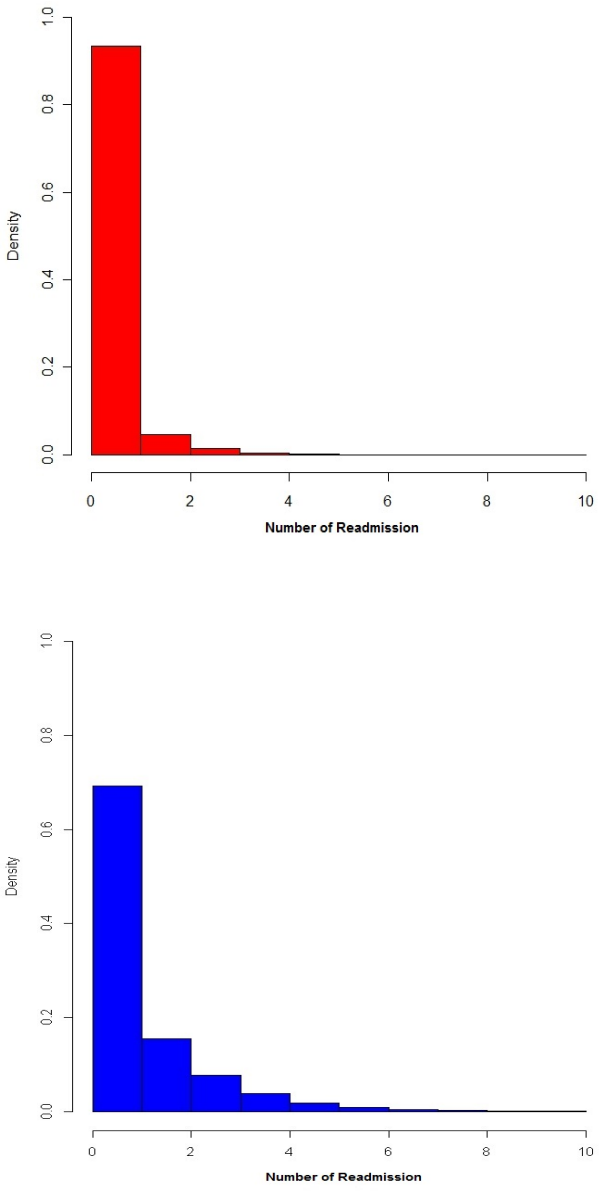

Figure 4.5: Histogram of readmissions per patient within a year since the first admission. Lombardia and England dataset.

(35 and 38 for Lombardia data, 36 and 43 for English data, Poisson and Negative Binomial respectively) than Zero-Inflated models (69 and 76 for Lombardia data, 65 and 81 for England data, Poisson and Negative Binomial respectively). The best choice is a compromise: of course, we can't keep the models without

	Lombardia dataset			Lombardia dataset		
	2×LogLike	Zero Predicted	Gap (obs - pred)	2×LogLike	Zero Predicted	Gap (obs - pred)
Poisson	-54,010	25,473	-442 (-1.7%)	-655,839	82,529	-11,233 (-11.9%)
Neg Bin	-43,686	26,440	+525 (+2.0%)	-649,519	97,116	+3,351 (+3.57%)
Hurdle Poisson	-42,040	25,915	0	-644,440	93,763	+9 (+0.0001%)
Hurdle Neg Bin	-41,820	25,915	0	-624,800	93,763	+9 (+0.0001%)
Zero Infl Poisson	-39,200	26,366	+451 (+1.7%)	-637,000	99,882	+6,126 (+6.52%)
Zero Infl Neg Bin	-39,200	26368	+453 (+1.7%)	-625,400	102,832	+9,1265 (+9,73%)

Table 4.3: Estimate of goodness of fit for all counting models (-2×LogLikelihood and zeros predicted). Lombardia and England dataset.

zero-counting part (Poisson and Negative Binomial). On the other hand, we privilege the zero-prediction and the computational costs over the LogLikelihood estimation. That leads us to choose the Hurdle Negative Binomial in both dataset. However, in graphs we also keep the Zero-Inflated Negative Binomial model for a deeper analysis.

4.2.2 Outcomes of Chosen Counting Models

The same order adopted for Logistic regression is followed: we start examining the significance of each variable, and then we give a look to the effects.
We report the coefficients and the p-values of Hurdle model with a Negative Binomial counting distribution, because this is the model that we retain more suitable to our data. It is also interesting to compare these coefficients with the one of Logistic regression of first readmission, because the Zero-Hurdle part corresponds to the event of non-readmission in Logistic regression (of course, within a year instead of 30 days).
First of all, the Zero-Hurdle part is analysed: in UK dataset, all covariates are relevant to explain the zero hurdle part, and that is mostly aligned with the results of Logistic regression. Some predictor not meaningful in the short term, becomes important in the long term (*wtloss*, *coagulopathy*, *compdiabetes*, *psychosis* and *pulmcirc*)
In the Lombardia case, the major part of the main predictors is still the same of Logistic regression, but we can record some dissimilarity: predictor like *metastatic*, *dementia*, *arrhythmia*, *pulmonarydz*, *anemia*, *electrolytes*, *hypertension* and *ICD* now become significant to predict another admission. That could be due to the extension of the readmission time (one year instead of 30 days). Other variables, instead, lose relevance, as *coagulopathy* and *liver*.
For the counting part, the comorbidities that impact on the number of readmis-

sion are *renal*, *hemiplegia*, *arrhythmia*, *pulmonarydz* and *hypertension*, and the most important procedures are ICD and $CABG$.
On the other hand, England dataset keeps quite all comorbidities and procedures as predictors for the counting part, except for $SHOCK$ and for *hivaids*, which is never important in these analysis. A similar significance of coefficients is observable in Zero-Inflated models, for both dataset.
After having found the significant covariates, we try to understand the impact fo covariates on the probability of readmission and on the number of readmission too.
In this analysis, of course, the major instruments still remain the confidence intervals. For a complete overview, we have plotted also the results coming from Zero-Inflated models with the Negative Binomial counting distribution. We have to be careful especially in the interpretation of the Zero-Inflation and Zero-Hurdle part. In fact, for the zero part we are using a logit model in both cases, but with a substantial difference: in Hurdle model, p represents the probability of overcoming the hurdle in zero (a successful event), while in Zero-Inflation model p represents the probability of belonging to the zero-inflation. The estimation of coefficients for the counting part is the same in both the models. This specification is necessary to understand the values of the Odds Ratio.
The first overview is given to the effect of comorbidities in Zero counting part (Figure 7.2). At a first sight, it is interesting the symmetry in the England dataset, where the comorbidities play an important role in augmenting the probability of readmission. In fact, in the case of Hurdle models, all coefficients of comorbidities are higher than 1 (so, the comorbidities affect positively the probability of readmission). In Zero-Inflated model, instead, all coefficient are significantly lower than 1 (so, if a patient is not affected by a comorbidity, the probability of not being readmitted increases). Moreover, the modulus of Zero-Inflation coefficients (not Odds ratio) is higher on average, if compared with Hurdle coefficients. That probably causes the major number of zeros predicted.
In the case of Lombardia, we can observe the same behaviour in the significant comorbidities (*dementia*, *renal*, *arrhythmia*, *compdiabetes*, *pulmcirc* and *hypertension*) and a discordant effect only for *hemiplegia*.
Very important is the joined effect of *metastatic*, *renal*, *arrhythmia* and *compdiabetes* in both dataset, because in Hurdle model they have the same impact on the probability of readmission. This is an interesting similarity between Lombardia and England.
Different is the compared effect of other covariates: *dementia*, *anemia*, *electrolytes*, *liver* and *hypertension* are more relevant in UK than in Italy, while the effect of *pulmcirc* is the contrary. Among the comorbidities that mostly affect Heart Failure patients (*renal*, *arrhythmia*, *pulmonarydz* and *hypertension*), we can observe that *renal* and *arrhythmia* have a similar impact (increasing the probability of readmission of 95% and 75% respectively). The comorbidities *pulmonarydz* and *hypertension*, on the other hand, have a higher impact on the England dataset (84% and 86%) than in the Lombardia one (38% and 32%).
Moreover, *renal* is the only predictor for the first readmission in short-term (30 days) and long-term (1 year), both for Lombardia and for England. In short term it increases the probability of readmission of 49% (Lombardia) and 33% (England); then, it highly rise its value. It is interesting noting that, in the Lombardia dataset, some covariates change behaviour from the short-term to the long-term. Some of them, especially, become meaningful predictors only

when considering a longer forecast.
In the England dataset, instead, the predictors for the long-term first readmission are the same of short-term, plus others in addiction.

After the analysis of zero component, we go into details of counting component. In this case, the longitudinal analysis (same dataset, different models) is much more easier. We can immediately notice that the trend of Confidence interval is similar for the counting part of Hurdle and Zero-Inflated model. The Lombardia dataset has, on average, a wider length of the intervals in the Hurdle model than in Zero-Inflated model, but this isn't a substantial matter for our analysis.
Also in this case, the covariates that mostly influence the number of readmission are *renal*, *arrhythmia*, *pulmonarydz*, *hemiplegia* and *hypertension*. *Arrhythmia* and *pulmonarydz* condition both the dataset with a similar weight. *Renal* and *hypertension* have a positive leverage in the counting part, but with a substantial difference: in the first case, the mentioned disease is much more relevant for Lombardia than for England (47% versus 17%); in the latter case the impact is the opposite (11% versus 36%). On the other hand, *Hemiplegia* has a completely opposite effect: it influence positively the counting part in England, while in Lombardia the effect is negative.
We also give an overview of the implication of procedures on readmissions of patients from the Table 4.4 and 4.5. Differently from the Logistic regression, in the Lombardia dataset the probability of being readmitted more than once in a year is strongly influenced by ICD, $PTCA$ and $SHOCK$, and they still influence further readmissions. $CABG$, instead, has the opposite effect for any kind of readmission (this behaviour is consistent with logistic regression).
In the England dataset, instead, all procedures make a patient be readmitted once at least. Once readmitted, further readmissions are influenced by ICD and $PTCA$.

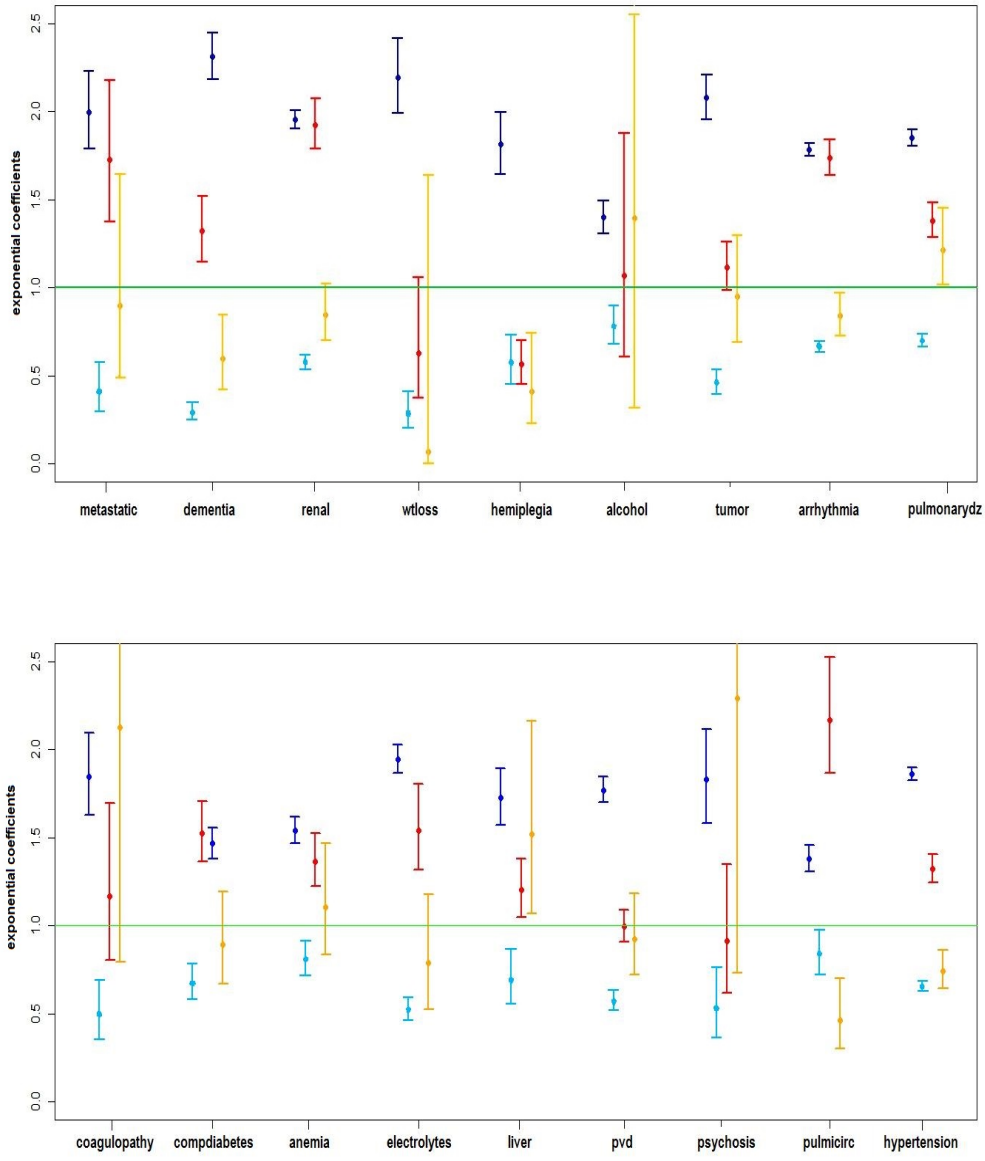

Figure 4.6: Confidence intervals for Zero-Hurdle (Zero counting part in Hurdle model) exponential coefficients (blue: England, red: Lombardia), and Zero-Inflation (Zero inflation part in Zero-Inflated model) coefficients (light blue: England, orange: Lombardia).

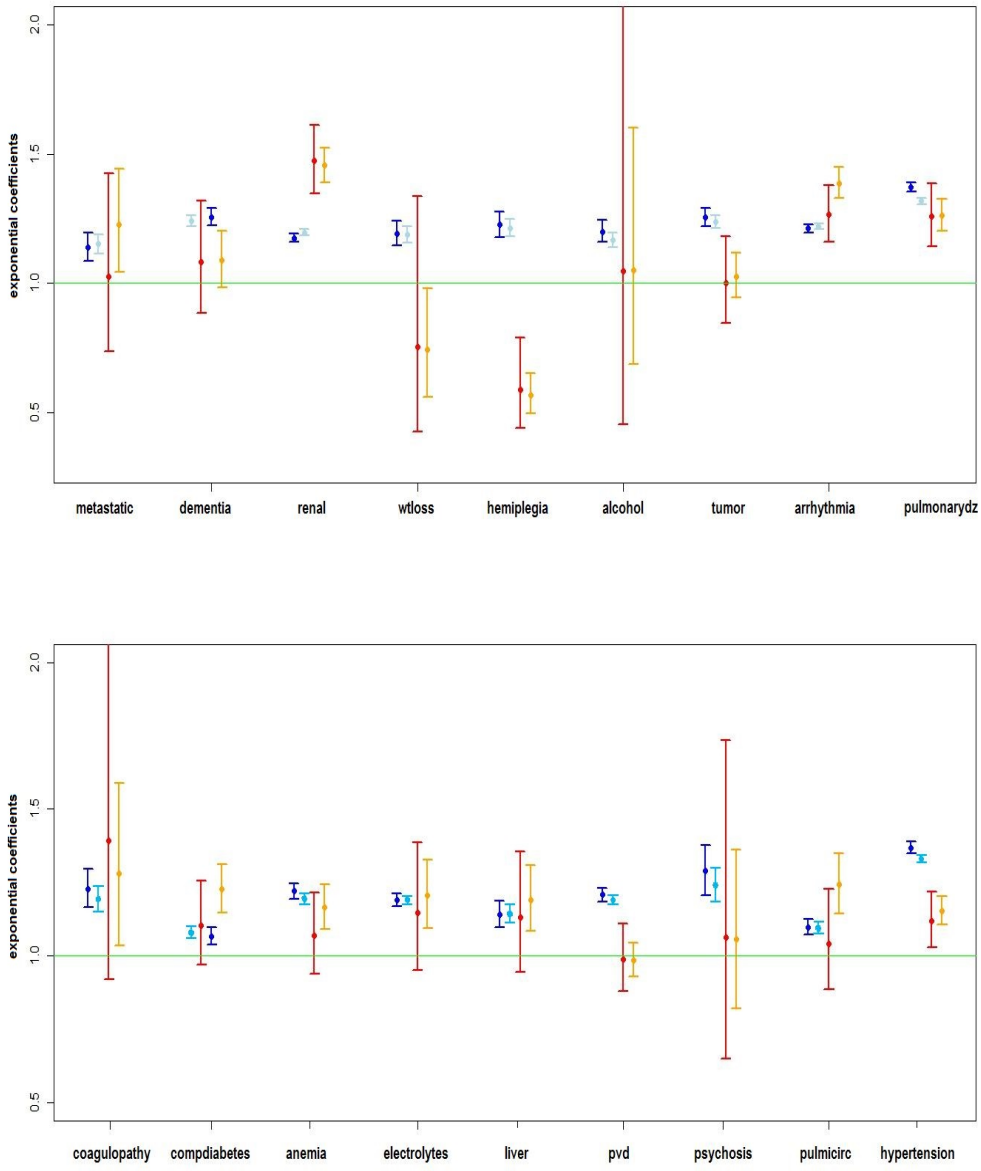

Figure 4.7: Confidence intervals for comorbidities exponential coefficients in counting part of Hurdle model (blue: England, red: Lombardia), and of Zero-Inflation model (light blue: England, orange: Lombardia).

	Lombardia dataset		**England dataset**	
	coefficient	pval	coefficient	pval
Sex	1.10(1.04-1.18)	0.002(**)	1.02(1.00-1.04)	0.027(*)
Age	1.00(1.00-1.01)	<0.001 (***)	0.99(0.99-0.99)	<0.001(***)
LOS	1.06 (1.05-1.058)	<0.001 (***)	1.05(1.05-1.05)	<0.001(***)
Procedures				
ICD	6.21(5.32-7.26)	<0.001(***)	1.50(1.14-1.96)	0.00294(**)
CABG	0.29(0.24-0.34)	<0.001(***)	2.48(1.00-6.14)	0.04871(*)
PTCA	1.40(1.26-1.57)	<0.001(***)	4.04(3.52-4.64)	<0.001(***)
SHOCK	2.52(2.05-3.10)	<0.001(***)	2.39(1.74-3.28)	<0.001(***)
Comorbidities				
Metastatic	1.72(1.37-2.17)	<0.001(***)	1.99(1.78-2.22)	<0.001(***)
Dementia	1.32(1.14-1.52)	<0.001(***)	2.31(2.18-2.44)	<0.001(***)
Renal	1.92 (1.78-2.07)	<0.001(***)	1.95 (1.90-2.02)	<0.001(***)
Wtloss	0.62(0.37-1.05)	0.0867(.)	2.19(1.99-2.41)	<0.001(***)
Hemiplegia	0.56(0.45-0.70)	<0.001(***)	1.81(1.64-1.99)	<0.001(***)
Alcohol	1.07(0.60-1.87)	0.785	1.39(1.30-1.49)	<0.001(***)
Tumor	1.11(0.98-1.25)	0.0678(.)	2.07(1.95-2.20)	<0.001(***)
Arrhythmia	1.73(1.63-1.83)	<0.001(***)	1.78(1.44-1.81)	<0.001(***)
Pulmonarydz	1.38(1.28-1.48)	<0.001 (***)	1.84(1.80-1.89)	<0.001(***)
Coagulopathy	1.16(0.80-1.69)	0.407	1.84(1.62-2.09)	<0.001(***)
Compdiabetes	1.52(1.36-1.70)	<0.001(***)	1.46(1.37-1.55)	<0.001(***)
Anemia	1.36(1.22-1.52)	<0.001(***)	1.54(1.46-1.61)	<0.001(***)
Electrolytes	1.53(1.31-1.80)	<0.001(***)	1.94(1.86-2.02)	<0.001(***)
Liver	1.20(1.04-1.38)	0.006	1.72(1.57-2.11)	<0.001(***)
Pvd	0.99(0.90-1.09)	0.869	1.76(1.69-1.84)	<0.001(***)
Psychosis	0.91(0.61-1.34)	0.644	1.82(1.57-2.11)	<0.001(***)
Pulmcirc	2.16(1.86-2.52)	<0.001(***)	1.37(1.30-1.45)	<0.001(***)
Hypertension	1.32(1.24-1.40)	<0.001 (***)	1.86(1.82-1.89)	<0.001(***)

Table 4.4: Zero coefficients for Hurdle model (binomial with logit link). Lombardia and England dataset.

	Lombardia dataset		**England dataset**	
	coefficient	pval	coefficient	pval
Sex	1.02(0.95-1.13)	0.395	0.98(0.97-1.00)	0.105
Age	0.99(0.99-1.00)	0.128	0.98(0.98-0.98)	<0.001(***)
LOS	1.02(1.01-1.02)	<0.001(***)	1.01(1.01-1.01)	<0.001(***)
Procedures				
ICD	1.87(1.63-2.15)	<0.001(***)	1.20(1.07-1.35)	0.001 (**)
CABG	0.59(0.49-0.71)	<0.001(***)	0.97(0.74-1.27)	0.849
PTCA	1.12(0.98-1.27)	0.016(*)	1.38(1.31-1.46)	<0.001(***)
SHOCK	1.07(0.88-1.29)	0.721	0.94(0.83-1.06)	0.324
Comorbidities				
Metastatic	1.02(0.73-1.42)	0.873	1.13(1.08-1.19)	<0.001(***)
Dementia	1.08(0.88-1.31)	0.433	1.25(1.22-1.29)	<0.001(***)
Renal	1.47(1.34-1.61)	<0.001(***)	1.17(1.15-1.19)	<0.001(***)
Wtloss	0.75(0.42-1.33)	0.342	1.19(1.15-1.23)	<0.001(***)
Hemiplegia	0.59(0.44-0.79)	<0.001(***)	1.22(1.17-1.27)	<0.001(***)
Alcohol	1.04(0.45-2.40)	0.911	1.20(1.15-1.24)	<0.001(***)
Tumor	1.00(0.84-1.18)	0.998	1.25(1.21-1.29)	<0.001(***)
Arrhythmia	1.26(1.16-1.37)	<0.001(***)	1.21(1.19-1.23)	<0.001(***)
Pulmonarydz	1.25(1.14-1.38)	<0.001(***)	1.33(1.35-1.39)	<0.001(***)
Coagulopathy	1.39(0.92-2.10)	0.115	1.22(1.16-1.29)	<0.001(***)
Compdiabetes	1.10(0.96-1.25)	0.150	1.06(1.03-1.09)	<0.001(***)
Anemia	1.06(0.93-1.21)	0.343	1.22(1.19-1.24)	<0.001(***)
Electrolytes	1.14(0.95-1.38)	0.157	1.19(1.16-1.21)	<0.001(***)
Liver	1.13(0.94-1.35)	0.176	1.14(1.09-1.18)	<0.001(***)
Pvd	0.98(0.87-1.10)	0.824	1.20(1.18-1.23)	<0.001(***)
Psychosis	1.06(0.65-1.73)	0.803	1.28(1.20-1.37)	<0.001(***)
Pulmcirc	1.04(0.88-1.22)	0.616	1.09(1.07-1.12)	<0.001(***)
Hypertension	1.11(1.02-1.21)	0.002(**)	1.36(1.34-1.38)	<0.001(***)

Table 4.5: Counting coefficients for Hurdle model (Negative Binomial). Lombardia and England dataset.

4.3 Results of Multi-State Models

The last step of this thesis work is the hardest one, too. A new class of models has been implemented to discover other quantities of interest for the whole process . Since now, the unique quantity of interest has been the influence of covariates on probability of first/further readmissions of Heart Failure patients. Only "illness" state has been inspected, while the transitions to death or recovery for this disease have never been considered (although important).
Multi-State models, instead, are suitable when we want to know the transition rate (or related quantities like probability of transition, mean sojourn time, total length of stay) between different states (illness, recovery and death). This development is due to a more complex interpretation of our data: now the clinical history of each Heart Failure patient belongs to a Markov process. That consideration allows the examination of the quantities mentioned above.
Moreover, a forward step is available: inspecting the impact individual(i) level covariates z_i on the transition rate between states ($q_{rs}(z_i)$=$q_{rs}^{(0)}\exp(\beta_{rs}z_i)$). That means that Multi-State models are a useful integration of the previous ones. Logistic regression and Hurdle/Zero-Inflated models, indeed, find out the significant covariates and the related impact on the readmission. Multi-State models go through this analysis and also explain their impact on the transitions, giving a wider overview of Heart Failure disease.
As mentioned in Chapter 3.3.2, the implemented models are two. The first one considers only three states: "Alive Inside the hospital" (equivalent to a generic admission for Heart Failure), "Alive Outside the hospital" (equivalent to a live discharge) and "Death". The second model, instead, is a development of the previous one, because each admission/discharge is specified until the third admission.

4.3.1 Three states model

The "Three states model" looks at the relation between admission-discharge-death without considering the specific number of admission. This model goes beyond Logistic regression and Counting models too, because now the covariates play a role in the transition between admission-discharge, discharge-death or admission-death, while before they were influential only in a subsequent admission within a fixed period.
Before analysing the effect of covariate, we inspect the first output of `msm()` package: the transition intensity matrix. Transition intensity matrix is the principal instrument to build all others quantity of interest. Instead of giving our evaluations on its basis, we start from the probability of each state being the next (instead of the transition intensity). As we can see in Table 4.6, in English dataset the probability of readmission ("Alive OUT" to "Alive IN") and discharge ("Alive IN" to "Alive OUT") are higher than in the Lombardia dataset. In the Lombardia dataset, instead, the transition from "Alive OUT" to "Death" as well as the death inside the hospital ("Alive IN" to "Death") are higher than in the England dataset. Our process, of course, is continuous in the time, that is why looking at the changes of probability of transition along the time is interesting too.
As we can see in Figure 4.8, both in the short and in the long term, the prob-

		Lombardia dataset	
	Alive In Hospital	Alive Out Hospital	Death
Alive In Hospital	0	0.912	0.083
Alive Out Hospital	0.657	0	0.343
		England dataset	
	Alive In Hospital	Alive Out Hospital	Death
Alive In Hospital	0	0.925	0.075
Alive Out Hospital	0.932	0	0.067

Table 4.6: Probability of each state being next, conditional to the change of state. Multi-State model with three states. Lombardia and England dataset.

ability of readmission is higher for an English patient than for an Italian one. Moreover, this transition is much more frequent in UK than in Italy.
Once a patient is admitted, it becomes interesting looking to the probability of the following possible state. In Figure 4.9, the probability of transition from "Alive inside the Hospital" to "Death" or to "Alive outside the Hospital" is shown. As we can see, Lombardia patients are inclined to die instead of being discharged alive in a long term (while it is the opposite for the England dataset); in a short-run, moreover, this propensity is unchanged. As conclusion, Lombardia patients are more likely to die inside the hospital than English ones, instead of being discharged alive. Nevertheless, these transitions in English dataset are faster than in the Lombardia one.
These feature didn't come out from the previous models, that is why we have reported it now.

Another interesting tool provided by `msm()` package is the appraisal of the Mean Sojourn Time and of Total Length of Stay (referred to each patient). As we can see in Table 4.7, the total length of stay as well as the Mean sojourn time is lower for the England dataset than for Lombardia one. This information, joint to the probability of transition, may be an handy and simple indicator to evaluate the efficiency of hospitals. Of course, on the other hand, it should be used in connection with other detailed material. On the other hand, we can see that the Total Length of Stay alive in Hospital is higher in England than in Italy, which means that English patients are more likely to be readmitted. Furthermore, the Total Length of Stay Alive out of Hospital is still higher in England than in Lombardia, which probably means that the death rates are lower in England than in Lombardia.

Anyway, the most important analysis that we can do is related to the impact

	Lombardia dataset	**England dataset**
Mean Sojourn Time		
Alive In Hospital (days)	13.96	10.64
Alive Out Hospital (days)	456.28	175.28
Total LOS		
Alive In Hospital (days)	34.90	77.48
Alive Out Hospital (days)	1040.72	1180.83

Table 4.7: Mean Sojourn Time and Total Length of Stay. Multi-State model with three states. Lombardia and England dataset.

of covariates on the transition rate, as done with the previous models. Once we have gained the general trend of Admission-Discharge-Death of Heart Failure

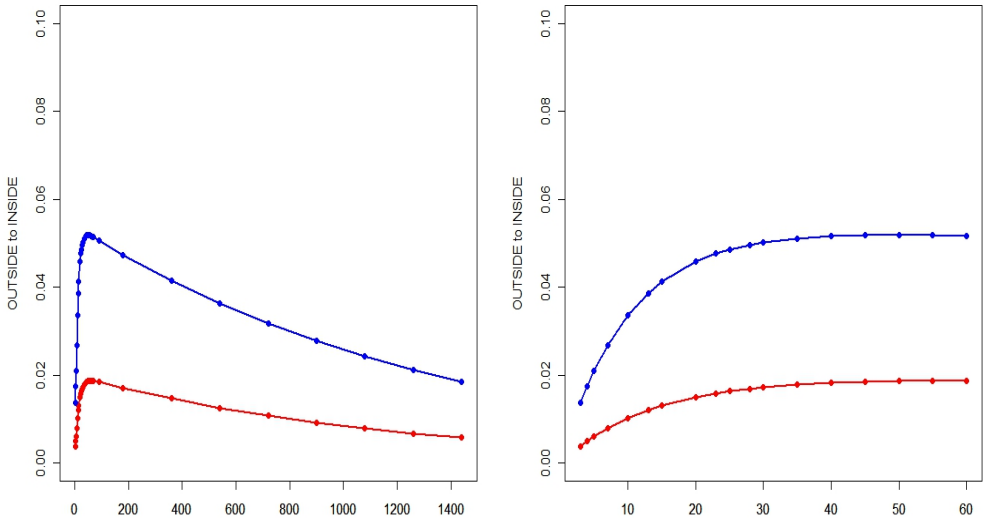

Figure 4.8: Trend of probability of readmission in Heart Failure patients. Lombardia dataset (red) and England dataset(blue). From 0 to 4 years (left) and from 0 to 2 months (right).

patients, indeed, it becomes interesting what may happen if a patient is affected by a specific disease (so, the impact of a disease/procedure on a specific transition).
Table 4.8, Table 4.9 and Table 4.10 report the exponential hazard ratio of the most interesting transitions (so: discharge, death inside the hospital and readmission).
First of all, we look at comorbidities impact. In Table 4.8, we can see the effect of comorbidities in the transition from "Alive Inside the Hospital" to "Alive Outside the Hospital". *Metastatic*, *renal*, *hemiplegia*, *compdiabetes*, *electrolytes* and *pulmcirc* are coherent for both the dataset, because they are significant and, in accordance with the common sense, they reduce the transition intensity of discharge. The decreasing influence of comorbidities, moreover, is always evident for those significant comorbidities in Italian dataset. In the England dataset, some comorbidity as an opposite effect: *alcohol*, *tumor*, *pulmonarydz*, *coagulopathy*, *pvd* and *hypertension*, indeed, have the effect of increasing the transition intensity of discharge. Remarkable is the behaviour of *hypertension* (always meaningful in previous analysis). In Italian dataset, *hypertension* is not important, while in the England dataset it has the effect of increasing the discharge rate. That behaviour may be due to the high prevalence.
Different is the effect of comorbidities on the death inside the hospital (Table 4.9). In that case the discrepancy between the two dataset is much more evident. *Metastatic*, *dementia* and *renal* are the only comorbidities with the same significance and effect in both dataset (they, of course, increase the transition rate). *Compdiabetes* and *pulmcirc* significantly decrease the transition rate,

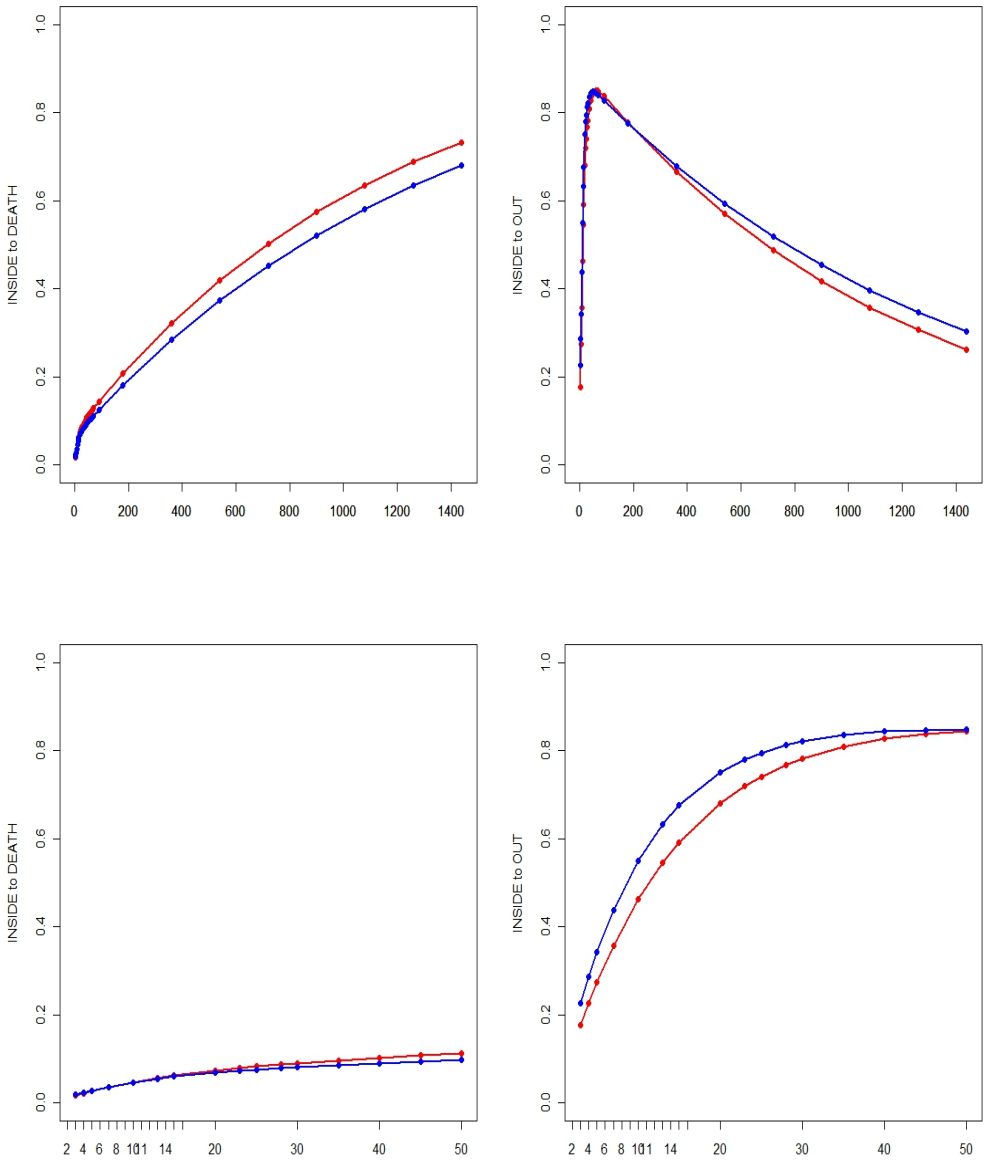

Figure 4.9: Trend of probability of death inside the hospital(left) and discharge (right) in Heart Failure patients. Lombardia dataset (red) and England dataset(blue). From 0 to 4 years (above) and from 0 to 2 months (below).

while *tumor*, *arrhythmia* and *electrolytes* have a discordant effect: *electrolytes* and *tumor* are higher than 1 in Lombardia dataset, while *arrhythmia* is lower. For the England dataset the influence is the opposite.
The last transition that we consider is from "Alive Outside the Hospital" to Alive Inside the Hospital (Table 4.10), that is equal to a readmission (the first state that a patient occupy, in fact, is his first admission). It is very interesting that these results are a very similar to the ones found in the counting part of Hurdle models. If we focus on English dataset, indeed, we can observe that all comorbidities have a positive impact in increasing the transition intensity; the weight of this growing effect is specular to the Odd Ratios found in Hurdle component, but also in the Logistic regression of first readmission (see, for example, *alcohol*, *pulmonarydz* or *psychosis*). We can say the same for the Lombardia dataset, in which the most evidence parallelism is observable in the decreasing effect of *hemiplegia* or in the highly increasing effect of *renal* and *compdiabetes*.
The last analysis is focused on the effect of procedures on the transition rates. A first thing that we can notice is related to England the dataset. As we can see, indeed, none procedure as an impact on the transitions that we are considering. This is important mainly when we consider the transition between "Alive Outside the Hospital" to "Alive Inside the Hospital", which is our readmission in Logistic regression. In this case we can't see the same parallelism recorded for comorbidities. In the case of the Lombardia dataset, instead, we observe that procedures always have an important impact in all transitions. We start from Table 4.8: in the transition of discharge, $CABG$, $PTCA$ and $SHOCK$ have a decreasing impact on the transition intensity, maybe due to their being invasive. That is true especially for $CABG$ and $SHOCK$, which take lower values than $PTCA$.
As we have seen in the previous tables (Table 4.4 and Table 4.5), some comorbidities increase the chance of both one readmission and many readmissions, in both the countries too. These comorbidities are *renal*, *pulmonarydz*, *arrhythmia*, *hemiplegia* and *hypertension*. Four of these are the most frequent in both the populations (*renal*, *pulmonarydz*, *arrhythmia* and *hypertension*), while *hemiplegia* is not highly frequent, but keeps its relevance in both dataset.

	Lombardia dataset		England dataset	
Procedures				
ICD	1.03 (0.98- 1.08)		1.00 (0.92-1.07)	
CABG	0.29 (0.28- 0.31)	*	0.99 (0.84-1.18)	
PTCA	0.70(0.68- 0.73)	*	1.00 (0.95-1.04)	
SHOCK	0.21 (0.19- 0.24)	*	0.99 (0.93-1.07)	
Comorbidities				
Metastatic	0.80 (0.74-0.86)	*	0.97(0.95-0.99)	*
Dementia	0.93 (0.89-0.97)		0.84 (0.83-0.85)	*
Renal	0.86 (0.84-0.87)	*	0.78 (0.77-0.78)	*
Wtloss	0.48 (0.42-0.55)	*	1.02 (1.00-1.03)	
Hemiplegia	0.59 (0.56-0.62)	*	0.93 (0.91-0.94)	*
Alcohol	1.00 (0.86-1.16)		1.22 (1.20-1.23)	*
Tumor	0.94 (0.91-0.97)	*	1.09(1.08-1.10)	*
Arrhythmia	0.94 (0.93-0.96)	*	0.86 (0.86-0.86)	*
Pulmonarydz	0.90 (0.88-0.92)	*	1.04 (1.04-1.05)	*
Coagulopathy	0.82 (0.75-0.90)	*	1.04 (1.02-1.06)	*
Compdiabetes	0.94 (0.92-0.97)	*	0.94 (0.93-0.95)	*
Anemia	0.88 (0.86-0.91)	*	1.00 (0.99-1.01)	
Electrolytes	0.91 (0.88-0.95)	*	0.90 (0.89-0.91)	*
Liver	0.94 (0.91-0.98)	*	0.98 (0.96-0.99)	
Pvd	0.87 (0.84-0.89)	*	1.08 (1.07-1.09)	*
Psychosis	0.93 (0.85-1.03)		1.03 (1.00-1.05)	
Pulmcirc	0.81 (0.79-0.85)	*	0.95 (0.94-0.96)	*
Hypertension	1.01 (1.00-1.03)		1.06 (1.06-1.07)	*

Table 4.8: Exponential hazard rate: Alive Inside the Hospital to Alive Outside the Hospital (live discharge).

	Lombardia dataset		England dataset	
Procedures				
ICD	0.34 (0.26- 0.44)	*	0.99 (0.76-1.30)	
CABG	0.24 (0.19-0.29)	*	0.99 (0.55-1.80)	
PTCA	0.31 (0.26-0.36)	*	1.00 (0.86-1.16)	
SHOCK	7.14 (6.67- 7.65)	*	1.00 (0.78-1.28)	
Comorbidities				
Metastatic	2.51 (2.18-2.88)	*	1.07 (0.99-1.15)	*
Dementia	2.21 (2.02-2.43)	*	1.04(1.00-1.07)	*
Renal	1.24 (1.17-1.31)	*	1.84 (1.81-1.87)	*
Wtloss	1.11 (0.84-1.47)		0.75(0.71-0.79)	*
Hemiplegia	0.93 (0.81-1.07)		0.76 (0.72-0.80)	*
Alcohol	0.97 (0.57-1.63)		0.78 (0.74-0.82)	*
Tumor	1.36 (1.24-1.48)	*	0.85 (0.82-0.89)	*
Arrhythmia	0.81 (0.77-0.85)	*	1.41(1.38-1.43)	*
Pulmonarydz	1.02 (0.96-1.08)		1.16 (1.14-1.18)	*
Coagulopathy	1.03 (0.82-1.28)		0.84 (0.78-0.91)	*
Compdiabetes	0.90 (0.82-0.99)	*	0.77 (0.74-0.79)	*
Anemia	0.87 (0.80-0.95)	*	0.57 (0.55-0.58)	*
Electrolytes	1.40(1.26-1.55)	*	0.67 (0.66-0.69)	*
Liver	0.90 (0.80-1.02)		0.89 (0.85-0.95)	*
Pvd	0.99 (0.92-1.07)		0.64 (0.62-0.66)	*
Psychosis	0.94 (0.68-1.29)		0.88 (0.80-0.98)	*
Pulmcirc	0.82 (0.73-0.92)	*	0.67 (0.65-0.70)	*
Hypertension	0.72 (0.68-0.76)	*	1.01 (0.99-1.03)	

Table 4.9: Exponential hazard rate: Alive Inside the Hospital to Death (death inside the hospital).

	Lombardia dataset		England dataset	
Procedures				
ICD	1.25 (1.10-1.42)	*	0.99 (0.92-1.08)	
CABG	0.90 (0.83-0.98)	*	0.99 (0.77-1.28)	
PTCA	1.11 (1.05-1.17)	*	0.99 (0.95-1.04)	
SHOCK	1.12 (0.97- 1.29)		0.99 (0.91-1.09)	
Comorbidities				
Metastatic	1.23 (1.08-1.40)	*	1.11 (1.08-1.13)	*
Dementia	0.77 (0.72-0.82)	*	1.19 (1.17-1.20)	*
Renal	1.45 (1.42-1.49)	*	1.29 (1.29-1.30)	*
Wtloss	0.86 (0.69-1.06)		1.15 (1.13-1.17)	*
Hemiplegia	0.86 (0.79-0.92)	*	1.24 (1.22-1.26)	*
Alcohol	0.91 (0.74-1.12)		1.27 (1.25-1.28)	*
Tumor	1.00 (0.96-1.05)		1.13 (1.11-1.14)	*
Arrhythmia	1.19 (1.16-1.22)	*	1.17 (1.17-1.18)	*
Pulmonarydz	1.22 (1.19-1.25)	*	1.32 (1.31-1.32)	*
Coagulopathy	1.57 (1.40-1.76)	*	1.13 (1.11-1.15)	*
Compdiabetes	1.33 (1.29-1.38)	*	1.22 (1.21-1.24)	*
Anemia	1.15 (1.11-1.19)	*	1.16 (1.16-1.17)	*
Electrolytes	1.10 (1.04-1.17)	*	1.21 (1.21-1.22)	*
Liver	1.02 (0.97-1.08)		1.18 (1.16-1.20)	*
Pvd	1.04 (1.02-1.08)	*	1.13 (1.12-1.13)	*
Psychosis	0.95 (0.83-1.10)		1.49 (1.5-1.53)	*
Pulmcirc	1.47 (1.40-1.54)	*	1.15 (1.14-1.16)	*
Hypertension	1.08 (1.05-1.10)	*	1.18 (1.17-1.18)	*

Table 4.10: Exponential hazard rate: Alive Outside the Hospital to Alive Inside the Hospital (readmission).

4.3.2 Multi-State Model with detailed admissions

After having inspected the easier model, we want to go deeper in the possibilities that Multi-State model gives. That has brought to the decision of giving much more importance to the specific transition between subsequent admissions (and relative discharges), in order to have a complete overview that may integrate the results collected by Logistic regression (first and second readmission) and by Counting model for readmissions.
As previous done, we initially focus on the generalities of the process, just related to probability of transition and similar quantities. Then, we will focus on the contribute that the most important covariates may give to specific transitions.
As earlier explained in Chapter 3.3.2, the states that a patient can occupy are the following: first, second and third admission and relative discharges, fourth admission (that includes also further admissions) and, at last, death, which is reachable from all the mentioned states (see Figure 3.6 to refresh the admitted transitions).
We will follow the same outline of section 4.3.1, starting from the probability of each state being the next. As we can see in Table 4.11, the two dataset present similarities and dissimilarities as well. The main similarity is observable in the probability of discharge for a Heart Failure patient: in both dataset, indeed, the probability of being discharged alive or, consequentially, dying during the admission is quite similar (around the 9 %). We have to note, however, that in the case of English dataset the probability of death inside the hospital at first admission is lower than in the following ones (this transition wasn't considered in the previous models, but it doesn't change our purpose). The principal difference lies in the probability of readmission, because for Italian patients is much likely to not be readmitted than for English patients. Nevertheless, this behaviour is inclined to change along the readmissions, because as much the number of hospitalization increases much more the readmission probability increases. The trend for the England dataset, on the other hand, is the opposite (but it is much less significant).
The last overview that we can give to the whole process is related to Mean Sojourn Time. We have not reported the Total Length of Stay because, differently from the previous one, in this model we don't have reverse transitions. As we can see in Table 4.12, Italian admissions for Heart Failure, on average, last more than the England one. On the other hand, the permanence alive outside the hospital is higher. This behaviour may underline the inclination of not being readmitted again (as mentioned above).

Of course, our main purpose is finding the role of each covariates on subsequent admissions. As done in the previous models, it would be better to put all covariates of interest inside the model. Unfortunately, due to computational reasons, that's no longer possible to do. The problem lies in the complication of the model and in the big amount of data. As much as the number of states (so, parameters) increases, the algorithm can't allocate a such big space of memory, especially if we want to consider covariates too. That can't run the chosen model.
To allow the program run, we have made a choice based on the previous results, in order to give less covariates in input to the model. The maximum number of covariates that we could put inside the model for the the England dataset is 4, while for the Lombardia dataset is 7. That is why we have decide to differ-

Lombardia dataset								
	1adm	**1disc**	**2adm**	**2disc**	**3adm**	**3disc**	**4adm**	**Death**
1adm	0	0.92	0	0	0	0	0	0.08
1disc	0	0	0.60	0	0	0	0	0.40
2adm	0	0	0	0.91	0	0	0	0.09
2disc	0	0	0	0	0.68	0	0	0.32
3adm	0	0	0	0	0	0.91	0	0.09
3disc	0	0	0	0	0	0	0.72	0.28
4adm	0	0	0	0	0	0	0	1

England dataset								
	1adm	**1disc**	**2adm**	**2disc**	**3adm**	**3disc**	**4adm**	**Death**
1adm	0	0.98	0	0	0	0	0	0.02
1disc	0	0	0.94	0	0	0	0	0.06
2adm	0	0	0	0.90	0	0	0	0.10
2disc	0	0	0	0	0.93	0	0	0.07
3adm	0	0	0	0	0	0.90	0	0.10
3disc	0	0	0	0	0	0	0.92	0.08
4adm	0	0	0	0	0	0	0	1

Table 4.11: Probability of each state being next, conditional to the change of state. Multi-State model with detailed admissions. Lombardia and England dataset.

	Lombardia dataset	**England dataset**
1adm (days)	14.0 ± 0	10.5 ± 0
1disc (days)	591.2 ± 3.6	263.5 ± 0.5
2adm (days)	13.5 ± 0.1	10.9 ± 0
2disc (days)	395.1 ± 3.7	195.8 ± 0.5
3adm (days)	13.9 ± 0.1	11.1 ± 0
3disc (days)	298.4 ± 4.0	160.1 ± 0.4
4adm (days)	389.4 ± 7.9	303.2 ± 1.3

Table 4.12: Mean Sojourn Time of Multi-State model with detailed admissions. Lombardia and England dataset.

entiate the procedure (4 for both datasets) from the comorbidities (19 for both datasets). Among comorbidities, we have chosen thanks to a double criterion. First of all, we have considered the ones that were most important both in Zero and in counting part of Hurdle and Zero-Inflated models, in order to widen our time perspective. Secondary, among these comorbidities, we have chose the most frequent in Lombardia and England populations. The chosen pathologies are as follows: *renal*, *arrhythmia*, *pulmonarydz* and *hypertension*. We have to keep in mind that the following results may be affected by this choices, because we are not considering the combined effect of all covariates.
We start considering the effect of the procedures. In Table 4.13 we can see the effect of ICD and $CABG$, while in Table 4.14 we can see the effects of $PTCA$ and $SHOCK$. We will mainly focus on the effect on readmissions, which is the most important transition for this research, but we will give a look to the other transitions as well.
In the England dataset, in no one case the procedures have an impact in explaining the readmissions. The same behaviour is observable in the contribution to death inside the hospital, except for $SHOCK$, which has an impact in accelerating this process from the second readmissions onwards. The impact of

procedures in the England dataset, moreover, is weakly visible in the transition of discharge: $PTCA$ and ICD seem to accelerate the transition of discharge, while $SHOCK$ and $CABG$ have an opposite effect. This result wasn't caught in the previous Multi-State model.
In the Lombardia dataset, the procedures play a role to explain the transitions, much more than in the England dataset. In readmission transitions, for example, the impact of ICD and $CABG$ is similar: they accelerate the first readmission, but they influence oppositely further readmissions. $PTCA$ and $SHOCK$, on the other hand, are influential only in accelerating the first readmission and not the following ones. These results seem to contrast the results found in Logistic regression (especially with the models with first and and second readmissions), but we have to keep in mind that in Logistic regression we were analysing the probability of readmission within 30 days from the discharge, while in that case we are not considering this time limit. This impact, however, is different also in Multi-State model with three states.
When considering the death inside the hospital, the results are aligned with the ones found in the first Multi-State model: ICD, $CABG$ and $PTCA$ reduce the transition intensity, while $SHOCK$ highly increases it. In quite all processes of discharge, instead, the procedures are inclined to slow down the transition, as expected by common sense. This range of inspecting the impact of covariates on all transitions was previously impossible to gain.

We conclude our overview with the impact of comorbidities on transition

	Lombardia dataset		**England dataset**	
	ICD	**CABG**	**ICD**	**CABG**
Discharge Alive				
1adm to 1disc	0.70 (0.64-0.77)	0.31 (0.29-0.33)	0.87 (0.75-1.00)	0.61 (0.44-0.85)
2adm to 2disc	1.24 (1.15-1.35)	0.34 (0.30-0.39)	1.34 (1.15-1.56)	0.72 (0.50-1.03)
3adm to 3disc	1.43 (1.28-1.60)	0.28 (0.21-0.36)	1.16 (0.98-1.38)	0.86 (0.51-1.45)
Death In Hospital				
1adm to Death	0.30 (0.19-0.48)	0.20 (0.16-0.25)	0.94 (0.32-2.80)	0.97 (0.12-7.63)
2adm to Death	0.37 (0.24-0.58)	0.34 (0.23-0.51)	0.88 (0.49-1.59)	0.91 (0.33-2.46)
3adm to Death	0.55 (0.31-0.97)	0.62 (0.38-1.02)	0.85 (0.48-1.53)	0.98 (0.23-4.12)
4adm to Death	0.53 (0.40-0.69)	0.92 (0.44-1.95)	0.95 (0.62-1.45)	0.95 (0.29-3.12)
Readmission				
1disc to 2adm	1.27 (1.13-1.42)	1.11 (1.01-1.23)	1.05 (0.90-1.22)	1.00 (0.66-1.56)
2disc to 3adm	0.90 (0.80-1.00)	0.80 (0.64-0.99)	0.93 (0.78-1.10)	0.99 (0.61-1.62)
3disc to 4adm	0.74 (0.64-0.85)	0.73 (0.55-0.97)	0.99 (0.52-1.90)	1.01 (0.34-3.01)

Table 4.13: Exponential hazard rate of ICD and CABG. Lombardia and England dataset.

intensities. As explained above, we have chosen the four comorbidities (*renal*, *arrhythmia*, *pulmonarydz* and *hypertension*) depending on the output of Hurdle models and on the proportion of the pathology. As done with the procedures, in Table 4.15 (Lombardia dataset) and in Table 4.16 (England dataset) we report the Exponential Hazard Rates for all the transition considered.
We start from the readmissions: *renal* and *pulmonarydz* are very influential in accelerating the transition in Lombardia and in England as well. Furthermore, the effect of these covariates is similar. The trend seems to be decreasing in both dataset for *renal* disease, while *pulmonarydz* keeps a constant influence (especially in the first and second readmission). The comorbidities *arrhythmia* and *hypertension*, instead, haven't an important impact on the transition intensities

	Lombardia dataset		England dataset	
	PTCA	**SHOCK**	**PTCA**	**SHOCK**
Discharge Alive				
1adm to 1disc	0.72 (0.69-0.76)	0.27(0.24-0.30)	1.33 (1.25-1.42)	0.53 (0.45-0.62)
2adm to 2disc	0.75 (0.68-0.82)	0.16 (0.13-0.21)	1.07 (0.98-1.18)	0.44 (0.35-0.54)
3adm to 3disc	0.80 (0.70-0.92)	0.16 (0.11-0.23)	1.14 (1.01-1.27)	0.76 (0.61-0.95)
Death In Hospital				
1adm to Death	0.28 (0.24-0.34)	7.25 (6.60-7.96)	1.17 (0.69-1.96)	1.17 (0.50-2.74)
2adm to Death	0.44 (0.31-0.63)	8.83 (7.70-10.13)	0.96 (0.72-1.28)	2.53 (1.93-3.32)
3adm to Death	0.61 (0.37-1.00)	8.26 (6.82-10.00)	1.00 (0.70-1.43)	1.87 (1.23-2.85)
4adm to Death	0.58(0.44-0.77)	4.31 (3.30-5.62)	1.05 (0.86-1.28)	1.71 (1.27-2.30)
Readmission				
1disc to 2adm	1.24 (1.16-1.32)	1.28 (1.07-1.52)	1.05 (0.90-1.22)	1.03 (0.89-1.21)
2disc to 3adm	1.02 (0.90-1.15)	0.97 (0.68-1.39)	0.93 (0.78-1.10)	1.01 (0.82-1.24)
3disc to 4adm	1.14 (0.95-1.36)	0.89 (0.39-2.02)	0.99 (0.52-1.90)	1.23 (0.94-1.60)

Table 4.14: Exponential hazard rate of PTCA and SHOCK. Lombardia and England dataset.

(except for *arrhytmia* in the England dataset for the first readmission). This result is different from the first Multi-State model and from all previous models, in which these latter covariates were significant for both the dataset. If we give a look to the discharges, instead, we note that *renal* and *arrhythmia* are significant in similar way, because the decrease the transition intensity among all kind of that transition. In the Lombardia dataset, this behaviour is unchanged for the leftover comorbidities (*pulmonarydz* and *hypertension*), while in English dataset *pulmonarydz* and *hypertension* invert their contribution. This behaviour is exactly the same that we can see in the first Multi-State models (with three states).
We conclude looking at the contribution of comorbidities to the Death Inside the hospital. In the England dataset, a Heart Failure patient affected by *renal*, *pulmonarydz* and *arrhythmia* diseases is inclined to die faster. That is true for Lombardia patients too, but only if they are affected by *renal* disease. Interesting is the role of *hypertension* in the Lombardia dataset, because this comorbidities decreases the transition intensity of Death Inside the hospital.

	Lombardia dataset			
	renal	**arrhythmia**	**pulmonarydz**	**hypertension**
Discharge Alive				
1adm to 1disc	0.81 (0.79-0.84)	1.02(1.00-1.04)	0.91 (0.88-0.93)	0.96 (0.94-0.99)
2adm to 2disc	0.82 (0.79-0.85)	0.96 (0.92-0.99)	0.86 (0.83-0.89)	0.95 (0.92-0.99)
3adm to 3disc	0.80 (0.76-0.84)	0.91 (0.87-0.97)	0.89 (0.85-0.94)	0.98 (0.93-1.03)
Death In Hospital				
1adm to Death	1.09 (0.99-1.20)	0.66 (0.61-0.71)	1.01 (0.93-1.11)	0.62 (0.58-0.68)
2adm to Death	1.23 (1.10-1.38)	0.98 (0.88-1.09)	0.86 (0.76-0.96)	0.70 (0.63-0.78)
3adm to Death	1.26 (1.08-1.47)	0.88 (0.75-1.03)	1.05 (0.90-1.23)	0.62 (0.53-0.72)
4adm to Death	1.13 (1.05-1.23)	1.12 (1.03-1.22)	1.06 (0.98-1.15)	0.96 (0.88-1.02)
Readmission				
1disc to 2adm	1.23 (1.17-1.28)	0.98 (0.94-1.01)	1.08 (1.04-1.12)	1.03 (0.99-1.06)
2disc to 3adm	1.14 (1.08-1.20)	0.98 (0.94-1.03)	1.10 (1.05-1.15)	0.95 (0.91-1.00)
3disc to 4adm	1.17 (1.10-1.25)	1.04 (0.97-1.11)	1.16 (1.09-1.24)	0.96 (0.90-1.02)

Table 4.15: Exponential hazard rate of *renal*, *arrhythmia*, *pulmonary disease* and *hypertension*. Lombardia dataset.

	England dataset			
	renal	**arrhythmia**	**pulmonarydz**	**hypertension**
Discharge Alive				
1adm to 1disc	0.73 (0.72-0.74)	0.85(0.84-0.86)	0.99 (0.98-1.00)	1.03 (1.02-1.04)
2adm to 2disc	0.73 (0.72-0.74)	0.85 (0.85-0.86)	1.01 (1.00-1.02)	1.05 (1.04-1.06)
3adm to 3disc	0.76 (0.75-0.77)	0.85 (0.84-0.86)	1.04 (1.02-1.05)	1.04 (1.03-1.05)
Death In Hospital				
1adm to Death	2.24 (2.09-2.40)	1.81 (1.74-1.88)	1.33 (1.23-1.44)	1.28 (1.20-1.37)
2adm to Death	1.63 (1.58-1.68)	1.20 (1.16-1.23)	1.03 (0.99-1.06)	0.77 (0.75-0.79)
3adm to Death	1.55 (1.50-1.61)	1.19 (1.15-1.23)	1.03 (1.00-1.07)	0.78 (0.76-0.81)
4adm to Death	1.28 (1.25-1.30)	1.17 (1.15-1.19)	0.96 (0.94-0.98)	0.96 (0.94-0.98)
Readmission				
1disc to 2adm	1.27 (1.25-1.28)	1.07 (1.06-1.08)	1.09 (1.08-1.10)	1.01 (1.00-1.02)
2disc to 3adm	1.16 (1.14-1.17)	1.01 (1.00-1.02)	1.10 (1.08-1.11)	1.00 (0.99-1.01)
3disc to 4adm	1.12 (1.10-1.13)	1.00 (0.99-1.01)	1.09 (1.07-1.10)	0.99 (0.98-1.01)

Table 4.16: Exponential hazard rate of *renal*, *arrhythmia*, *pulmonary disease* and *hypertension*. England dataset.

Chapter 5

Conclusive Remarks

This work explores the problem of predicting the readmissions of Heart Failure patients, comparing the results of two datasets: the first one from Lombardia (and Italian region) and the second from England (a county within the United Kingdom). This being main motivating reason of the project above described.
Heart Failure, indeed, is very common in our society and, being a chronic disease, leads patients to be readmitted more than once. The challenge for researchers and hospitals is to find out suitable predictors for further readmissions. The reason lies in improving the hospital care (through the upgrade of therapies and the targeting of interventions) and, consequentially, in saving money.
This purpose is pursued by using statistical methods, which have proven suitable tools for several reasons: they take advantage of already existing datasets (for example, administrative database) and they easily adapt to the response of interest, in order to find the enquired responses. The prediction of readmissions is a topic strongly supported by literature. For example,Bartolomeo et al. (2008) have enquired the problem of readmission in Chronic Obstructive Pulmonary Disease (COPD); Bottle et al. (2014), Ieva et al. (2015), Philbin and DiSalvo (1999) and Postmus et al. (2012) focused on the problem of readmission of Heart Failure patients, using different statistical tools.
However, in this work we have introduced a new enriching point of view: the comparison between different datasets. Many reasons justify this choice: to begin with, we gain a stronger perspective on the problem of readmissions; moreover, we contrast the phenomenon of Heart Failure in two countries and, finally, we can compare the efficiency of the Health Systems in facing the problem of readmissions. Highlighting similarities and dissimilarities is the motivating reason that guides this work.
Furthermore, thanks to the statistical methods applied, this project gains a wide perspective: in Logistic regression, our response variable is the readmission within 30 days from the discharge; in Hurdle an Zero-Inflated models our response is the number of readmissions per patient in a year; in Multi-States models we have analysed the process of admission-discharge-death itself, enquiring the transition intensities among different states. In all these models, the aim was finding out the weight of covariates on the response of interest and enquiring the process itself. Across models, therefore, a multiface insight of the problem of Heart Failure readmissions is provided.
Let us now give a summary of the results across dataset and implemented mod-

els.
The first basic comparison has been done between the features of populations, which has highlighted a similarity in the distribution of age and in the proportion of diseases. Indeed, the comorbidities with high percentage in England dataset (*renal*, *arrhythmia*, *pulmonarydz* and *hypertension*) present similar values in Lombardia dataset. Except for *electrolytes* (higher in England than in Lombardia), the same parallelism is recorded for less frequent pathologies. The main difference between Lombardia and England lies in the proportion of procedures done: in Italian dataset they are always double than England one, even if the percentage are not high.
The implementation of the models has disclosed the following features: *renal* is the most important predictor in all models implemented and in both datasets too. It substantially gives a positive contribution in readmissions both in short-run (Logistic regression) and long-run (Hurdle and Zero-inflated models and Multi-State models). This results is already confirmed by literature (see Postmus et al. (2012) and Bottle et al. (2014)).
The comorbidities *arrhythmia*, *pulmonarydz* and *hypertension* are important as well in most models for both the datasets, especially when dealing with models that do not require a short time limit to consider a readmission (Counting models and Multi-state models). Furthermore, these pathologies are the most frequent in both populations. We can assert that a patient affected by these pathologies are really more likely to be readmitted. Controlling these diseases may decrease the risk of further Heart Failure and, consequently, of additional readmissions, which may be a benefit both for patients and for hospitals. So, this first outcome is important, for it points out an existing *fil rouge* in the readmission of Heart Failure patients for both countries. The mentioned comorbidities, indeed, may represent "universal" information, that can constitute a common basis on which to add specific "local" information.
The effect of some pathologies is different from Lombardia to England. In the Lombardia dataset, for example, *compdiabetes* and *pulmcirc* are significant in most models, while in England their effect is not significantly relevant. They especially give a contribution to the first readmission both in short-run and long-run, increasing significantly the probability of readmission.
In the England dataset, instead, the same behaviour is followed by *alcohol*, *tumor*, *dementia* and *pvd* (which are not relevant for Lombardia population). These comorbidities, indeed, are good predictors in explaining the short-term and long-term readmissions.
These are the pathologies that mostly differentiate the Heart Failure population of Lombardia from the one of England. Enquiring the causes that lead to this different impact could be fascinating as well and stimulating for further explorations.
Comorbidities are quantities associated to the health population (due to geographical or cultural reasons), and they are not directly linked to the efficiency of hospitals in facing Heart Failure readmissions. The competence of hospitals, however, may be analysed through other quantities: the impact of procedures and the timing of admission-discharge-death. Of course, these analyses are not absolute, but they help in gaining a former idea.
The impact of procedures is curious: in England dataset, for example, procedures have a high impact in the short-term readmissions (like in Logistic regression) and in Counting models, while they lose relevance in all Multi-State

models. For example, $CABG$, $PTCA$ and $SHCOK$ increase the probability of readmission in a short term. In a longer perspective, this behaviour involves also ICD, that was a decreasing factor in readmissions within 30 days. In the Lombardia dataset, instead, the procedures are relevant also in the first Multi-State model. The effect of $CABG$, differently from England, decreases the probability of readmissions, while the others procedures are a risk factor for further readmissions, samely as in the UK.
Multi-State models have also show up peculiarities of the trend of the process: we can see that in England the inclination of being readmitted is higher than in Lombardia, while in Italy the probability of dying inside the hospital is higher and the Length of Stay, on average, is longer than in the UK. All these reflections ar valid in the short and in the long term.
We can assert that the use of different models really helps to gain a complete view of the process, and this gives strength to our analysis. Logistic regression, indeed, is simple to implement and allows to fix the time limit for a readmission, but it can't take track of the clinical history of the patients. Hurdle and Zero-Inflated models, instead, are able to supply to the said limit, because they summarize the clinical history of Heart Failure patients, but, summarizing the informations, we lose specific transitions and we shrink the time limits. Multi-State models, in processes related to chronic disease, are really helpful, because they can give much more information on transitions and on the impact of comorbidities in specific transitions. The only limitation lies in computational reason, because the more the model become complex, the less are the covariates that can be inserted (differently from the previous models, which are simpler).
A last limitation is due to the source of our data. This study, indeed, has taken advantage of administrative data. As already known by literature (see Service (2004) and Smith et al. (2004)) administrative data are useful and practical: they do not require excessive costs to be collected, they are regularly updated, they can provide historical informations); however, they present some limitation: the information collected, for example, is restricted to data required for administrative purpose, they may lack contextual background information, they can contain missing data and so on. All these limitations can affect our results, especially when facing medical purposes.
Further studies may be done to explore other aspects of the Heart Failure readmissions, or to prevent the death inside the hospital due to this disease. Moreover, it could be interesting to extend this analysis to other countries as well. An interesting comparison may be done on different countries of the European Union, to highlight differences among the nations that share a similar cultural background. The reason for it to be useful could be traced in a help to focus on the differences of Health Systems. An analysis across countries belonging to different continents, on the other hand, may highlight contrasting features of populations.
Useful would be also to apply this fan of models to other chronic disease (for example, to pulmonary disease, see Bartolomeo et al. (2008)). Another fascinating aspect could be the introduction of other kind of covariates, maybe associated to social or economical factors (for example, related to the social status) and present in administrative data as well. Of course, several other improvements to this work may be done, but our hope is to create a plot outline that could unify different effective methods, already existing and useful for the readmission of Heart Failure patients.

Chapter 6

Code

In this latter section, we report the main parts of **R** code, used to compute our models. We have not written the codes related to the data adjustment, even if they are a considerable part of the work. However, in Chapter 3 we've already described the necessary passages.

Logistic Regression

Function that fits the a general linear model with logit link. The outcome variable is an the indicator of readmission within 30 days, the covariates are the anthropological and clinical variables.

```
load("logistic_first.Rdata")
attach(logistic_first)

log_first<-glm(Y_all~age+sex+INTRAH_days+ICD+CABG+PTCA+SHOCK
               +metastatic+dementia+renal+wtloss+hemiplegia
               +alcohol+tumor+arrhythmia+pulmonarydz
               +coagulopathy+compdiabetes+anemia+electrolytes
               +liver+pvd+psychosis+pulmcirc+hivaids
               +hypertension,family=binomial(link="logit"))
summary(log_first)

detach(logistic_first)
```

Hurdle and Zero-Inflated Models

Functions that fits Counting models. We have reported the Hurdle model with Poisson as counting part and Zero-Inflated model with Negative Binomial as counting part. The outcome is the number of readmissions within a year since the first admission, while the covariates are the summarized anthropological and clinical variables.

```
library(pscl)

h1<-hurdle(n_readm ~ sex+age_first+INTRAH_days+ICD+CABG
           +PTCA+SHOCK+metastatic+dementia+renal+wtloss
           +hemiplegia+alcohol+tumor+arrhythmia
           +pulmonarydz+coagulopathy+compdiabetes+anemia
           +electrolytes+liver+pvd+psychosis+pulmcirc
           +hypertension, dist='poisson')
summary(h1)

z2<-zeroinfl(n_readm ~ sex+age_first+INTRAH_days+ICD+CABG
             +PTCA+SHOCK+metastatic+ dementia+ renal+wtloss
             +hemiplegia+alcohol+tumor+arrhythmia+pulmonarydz
             +coagulopathy+compdiabetes+anemia+electrolytes
             +liver+pvd+psychosis+pulmcirc+hypertension,
             dist='negbin')

summary(z2)

##### number of zeros predicted

round(c("Obs"=sum(n_readm<1),
        "Poisson"=sum(dpois(0,fitted(p1i))),
        "NB"=sum(dnbinom(0,mu=fitted(bn1i),size=bn1i$theta)),
        "Hurdle_Poisson"=sum(predict(h1i,type="prob")[,1]),
        "Hurdle_NB"=sum(predict(h2i,type="prob")[,1]),
        "ZI_Poisson"=sum(predict(z1i,type="prob")[,1]),
        "ZI_NB"=sum(predict(z2i,type="prob")[,1])
))
```

Multi-State Models

Function that fits Multi-State models. Below, the first model: the outcome is the transition between different states, depending on the transition times and flagged by the identity of patients. Furthermore, the related functions to extract quantities of interest.

```
library(msm)

### empirical statestable

statetable.msm(multi_state,ID_paz,data=data)

### matrix of transitions

matrix_IOD<-rbind(c(0,1,1),c(1,0,1),c(0,0,0))
rownames(matrix_IOD)<-c("Alive_In","Alive_Out","Dead")
colnames(matrix_IOD)<-c("Alive_In","Alive_Out","Dead")
```

```
### FIRST multi-state model (plain and with covariates)

msm_1<-msm(state ~ days ,subject=ID_paz,data=data,
           qmatrix=matrix_IOD,gen.inits=TRUE,
           death=3, exacttimes=TRUE,method="BFGS",
           control=list(fnscale=400000))
msm_1


msm_cov_1<-msm(state ~ days,subject=ID_paz,data=data,
               covariates= ~ dementia+renal+dementia
               +wtloss+hemiplegia+alcohol+tumor+arrhythmia
               +pulmonarydz+coagulopathy+compdiabetes
               +anemia+electrolytes+hypertension+pulmcirc,
               qmatrix= matrix_IOD,gen.inits=TRUE,
               exacttimes=TRUE, method="BFGS",
               control=list(fnscale=500000))

### hazard ratio

hazard.msm(msm_cov_1)


### transition probability matrix

p3<-pmatrix.msm(msm_1,t=3,ci="normal")

### mean sojourn time

mean_soj_time<-sojourn.msm(msm_1)

### total lenght of stay

total_los<-totlos.msm(msm_1)

### expected first passage time
efpt<-efpt.msm(msm_1)
```

Below, the second Multi-State model implemented.

```
### SECOND multi-state model (plain)

matrix_IOD<-rbind(c(0,1,0,0,0,0,0,1),
                  c(0,0,1,0,0,0,0,1),
                  c(0,0,0,1,0,0,0,1),
                  c(0,0,0,0,1,0,0,1),
                  c(0,0,0,0,0,1,0,1),
                  c(0,0,0,0,0,0,1,1),
                  c(0,0,0,0,0,0,0,1),
                  c(0,0,0,0,0,0,0,0))
```

```
rownames(matrix_IOD)<-c("1","1a","2","2a",
                        "3","3a","4","Death")
colnames(matrix_IOD)<-c("1","1a","2","2a",
                        "3","3a","4","Death")


msm_2<-msm(new_state ~ days, subject=ID_paz, data=data_msm,
           qmatrix=matrix_IOD, gen.inits=TRUE, method="BFGS",
           control=list(fnscale=400000,maxit=10000000),
           exacttimes=TRUE)
```

Bibliography

D.C Atkins, S.A. Baldwin, C. Zheng, R.J. Gallop, and C. Neighbors. A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychol Addict Behav.*, 27(1):166–177, March 2013.

No authors listed. Heart failure. *Nursing Standard*, 29(15):18–25, 2014.

N. Bartolomeo, P. Trerotoli, A. Moretti, and G. Serio. A markov model to evaluate hospital redamissions. *BMC Medical Research Methodology*, 23(8), 2008.

J. Beyersmann, M. Schumacher, and A. Allignor. *Competing Risks And Multistate Models with R.* Springer, 2012.

A. Bottle, P. Aylin, and D. Bell. Effect of the redamission primary diagnosis and time interval in heart failure patients: analysis of english administrative data. *European Journal of Heart Failure*, (16):846–853, 2014.

A. Buu, R. Lib, X. Tanc, and R.A. Zuckera. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Stat Med.*, 31(29):4074–4086, 2012.

J. Castaeda and B. Gerrits. Appraisal of several methods to model time to multiple events per subject: Modelling time to hospitalizations and death. *Revista Colombiana de Estadstica*, 33(1):43–61, 2010.

M.J. Faddy. Extended poisson process modelling and analysis of count data. *Biometrical Journal*, 39(4), 1997.

G. Grover, A.K. Gadpayle, P.K. Swain, and B. Deka. A multistate markov model based on CD4 cell count for HIV/AIDS patients on antiretroviral therapy (ART). *International Journal of Statistics in Medical Research*, (2):144–151, 2013.

P. Hougaard. Multi-state models: A review. *Lifetime Data Analysis*, (5):239–264, 1999.

M.C Hu, M. Pavlicova, and E.V. Nunes. Zero-inflated and hurdle models of count data with extra zeros: Examples from an hiv-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse*, (37):367–375, 2011.

F. Ieva, C.H. Jackson, and L.D. Sharples. Multi-state modelling of repeated hospitalisation and death in patients with heart failure: the use of large administrative databases in clinical epidemiology. statistical methods in medical research. to appear. 2015.

S. Jackman. pscl: Classes and methods for r developed in the political science computational laboratory. *Journal of Statistical Software*. URL `http://pscl.stanford.edu/`. Department of Political Science, Stanford University. Stanford, California.

C. Jackson. Multi-state modelling with R: the msm package. *MRC Biostatistics Unit, Cambridge, UK*, 2014.

C.H. Jackson. Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, 38(8):1–29, 2011. URL `http://www.jstatsoft.org/v38/i08/`.

E.F Philbin and T.G. DiSalvo. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology*, 33(6):1560–1566, 1999.

D. Postmus, D.J. van Veldhuisen, T. Jaarsma, M.L. Luttik, J. Lassus, A. Mebazaa, M.S. Nieminen, V.-P. Harjola, J. Lewsey, E. Buskens, and H.L. Hillege. The COACH risk engine: a multistate model for predicting survival and hospitalization in patients with heart failure. *European Journal of Heart Failure*, (14):168–175, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

ADLS: Administrative Data Liason Service. Administrative data introduction. 2004. URL `http://www.adls.ac.uk/adls-resources/guidance/introduction/`.

G. Smith, M. Noble, C. Anttilla, L. Gill, A. Zaidi, G. Wright, C. Dibben, and H. Barnes. The value of linked administrative records for longitudinal analysis. *ESRC National Longitudinal STrategy Committee*, 2004.

F. Willekens. *Multistate Analysis Of Life Hystories With R*. Springer, 2014.

A. Zeileis, C. Kleiber, and S. Jackman. Regression models for count data in R. *Journal of Statistical Software*, 27(8), 2008. URL `http://www.jstatsoft.org/v27/i08/`.

E. Zheng. A predictive model for readmission of patients with congestive heart failure: A multi-hospital perspective.