UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

CHU
Centre Hospitalier Universitaire
Dijon Bourgogne

Master's Thesis

A thesis submitted in partial fulfillment of the requirements
for the Master of Science degree in Mathematics, MIGS.

Mathématiques pour l'Ingénierie, alGorithmique, Statistique

Analysis of factors associated with 30-day readmission following initial hospitalization for prostate surgery

Justine Sauce

Date of Submission: August 31, 2023

Accomplished as part of an apprenticeship program between
the *Université de Bourgogne*
and the *CHU Dijon Bourgogne*.

Supervised by

Jonathan Cottenet
Catherine Quantin
Service de Biostatistiques et d'Information Médicale (SBIM)
CHU Dijon Bourgogne

Brice Nafetat
Qualité, Pilotage, Statistiques et Observatoire (QualiPSO)
Pôle Emploi Bourgogne Franche-Comté

Sep. 2022 - Aug. 2023

Preface

Creating a real link between theory and practice was the ultimate goal of the work-study program between *Université de Bourgogne* and *CHU Dijon Bourgogne*. The application and adaptation of statistical models to real situations can be much more tedious than it seems at the first glance. This Master's thesis is intended to provide an overview of both sides from the mathematical rigor required for good statistical modeling to the application, interpretation, and discussion of the results obtained. During my time at university, I acquired the necessary knowledge of statistical modeling to build models properly and ensure their reliability. During my apprenticeship, within the *Service de Biostatistiques et d'Information Médicale* (SBIM) at *CHU Dijon Bourgogne*, I was able to deepen my knowledge and skills, and I discovered the proposals of studies, the drafting of articles and the delicacy of each publication..

This was achieved through a research project conducted in collaboration with the *Service de Biostatistiques et d'Information Médicale* over a period of almost 8 months. The name of this project is « **Analysis of factors associated with 30-day readmission following initial hospitalization for prostate surgery** » and it will be the main focus of this Master's thesis. As with each project proposal within the *CHU Dijon Bourgogne*, the data, materials and methods used are governed by a specific proposal. This thesis will detail all these aspects, in the manner of a scientific publication outline, whilst adding an eccentric twist by revealing the hidden part of the iceberg: the details of the mathematics used. This part is, I believe, the heart and essence of a trust study, which is why it will be given great importance in this thesis.

This Master's thesis is divided into four main parts, one of which is dedicated to presenting the context of the project and the methodology to be adopted according to the proposal. The next one is dedicated to the statistical tools used in this research, summarizing the main mathematical principles behind each SAS procedure used, and examples where these methods were used to better illustrate the purpose of each of them. The third part focuses on the interpretation of the results, the conclusions, a discussion and possible future perspectives to which this study has led. Finally, the fourth and last part recapitulates all the challenges I faced throughout the study.

Acknowledgements

Dear readers,

With this Master's thesis, I turn the page on an intense academic career and begin an exciting new chapter with the start of my doctorate. It is with great pleasure that I take a moment to express my deep gratitude to the people who have contributed to the birth of this Master's thesis. My journey would not have been the same without the generous and constant support of countless people.

I am thinking in particular about the trust shown by the President of the University, Vincent Thomas, in welcoming me to the *Université de Bourgogne*. I would like to thank him warmly for giving me the opportunity to continue my studies in an environment that is so devoted to its students. My most sincere thanks go to my professors and mentors, Hervé Cardot, Catherine Labruère Chazal and many more, who guided me with patience and expertise throughout this academic adventure. Their availability, enlightened advice, and enriching discussions had a profound influence on the quality of this work and encouraged me to pursue research studies.

My gratitude also goes to Jonathan Cottenet for wise guidance and invaluable support throughout the thesis. His ideas and commitment to my work during this last year of the Master's apprenticeship were of primary importance.

Special thanks go to my family and my loved ones, whose unconditional support has been my source of strength and motivation. To them, who pushed me and kept me moving forward, to the way they supported me in every moment of doubt. Your constant encouragement helped me to persevere in the most demanding moments.

Finally, I would like to express my gratitude to each person who, in one way or another, contributed to the completion of this Master's thesis. Your combined efforts have made this phase an achievement of which I am truly proud.

With all my gratitude.

Contents

Appendices

Appendix A

Appendix B

Appendix C

Appendix D

List of Tables

List of Figures

Acronyms

AHRQ Agency for Healthcare Research and Quality. 2

ATIH *Agence Technique de l'Information sur l'Hospitalisation*. 1

CCAM *Classification Commune des Actes Médicaux*. 2–4

CCI Charlson Comorbidity Index. 4, 17

CHU *Centre Hospitalier Universitaire*. 1, 5

CMD *Catégorie Majeure de Diagnostics*. 43

DAS *Diagnostic Associé Significatif*. 2–4

DIM *Département d'Information Médicale*. 1

DP *Diagnostic Principal*. 2–4

DR *Diagnostic Relié*. 2–4

FDep French Deprivation index. 5, 32

FINESS *Fichier National des Établissements de Santé Sanitaires et Sociaux*. 2, 5

GHM *Groupe Homogène de Malades*. 2, 3

GLM Generalized Linear Model. 27

GLMM Generalized Linear Mixed Model. 27

HR Hazard Ratio. 12, 13

ICD-10 International Classification of Diseases, 10th Revision. 2–4

INSEE *Institut National de la Statistique et des Études Économiques*. 2, 5

LMM Linear Mixed Model. 27

MCO *Médecine, Chirurgie, Obstétrique et Odontologie*. 1, 3

PH Proportional-Hazards. 6, 11–13, 15, 16, 18, 23, 24, 26, 39

PMSI *Programme de Médicalisation des Systèmes d'Information*. 1, 2, 4, 5, 37

RSA *Résumé de Sortie Anonymisé*. 2

SBIM *Service de Biostatistiques et d'Information Médicale*. 1

SNDS *Système National des Données de Santé*. 1

SNIIRAM *Système National d'Information Inter-Régimes de l'Assurance Maladie*. 37

Part I

Introduction to the research project

Chapter 1

Framework for the research project

The apprenticeship took place in a research unit at the *Service de Biostatistiques et d'Information Médicale* (SBIM) of the *CHU Dijon Bourgogne*. The subject of the present research is the analysis of factors that may be associated with 30-day rehospitalization following initial hospitalization for prostate surgery, using the *Programme de Médicalisation des Systèmes d'Information* (PMSI) database.

The following sections introduce the CHU and PMSI hospital databases. The aim is to provide the reader with some background on how the PMSI works and to emphasize the anonymous nature of the data used in this study. This is followed by an introduction to the research project and its objectives.

1.1 Research department at the CHU

In France, since 2008, each hospital's budget has depended on the medical activity described in the PMSI, which compiles discharge summaries of admissions. It collects data on all hospital (private or public) admissions in France in order to better manage the financing of healthcare establishments and organize the supply of care.

The organization of PMSI data compilation and processing in the *Médecine, Chirurgie, Obstétrique et Odontologie* (MCO) field is managed by the *Agence Technique de l'Information sur l'Hospitalisation* (ATIH). Prior to the release of (anonymized) medico-administrative data by ATIH, the data is produced and collected within each hospital, thanks to the *Département d'Information Médicale* (DIM), dedicated to managing the collection of this data.

1.1.1 *Service de Biostatistiques et d'Information Médicale* (SBIM)

At the *CHU Dijon Bourgogne*, the DIM is headed by Pr. Catherine Quantin, and complemented by a research unit - the *Service de Biostatistiques et d'Information Médicale* (SBIM) - whose aim is to analyze the CHU's healthcare offer in relation to its environment, and to carry out clinical or epidemiological research based on *Système National des Données de Santé* (SNDS[1]) data. It also participates in various medico-economic studies (e.g. analysis of hospital stay costs). All these missions converge towards better patient care: reduction in readmission/rehospitalization rates, lower mortality, and optimization of hospital stays. The objective is to promote the evaluation of care.

The research project covered by this Master's thesis took place under a research program led by the SBIM, in collaboration with the unit's statisticians and clinicians. The ultimate goal was also to improve patient follow-up through a better understanding of the risk factors associated with 30-day rehospitalization after prostate surgery.

[1] The SNDS brings together the main existing French health databases (including PMSI databases).

1.1.2 Introduction to PMSI data

The information collected in the context of the PMSI is protected by professional secrecy. An anonymous linkage of PMSI information collections has been implemented since 2001 (DHOS-PMSI-2001 circular n°106 of February 22, 2001) thanks to Catherine Quantin. It allows following the hospitalizations of the same patient. The anonymous linkage is based on the creation of a unique anonymous number for each patient. The hospitalizations of the same person can thus be identified but it is impossible to determine the identity of the person from its chain number. In this context, and with a national perspective, activity is recorded in the form of a *Résumé de Sortie Anonymisé* (RSA) produced for each hospitalization (stays and sessions), in each hospital. Each RSA is classified, using a specific classification algorithm, into a single *Groupe Homogène de Malades* (GHM). The classification of all stays in a hospital into GHMs determines the reimbursement rate, since stays classified in the same group have, by construction, similar resource consumption. The classification is also medical, its first level of classification is based on medical criteria (medical procedures or notorious reason for hospitalization).

The RSA contains information on the patient and its stay, both administrative and medical. These include the ***Fichier National des Établissements de Santé Sanitaires et Sociaux* (FINESS) number** (identifying the hospital where the stay took place), ***Numéro d'Index*** (enabling the identification of stays within the same establishment), ***Numéro de Séjour*** (temporal identifier for the stay) and ***Numéro Anonyme*** (unique pseudonymized identifier for each patient), which are identifiers used to order stays for each patient. A RSA includes the ***Diagnostic Principal*** (DP) (the reason the patient was admitted to the unit and/or hospitalized), ***Diagnostic Relié*** (DR) (all conditions that could be related to the principal diagnosis), and ***Diagnostic Associé Significatif*** (DAS) (all complications and comorbidities that could impact the course of the hospitalization), coded according to the World Health Organization's International Classification of Diseases, 10th Revision (ICD-10). It also contains **medical procedures**, coded according to the *Classification Commune des Actes Médicaux* (CCAM). Among the other information available, the following will be used for this study: **geographic code** (postcode), **gender** (consistency of information) and **mode of admission and mode of discharge** (e.g. home, another hospital).

The PMSI will be the main data source used to reconstitute patient characteristics. Concerning environmental factors, we cross-referenced geographical data (limited) from the PMSI with tables from the *Institut National de la Statistique et des Études Économiques* (INSEE[2]).

1.2 Research theme

1.2.1 Study background

Although the risk of rehospitalization has often been studied and is well documented in international publications, the reasons for rehospitalization remain poorly understood. It has already been shown that certain rehospitalizations have a deleterious effect on patients' well-being and lead to a considerable increase in hospital expenditure in the United States [3, 4]. A study by the Agency for Healthcare Research and Quality (AHRQ) estimates that in 2011 there were 3.3 million all-cause re-hospitalizations in the United States within 30 days of hospital discharge, representing $41.3 billion in hospital expenditures. A relevant question is whether hospital readmission can be an indicator of the quality of the health care system. In addition, a better understanding of the factors associated with rehospitalizations would allow us to develop strategies to avoid it.

1.2.2 Research objectives

We attempted to measure the impact of individual and environmental factors that may be associated with rehospitalization, taking into account not only age and length of stay, but also a number of individual clinical factors (e.g. comorbidities, Charlson Comorbidity Index), as well as socio-economic factors measured at an aggregate level (so-called environmental factors) such as a deprivation index, and the public or private status of the hospital. Based on the results of these analyses, we tried to draw up an overview of the situation in order to provide a better understanding of its determinants. Our main objectives were first to assess the impact of individual factors using a Cox proportional-hazards model. Then, using multilevel logistic regression, we investigated the effects of environmental (socio-economic) factors on the risk of rehospitalization at 30 days.

[2] INSEE is a public organization responsible for producing, analyzing and publishing official statistics on the French economy and society.

Chapter 2

Methodology

The purpose of this chapter is to clarify matters, starting with the definition of the inclusion criteria. Then we'll discuss the factors that should be considered, and finally we'll look at the statistical methods that can be considered for the analyses.

2.1 Inclusion and exclusion criteria

We considered patients over 18 years old, hospitalized for prostate surgery between January 2012 and November 2014, with the aim of tracking 30-day rehospitalizations. Patients included in the study must have one of the inclusion pathologies, referenced as DP, DR or DAS, and must have undergone one of the surgical procedures specified in the proposal, both identified respectively by the ICD-10, and CCAM codes listed in Appendix A.1.

Patients who had undergone prostate surgery in the year (365 days) prior to the initial, also called index, hospitalization were excluded from the study. Deaths recorded during the index hospitalization were excluded due to their extremely low percentage of 0.16% (449 deaths out of 275189 hospitalizations) of cases. Only MCO stays with home admission modes were selected, excluding stays and sessions corresponding to specific and/or iterative treatments. Iterative stays are identified by a specific GHM, the list of which is given in Appendix A.1.3. Figure 2.1 below, gives an overview of the flowchart of inclusions.

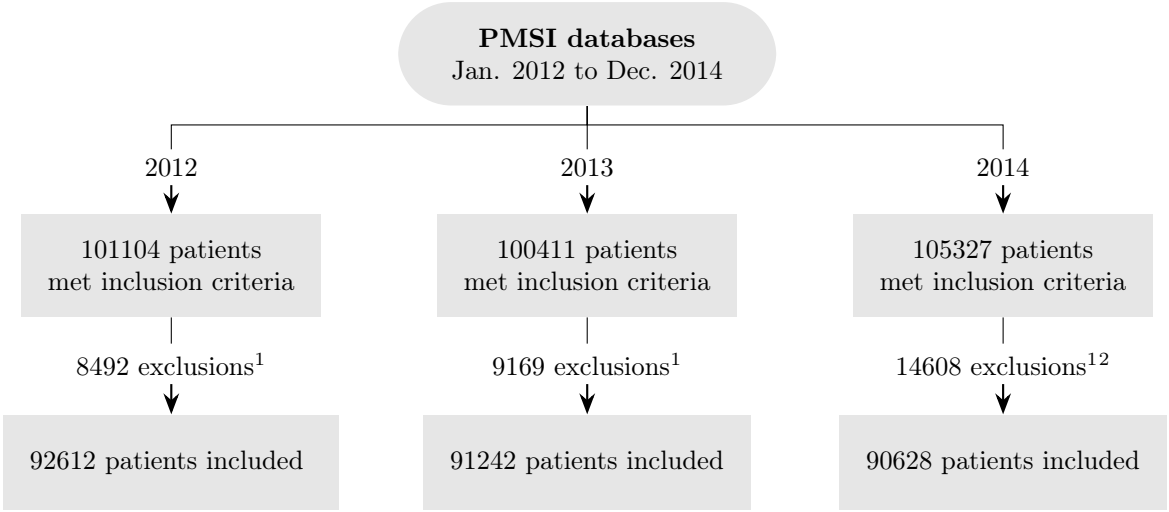

Figure 2.1: Flowchart of inclusions by year

Finally, 4009 duplicate stays were excluded, keeping only the patient's first hospitalization. According to these inclusion criteria, the study will be carried out on **270473 patients**.

[1] Already been hospitalized for the condition in question within 365 days, hospitalization not coming from home, deaths.
[2] Hospitalizations during December 2014.

2.2 Materials

We worked with SAS throughout the study, both to manipulate and analyze PMSI data. The construction of the outcome and new variables was necessary as the information is available but not coded appropriately for the study. A detailed explanation of the construction process for these variables is provided in the following subsections.

2.2.1 Outcome: 30-day rehospitalization

In the PMSI databases, we can find the start of the stay, identified by *Numéro de Séjour*, and the end of the stay, obtained by adding the length of stay to the *Numéro de Séjour*. Once the start and end of the stay have been correctly identified, we calculated the time elapsed between two hospitalizations. See Appendix A.2 for a schematic example and further explanation.

This enabled us to access two key pieces of information:

- The **outcome**: whether or not the patient was rehospitalized within 30 days.
- The time to rehospitalization, called Delay, which was used as a **time-to-event** variable, allowing us to apply survival models.

To construct the outcome, we constructed a variable coded 1 when the delay was between 1 and 30 days, and 0 otherwise. Only the time between the first rehospitalization and the index hospitalization was taken into account, as specified in the proposal.

The delay was introduced in the form of a new variable called `Delay`, corresponding to the time elapsed between the index hospitalization and the following one.

2.2.2 Individual factors (patient characteristics)

Concerning individual factors, we considered age and length of stay, the construction of the Charlson comorbidity index, commonly used when analyzing 30-day rehospitalization [20, 23]; and after studying the literature [24, 19], the Elixhauser index seemed a worthwhile idea to consider.

The **Charlson Comorbidity Index (CCI)** is the scoring system most widely used by researchers and clinicians to measure comorbidities. This measurement tool provides a weighted score of a patient's comorbidities by considering the level of severity of 17 (19 depending on the definition) predefined comorbid disorders, as well as the number of disorders present among them. Based on a similar principle, but with an approach focusing more on the number rather than the weight of comorbidities, we can define the **Elixhauser index**. The index is defined by 31 comorbidity groups, weighted uniformly. For both indices the weights are summed, the higher the score, the higher the expected hospital resource utilization and mortality rate, which is why they are useful in a rehospitalization study. They were built up using ICD-10 codes; the list of groups, associated codes and their weighting is given in Appendix A.3. These pathologies will be sought on patients' DP, DR and DAS recorded at index hospitalization.

Finally, for exploratory investigation we built a variable distinguishing different **ranks of cancer** at inclusion. The ranks considered were constructed on the basis of ICD-10 code referencing. Three groups were defined. A moderate rank group for benign prostatic hyperplasia and low-grade prostatic dysplasia. A high rank group, covering carcinoma in situ of the prostate (high-grade dysplasia), benign tumour of the prostate and unpredictable tumour (or tumour of unknown evolution) of the prostate. And the last group - very high rank - concerns exclusively malignant tumours of the prostate. In addition, a variable distinguishing **medical procedures** undergone at inclusion was constructed. It is based on CCAM groups. Three groups were also considered: anaesthetic procedures, surgical procedures and technical procedures. These partitions, as well as all the codes sought, are detailed in Appendix A.1.

2.2.3 Environmental factors

The second line of research focused on environmental factors. The aim was to include potential variations between care in different hospitals or different localities.

The first step was to construct an **urban/rural status indicator** for the household. The INSEE has categorized each commune as rural or urban on the basis of 2012 population census data. To define the degree of urbanization, INSEE has classified as urban units those communes or a set of communes comprising a continuous built-up area ($\leq$ 200m between two constructions) inhabited by at least 2,000 people. Communes with such urban units are known as "Urban municipalities", and other municipalities as "Rural municipalities". By cross-referencing INSEE tables with the geographic codes available in the PMSI tables, we were able to reconstruct the urban or rural status of the included patients' household.

In a second step, we also reconstituted an index of deprivation, using the **French Deprivation index (FDep)**. The FDep index was created to provide a general population geographical indicator of social disadvantage specifically adapted to health studies on the French population. The FDep index was defined as the first component of the principal component analysis (PCA) (67-70% of the total variance explained, depending on the period) of the following four variables: *median income per consumption unit in the household, rate of baccalaureate holders in the non-educated population aged 15 and over, unemployment rate in the working population aged 15 to 64, and rate of blue-collar workers in the working population aged 15 to 64*. This is applied at commune level, for more aggregated scales such as the postcode, department or region, the population-weighted average of the score for the commune is used. We used the 2009 references for 2012 and the 2013 references for 2013-2014. In addition, most studies [15, 9] divide the FDep index into quintiles, which makes it possible to choose the middle class (around 0) as the reference.

The final step, concerning environmental factors reconstitution, consisted in identifying the **public or private status** of the hospital where the index hospitalization took place. For this purpose, we used the FINESS number as a basis, and cross-reference it with the type of hospital for each stay; public status corresponding to public hospitals (e.g. *Centre Hospitalier Universitaire* (CHU)), and private status was for private for-profit/nonprofit hospitals.

Note: It was not possible to reconstitute environmental factors for all patients, due to missing or incomplete geographic codes. We obtained 8.03% of missing values.

2.2.4 Summary of factors and their coding

The report contains code extracts, hence the Table 2.1 below is provided for guidance. It illustrates the initial modalities of the factors and their label in code examples.

Variable	Label	Type
`RH30`	Rehospitalization within 30 days	Binary (0: No rehospitalization, 1: Rehospitalization)
`Delay`	Time-to-event (delay)	Quantitative in days
`Age`	Age	Quantitative in years
`LOS`	Length of stay	Quantitative in days
`CCI`	Charlson Comorbidity Index	Quantitative index
`ELX`	Elixhauser Comorbidity Index	Quantitative index
`CR`	Cancer Rank	Qualitative 3-level (Moderate, High, Very High)
`MP`	Medical Procedure	Qualitative 3-level (Technical, Anesthesia, Surgery)
`Rur`	Rural status	Binary (0: Urban municipalities, 1: Rural municipalities)
`FDep`	French Deprivation index	Quantitative index
`Pub`	Public status	Binary (0: Private hospital, 1: Public hospital)

Table 2.1: Label and type of variables constructed for each factor

2.3 Statistical methods

We can divide this project into three main lines of research.

- **Univariate Analysis:** Association of factors with rehospitalization.
- **Multivariate Analysis:** Association of individual factors with the risk of rehospitalization by multivariate Cox regression analysis.
- **Multilevel Analysis:** Association of individual and environmental factors with the risk of rehospitalization by multilevel logistic regression.

The first line of research, corresponds to the initial descriptive phase of the factors considered for this study. Each factor corresponds to a possible covariate, and we needed to know how these covariates behave, study them and analyze them. Analyses of the distribution of quantitative and qualitative covariates according to their modality; and tests dedicated to the study of relationships between covariates and with the outcome, were carried out. Univariate analysis was essentially based on **Chi-square** and **Student's t-tests**.

The aim of the second line was to propose a model, adapted to survival data, enabling us to estimate and characterize the association between outcome and the individual factors selected in the first line of research.

In epidemiology, we are often led to describe and identify the factors associated to an event, and this is the aim of this research project. The outcome has two modalities: occurrence or non-occurrence of the event studied, in this case, whether or not the patient is re-hospitalized within 30 days. Two points of view can be adopted in this type of research. We can be interested in the probability of occurrence of the event, or in the probability of not yet having experienced the event. Given the binary character of the outcome, these two points of view led us to consider two main approaches to statistical analysis: Logistic regression and the **Cox Proportional-Hazards model** (Cox PH model), also called Cox regression.

The logistic regression is designed to describe and identify the factors associated to the event whatever the time at which it occurs. This model gives a first idea of the path to follow in the search for factors. This is why we generally start with logistic models as a basis
To take into account time-to-event and the right censoring of the data, we have to consider a survival analysis such as the Cox proportional-hazards model. However, Cox regressions are more complicated to fit than logistic regressions, because of the proportional-hazards assumption, which will be detailed later. As this model is the most appropriate for the data used in this study and for analyzing the risk of rehospitalization at 30 days,we will limit this Master's thesis to detail and focus on interpreting the results of the Cox proportional-hazards model for individual factors.

In order to be more precise on the question of patients' risk of 30-day rehospitalization, it may be useful to take into account environmental factors. We did not seek to measure these effects, but we did want to take into account potential correlations, for example due to socio-economic conditions between patients. This was the subject of the last line of research, focusing on the construction of **Multilevel Logistic regressions**.

Part II

Survival Analysis

Chapter 3

Selection of relevant covariates

It is generally admitted that the construction of multivariate models must be guided by a thorough knowledge of the subject. A study must be carefully planned, guided by the research questions and the methods envisaged for data analysis. Since knowledge of the subject is often limited or, at best, fragile, it is necessary to build models based on the available data.

One of the key questions in covariate selection is which of the covariates should be included in the model. Virtually all statistical software contains covariate selection procedures. This has made their use very popular, especially with end-users who have no formal training in statistics. However, this widespread availability has been a breeding ground for many misunderstandings about the role and necessity of covariate selection.

The other key point to remember is balance. The statisticians need to reach a balance by including the correct number of covariates in the regression equation:

- **Too few:** Underspecified models tend to be biased.
- **Too many:** Overspecified models tend to be less precise.
- **Just right:** Models with the correct terms are not biased and are the most precise.

To achieve this balance, we carried out univariate analyses to ensure the relevance of the chosen covariates, and also used classification, which as we shall see, enabled us to filter out certain covariates; these methods are presented in the following sections.

3.1 Univariate Analysis

After an initial descriptive analysis of the covariates, using FREQ and MEANS procedures, we aim to test the association between the outcome and one covariate at a time.

Univariate analyses were performed to filter "by hand" the covariates to be included in the model. We kept those for which the association with the outcome is sufficiently strong, with the association considered significant for a p-value under 0.20. A threshold of 5% (p-value under 0.05) would be far too strict in this initial variable selection stage (D. Commenges and H. Jacqmin-Gadda, 2015, p.96 [6]).

The two statistical tests used are those mentioned in the methodology, the Chi-square test and the Student's t test. In fact, given the binary nature of our outcome, and the data available, these two tests proved to be the most suitable for carrying out the initial association analyses. The assumptions required for their application were easily verified thanks to the large number of patients included in the study and their distribution. Moreover, observations are independent, since one observation corresponds to exactly one patient.

The following subsections present the tests and associated SAS procedures, and justify their correct use in this study.

3.1.1 Pearson Chi-square test

Factors can be qualitative variables such as the presence of other pathologies coded by 0 (absence of pathology) or 1 (presence of pathology). This involves comparing two binary variables together (outcome and factors); in order to compare those kinds of variables, the statistic we used is the Pearson Chi-square (χ^2) one.

The Pearson Chi-square statistic is computed as,

$$\chi^2 = \sum_{i}^{N} \sum_{j}^{P} \frac{(n_{ij} - e_{ij})}{e_{ij}}$$

where, N corresponds to the number of modalities of the first variable and P to the number of modalities of the second variable; n_{ij} is the observed cell count in the ith row and jth column of the table; $n_{i.}$ represents row totals and $n_{.j}$ column totals and e_{ij} is the expected cell count in the ith row and jth column of the table, computed as,

$$e_{ij} = \frac{(n_{i.} n_{.j})}{n}.$$

Under the null hypothesis that the N row and P column variables are independent, χ^2 has an asymptotic Chi-square distribution with $(N-1)(P-1)$ degrees of freedom.

The assumptions underlying the use of this statistical test have been verified in our study and are listed down below:

- **Categorical variables:** The Chi-Square Test of Independence determines significant associations between 2 categorical variables.
- **Simple random sample:** The test assumes data is obtained from a random sample.
- **Mutually Exclusive Categories:** Variable categories must be mutually exclusive. This means that each subject fits into one and only one level of each variable.
- **Single Data Contribution:** Each patient may contribute to one and only one cell in the Chi-Square test.
- **Sample size:** The sample size is assumed to be sufficiently large. If a chi-square test is performed on too small sample size, the chi-square test will produce an invalid inference.
- **Expected cell count:** The expected frequency in each cell should be five or more in at least 80% of the cells. If this assumptions is not verified, we can attempt a Fisher's exact test[1].
- **Independence:** The observations are always assumed to be independent of each other.

Finally, it's pretty simple to run the Chi-square test with SAS, we just used the FREQ procedure with '`chisq`' option, specified in the TABLES statement. The syntax of the code evaluating the association, between cancer rank (`CR`) and the outcome: rehospitalization within 30 days (`RH30`), by using a Chi-square test, is shown opposite.

```
PROC FREQ DATA=Table;
    TABLES RH30*CR / chisq;
RUN;
```

The FREQ procedure computes several chi-squared tests for each two-dimensional table, the output statistic called Chi-square is the one corresponding to the Pearson Chi-square test.

[1] If $> 20\%$ of the cell frequencies are < 5, SAS will print a warning, and we should not use the chi-square test. Instead, we use the Two-sided Fisher's Exact Test (printed by default when the table is 2 x 2)

3.1.2 Student's *t*-test

On the other hand, factors can be quantitative variables such as age or length of stay, coded in a continuous way. This involves comparing a binary variable with a continuous variable. For this purpose, we used a Student's independent samples t-Test. It compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different.

In the case of an independent samples t-test, the degrees of freedom are calculated based on the sample sizes of both samples (n_1 and n_2). The formula is $df = n_1 + n_2 - 2$. The t-statistic to test whether the means are different can be calculated, using empirical mean, as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where,

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}}$$

is the pooled standard deviation of the two samples, using the unbiased estimators of the two population variance. It is designed so that its square is an unbiased estimator of the pooled variance, whether or not the population means are exactly the same.

Please note that the use of this statistic and this estimator of the pooled standard deviation must be preceded by a verification of the following assumptions.

- **Normality:** The data should be approximately normally distributed. While t-tests are considered robust to moderate deviations from normality, severe violations can have an impact on the accuracy of test results. In the case of very large samples, normality is assumed.
- **Independence:** The observations in the samples must be independent of each other.
- **Homogeneity of Variances:** For the independent samples t-test, the variances of the two compared populations must be equal or at least approximately equal. This assumption is commonly known as the homogeneity of variances. If this assumption is not met, other tests can be used, such as Welch's t-test, which does not require equal variances.

Within the scope of our study, theses assumptions are met for the continuous variables we wished to test; allowing us to apply the most basic t-tests and t-statistics to our research and data analysis. In any cas, SAS' TTEST procedure ensures that equality of variances is respected by running a test of variance homogeneity called '`Folded F`', and returns the results of the t-test in the event of equality and of an adapted t-test in the event of variance inequality.

The null hypothesis of the '`Folded F`' test is that the variances are equal; the alternative is that the variances are not equal. This test is useful to determine which output we'll rely on: '`Pooled`' (equal variance assumed) for a large p-value or '`Satterthwaite`' (equal variance not assumed) for a small p-value. The test statistics and formulas used are those commonly introduced, and can be found in the SAS Help Center, in the TTEST procedure details section.

An example of code testing the association between rehospitalization within 30 days (`RH30`) and patient age (Age) is shown below.

```
PROC TTEST DATA=Table;
    CLASS RH30;
    VAR Age;
RUN;
```

In this study, equality of variances was validated for each continuous covariate, so we relied only on pooled results. For this reason, we don't go back over the statistics and tests adapted for variance inequality, but SAS help is precise and detailed on this subject.

3.2 Classification

Sometimes it can be useful to filter out covariates and select those that best explain our results. In our study, we chose to consider only certain comorbidity groups, using Elixhauser's predefined comorbidity groups. This choice was the consequence of too strong correlations between the index itself and the other covariates, it will be discussed in section 5.3 dedicated to correlations and in detail in section 11.2.1. Thus, we needed to know which comorbidity groups were the most influential.

A well-known method is to classify variables using a decision tree, which highlights the most important variables. Decision tree algorithms use data to divide the set of all possible combinations of covariate values, into non-overlapping regions. These regions correspond to the end nodes of the tree. Each region is described by a set of rules, which are used to assign a new observation to a particular region. In the case of a classification tree, the predicted value for an observation is the most frequent level of the response variable in that region. By construction, the first nodes of the tree are made up of the most important covariates.

The HPSPLIT procedure is adapted for this purpose, it is a high-performance SAS procedure that builds tree-based statistical models for classification and regression. The syntax for building a classification-tree for `RH30` prediction from Elixhauser's 31 comorbidity groups (`ELX GRP 1`, `ELX GRP 2`, ..., `ELX GRP 31`) is as follows.

```
PROC HPSPLIT DATA=Patients;
    CLASS RH30;
    MODEL RH30 = ELX_GRP_1 ELX_GRP_2 ... ELX_GRP_31;
RUN;
```

The MODEL statement specifies `RH30` as the outcome and the variables to the right of the equal sign as the covariates. The inclusion of `RH30` in the CLASS statement designates it as a categorical outcome and requests a classification tree.

The following Table 3.1 is a short version of the variable importance table obtained from the classification of the 31 comorbidity groups (before any exclusion/correlation analysis).

		Learning	
Covariate	**Label**	**Relative**	**Importance**
ELX GRP 2	Cardiac Arrhythmia	1.0000	15.8256
ELX GRP 20	Solid Tumor without Metastasis	0.6827	10.8047
ELX GRP 14	Renal Failure	0.6491	10.2726
ELX GRP 1	Congestive Heart Failure	0.4014	6.3526
ELX GRP 19	Metastatic Cancer	0.3966	6.2766
ELX GRP 11	Diabetes Uncomplicated	0.3322	5.2575
ELX GRP 3	Valvular Disease	0.3309	5.2367
ELX GRP 25	Fluid and Electrolyte Disorders	0.1560	2.4682
ELX GRP 27	Deficiency Anemia	0.1372	2.1718
ELX GRP 26	Blood Loss Anemia	0.1356	2.1462
ELX GRP 16	Peptic Ulcer Disease excluding bleeding	0.1129	1.7868
ELX GRP 6	Hypertension Uncomplicated	0.0953	1.5076

Table 3.1: Classification tree for 12 Elixhauser comorbidity groups

This gives us an initial idea of the comorbidities that can be considered as covariates. We'll keep just 10 out of 31 groups, to ensure that the model is not too complex without missing something. We need to analyze correlations before including them in a model. This correlation analysis will be discussed again in chapter 5 and the associated challenges in section 11.2.1. In what follows, we focus on model selection and fitting.

Chapter 4

Cox proportional-hazards model

Once the covariates had been identified, we moved on to building the model itself. Readers are invited to refer to Appendix B.1 if necessary for a brief introduction to survival data, their characteristics and some of the basics needed to better understand the tools that will be presented here. As explained in the methodology (section 2.3), for the analysis of individual factors, we chose to consider a Cox proportional-hazards model. The semi-parametric version of the proportional-hazards regression model, proposed by Cox (D.R. Cox, 1972 [7]) and commonly called Cox PH model, is widely used in epidemiology to assess the effect of covariates and to analyze survival data with right censoring, because it takes into account the time-to-event.

4.1 Model specification

Let y_i be the outcome observation y for a patient i, $i = 1, \ldots, n$. For each patient i we observe the vector of p covariates $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ associated with this individual. The p covariates, $X_1, \ldots, X_p$, are considered deterministic and can be quantitative or qualitative. The design matrix associated is defined as $X = (1|X_1|\ldots|X_p) \in \mathcal{M}_{n,p}(\mathbb{R})$.

In a Cox proportional-hazards regression model, the effect measure is the **hazard rate**, which represents the risk of the event. The relationship between the hazard associated with the occurrence of an event and the vector of p covariates is as follows:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) \tag{4.1}$$

$$= \lambda_0(t) \exp(X_i^T \beta) \tag{4.2}$$

where $\lambda_0(t)$ is the baseline hazard, and the p-vector β contains the regression parameters.

For the purposes of this study, we will use the following notation for the **survival function** (4.3) and the **cumulative risk function** (4.4).

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right) \tag{4.3}$$

$$A(t) = \int_0^t \lambda(u)du \tag{4.4}$$

The β vector is originally estimated by maximizing a **partial likelihood** (4.5) proposed by Cox (D.R. Cox, 1975 [8]),

$$\mathcal{L}(\beta, X) = \prod_{i=1}^{p} \frac{\exp(X_i^T \beta)}{\sum_{l:\tilde{T}_l \geq t_i} \exp(X_l^T \beta)} \tag{4.5}$$

where, $t_1 < t_2 < \cdots < t_k$ are the different event times observed and $\tilde{T}$ is the survival time random variable for right-censored data.

4.2 Hazard ratio

There is a similarity between Cox's **hazard ratio** (HR) and the odds ratio in logistic regression, the difference being that the hazard ratio must be seen in terms of instantaneous risk. By definition, all subjects will experience the event if the follow-up time is long enough. We therefore interprete a group more at risk than another (HR > 1) if the time-to-event is shorter than for the other group and *vice-versa*.

Given a model with p covariates, the model is written as the relation (4.1) given above. We then have,

$$\mathrm{HR}_k(t) = \exp(\beta_k)$$

which is the hazard ratio associated with the covariate X_k adjusted on all other covariates, representing the risk of one patient compared with another differing only through the value of the covariate X_k.

Another essential point to address is the confidence interval associated with the hazard ratio. Since the estimators of the regression parameters are distibuted asymptotically according to a normal distribution, the construction of the 95% confidence interval for the regression parameter β_k is $\left[\hat{\beta}_k \pm 1.96 \times \hat{\sigma_k}\right]$ where $\hat{\beta}_k$[1] is the estimate of β_k and $\hat{\sigma_k}$ the estimate of σ_k. However, we are interested here in the **95% confidence interval of the hazard ratio**. Since this is equal to the exponential of the regression parameter, we obtain the following interval $\left[\exp(\hat{\beta}_k - 1.96 \times \hat{\sigma_k}); \exp(\hat{\beta}_k + 1.96 \times \hat{\sigma_k})\right]$. (D. Commenges and H. Jacqmin-Gadda, 2015, p.134-135 [6])

These two values are our main focus in this study. Their calculation is included in the dedicated SAS PHREG procedure, which we'll take a look at in the next section; before moving on to application conditions.

4.3 PHREG procedure

To perform a regression analysis of survival data based on the Cox proportional hazards model, SAS provides the PHREG procedure. This procedure fits the Cox PH model by maximizing the partial likelihood and computes the baseline survivor function by using the Breslow estimate expressed as follows (D.Y. Lin, 2008 [13]).

$$\hat{\mathcal{L}}(\beta, X) = \prod_{i=1}^{p} \frac{\exp(X_i^T s_i)}{\left[\sum_{l:\tilde{T}_l \geq t_i} \exp(X_l^T \beta)\right]^{m_i}}$$

where s_i is the vector of the sum of the covariate vectors of the m_i patients who experienced the event at time t_i.

In the case of many *ex-aequo* we can also use the Efron method by specifying it in the SAS PHREG procedure. However, this method much more time consuming.

The syntax of this procedure is as follows.

```
PROC PHREG DATA=Patients
    MODEL Delay*RH30(0) = Age CR / rl;
RUN;
```

In the MODEL statement, the variable `Delay`, is crossed with the censoring variable, `RH30` (the outcome), with the value that indicates censoring is enclosed in parentheses. The values of `Delay` are considered censored if the value of `RH30` is 0; otherwise, they are considered event times.

SAS output will automatically produce a table displaying individual model effects and a table about hazard ratios and estimates. We can use the HAZARDRATIO statement to obtain the hazard ratios for a main effect in the presence of interaction (see Chapter 6) and the option '`rl`' in the MODEL statement to obtain risk limits, which represents the 95% confidence interval of the HR.

[1] Parameter estimates will be marked with a hat.

4.4 Assumptions underlying

Let us return to an essential point of the Cox PH model, which will dictate a major thrust of this thesis. The Cox model makes no assumptions about the shape of the baseline risk - it is said to be freely variable - and focuses on the regression parameters. On the log-scale the relation (4.1) becomes,

$$\log(\lambda_i(t)) = \log\left[\lambda_0(t)\exp(\beta_1 X_{i1} + \beta_2 X_{12} + \cdots + \beta_p X_{ip})\right] \tag{4.6}$$
$$= \log(\lambda_0(t)) + \beta_1 X_{i1} + \beta_2 X_{12} + \cdots + \beta_p X_{ip} \tag{4.7}$$

In other words, the Cox PH model assumes that the effects of covariates are additive and linear on the log rate scale. This is just one of the two fundamental assumptions of the Cox PH model. The construction of an appropriate Cox PH model depends on the assumptions being met.

Methods for assessing the proper construction of a Cox PH model are an important part of any statistical analysis, as investigators can be seriously misled if conclusions are drawn on the basis of erroneous assumptions. The process of examining adequacy is a combination of graphical representations and more formal hypothesis testing. In what follows, we'll briefly present the hypotheses and stress the importance of verifying them. We'll present various methods for checking and overcoming the difficulties encountered, depending on the type of covariates.

4.4.1 Loglinearity

The loglinearity assumption is the one shown by the relation (4.6). The covariates must behave **linearly on a log-scale**. Depending on the type of covariate (quantitative or qualitative), the hypotheses will not be stated in the exact same way, even if the principle remains identical. We will therefore distinguish this section according to the type of covariate in order to offer greater clarity and examples.

4.4.1.1 Transforming qualitative covariates

In the case of qualitative variables, the logarithm of risk should increase by $\beta_k \in \mathbb{R}$ as we move from one class to the next. When covariates are binary-coded, the problem doesn't arise (D. Commenges and H. Jacqmin-Gadda, 2015, p.73 [6]). By construction, going from level 0 to level 1 increases by exactly β_k. However, as soon as the covariate has more than 2 levels, the question remains as to how to estimate a HR between the first level and a level at least two levels higher.

Let's look at the HR estimate for the prostate cancer rank covariate, we'll note β_{CR} the regression parameter associated to this covariate. According to ordinal coding, this covariate is coded as follows: (0) for moderate rank, (1) for high rank and (2) for very high rank.

$$\text{HR}_{(1 \text{ vs } 0)} = \frac{\lambda_{\text{High}}(t)}{\lambda_{\text{Moderate}}(t)} = \frac{\lambda_0(t)\exp(\beta_{\text{CR}}(1))}{\lambda_0(t)\exp(\beta_{\text{CR}}(0))} = \exp(\beta_{\text{CR}})$$
$$\text{HR}_{(2 \text{ vs } 0)} = \frac{\lambda_{\text{Very High}}(t)}{\lambda_{\text{Moderate}}(t)} = \frac{\lambda_0(t)\exp(\beta_{\text{CR}}(2))}{\lambda_0(t)\exp(\beta_{\text{CR}}(0))} = \exp(2\beta_{\text{CR}})$$

On the log-scale, this gives us,

$$\log(\lambda_{\text{High}}(t)) - \log(\lambda_{\text{Moderate}}(t)) = \log(\lambda_0(t)\exp(\beta)) - \log(\lambda_0(t)) = \log(\lambda_0(t)) + \beta_{\text{CR}} - \log(\lambda_0(t)) = \beta_{\text{CR}}$$
$$\log(\lambda_{\text{Very High}}(t)) - \log(\lambda_{\text{Moderate}}(t)) = \log(\lambda_0(t)\exp(2\beta)) - \log(\lambda_0(t)) = \log(\lambda_0(t)) + 2\beta_{\text{CR}} - \log(\lambda_0(t)) = 2\beta_{\text{CR}}$$

Thus, moving from level (0) - Moderate rank - to level (2) - Very High rank - increases the logarithm of risk by $2\beta_{\text{CR}}$. In general terms, this is not what we want since loglinearity assumes a variation of β_k between each level of the covariate X_k. On the other hand, by using a binary coding for qualitative covariates, i.e. by introducing several dummy variables to represent each level of the covariate, we will keep a comparison at one level and thus an increase of β_k per level.

For example, with the prostate cancer rank covariate, the aim is to estimate the HR between three groups.

$$\text{CR} = \begin{cases} 0 & \text{Moderate rank} \\ 1 & \text{High rank} \\ 2 & \text{Very High rank} \end{cases} \qquad \text{CR}_1 = \begin{cases} 1 & \text{if CR} = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{CR}_2 = \begin{cases} 1 & \text{if CR} = 2 \\ 0 & \text{otherwise} \end{cases}$$

In this context we only need two dummy variables, because if these two dummy variables are zero, the patient will be diagnosed with a moderate rank of prostate cancer, which is then the reference. Including only the dummy variables created to replace the ordinal covariate, the model can be written as follows.

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 \text{CR}_1 + \beta_2 \text{CR}_2)$$

One question that may arise at this point; **Do we have to create dummy variables for each level instead of keeping categorical coding?** In fact, doing this will enable to test our loglinearity hypothesis for the categorical coding. All we have to do is compare the AICs of the model built with categorical coding and the model built with dichotomous coding. If both are identical, there's no need to introduce dichotomous coding: we can keep the categorical coding. Otherwise, we'll have to work with dichotomous coding.

The advantage of SAS is that it does the dichotomous coding itself when using PHREG procedure via the CLASS statement. In PHREG, the levels of the categorical variables are determined by the CLASS statement. In the following instructions, the CR variable is declared as a class variable in the CLASS instruction. Parameterization is used via the '`(ref='Moderate')`' option, giving the model the Moderate rank as a reference

```
PROC PHREG DATA=Patients
    CLASS CR (ref='Moderate');
    MODEL Delay*RH30(0) = Age CR;
RUN;
```

The class variable CR generates two dummy variables as covariates and two regression coefficients are estimated for the CR covariate levels, as if we were using CR_1 and CR_2. Therefore this is equivalent to,

```
PROC PHREG DATA=Patients
    MODEL Delay*RH30(0) = Age CR1 CR2;
RUN;
```

4.4.1.2 Conditions on quantitative covariates

When dealing with quantitative covariates, loglinearity assumes that a change of one unit in the continuous covariate must have the same effect on the event under consideration, no matter what value we start with.

In our study, an obvious quantitative covariate to consider is age. Besides, it is often used to illustrate hypothesis violation, since quantitative coding will imply a regular variation with the risk of rehospitalization, which is potentially not always the case.

Having the following model,

$$\lambda(t) = \lambda_0(t) \exp(\beta \times \text{AGE})$$

On a logarithmic scale, this is equivalent to,

$$\log(\lambda(t)) = \log(\lambda_0(t)) + \beta \times \text{AGE}$$

For all times t, this is a straight line with intercept $\log(\lambda_0(t))$ and slope β. The logarithm of the rate increase (or decrease depending on β) for each unit increase in the age covariate. Thus, coding a covariate as a continuous variable presupposes loglinearity, using, for example, a comparison of two patients aged X and $X+1$ years (difference of 1 year) involves,

$$\mathrm{HR}_{1\ \mathrm{year}} = \frac{\lambda_0(t)\exp(\beta(X+1))}{\lambda_0(t)\exp(\beta X)} = \exp(\beta)$$

while for a 10-year difference we get,

$$\mathrm{HR}_{10\ \mathrm{year}} = \frac{\lambda_0(t)\exp(\beta(X+10))}{\lambda_0(t)\exp(\beta X)} = \exp(10\beta)$$

On a logarithmic scale, this implies that for a jump of 10 years, there is also a jump of 10 years in the regression parameter (β), indicating linear behaviour. Therefore, in order to incorporate a quantitative covariate into the model, we must first ensure that its behavior is linear with β on the log-scale.

To do so, there are a relatively simple method which is very easy to implement, and gives a visual idea of linearity. Indeed, if the covariate respects loglinearity, the plot of the intercept and slope $\hat{\beta}$ should be a straight line. The aim here is to get an idea of its behavior using a dozen points and the associated $\hat{\beta}$ estimates. To do this we propose to construct a 10-level categorical covariate using the deciles of our quantitative covariate (D. Commenges and H. Jacqmin-Gadda, 2015, p.74-75 [6]). We then fit a Cox PH model with the categorized covariate, producing a coefficient for each level. We'll plot $\hat{\beta}$ estimates against the midpoints of each covariate level, with $\hat{\beta} = 0$ for the reference category (Reference). If the graph reflects a linear line, we can assume the loglinearity assumption and thus use the covariate in a Cox PH model as a quantitative covariate.

Example: We had to test this hypothesis for several covariates, including age and length of stay. Figures 4.1 and 4.2 are example of the graph obtained by this first method. To give an idea, the code associated with the covariate Age is shown in the Appendix B.2.1.

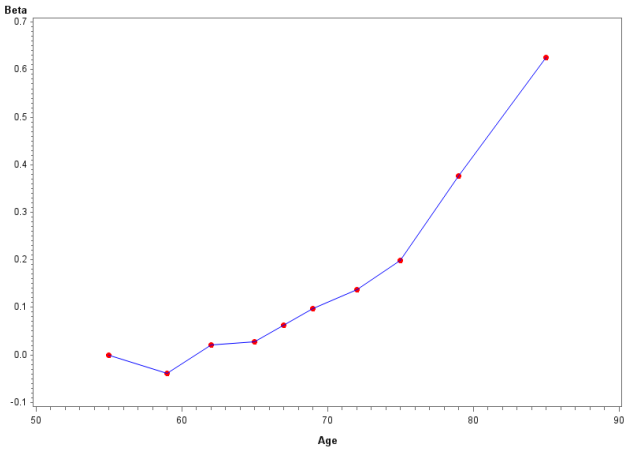

Figure 4.1: Beta estimates vs. Age covariate

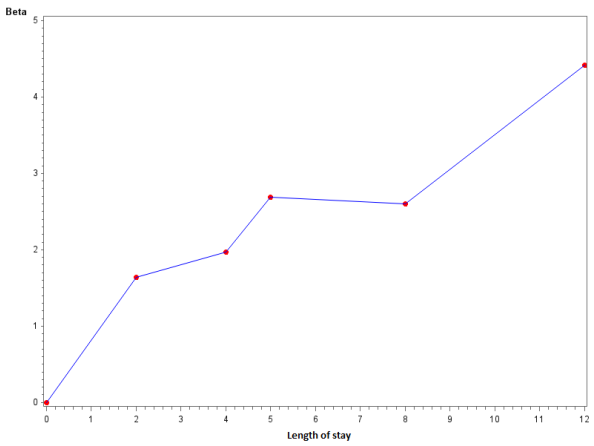

Figure 4.2: Beta estimates vs. LOS covariate

Figure 4.1 should represent a linear relationship but appears to grow much faster for higher values of the age covariate, which may indicate that, in this form, the covariate does not meet the log-linearity assumption. For the covariate length of stay (LOS), the relationship presented in Figure 4.2 may appear sufficiently linear to suggest that the hypothesis can be verified (but this is an impression and not a conclusion).

Yet this is only a first and visual impression, there are other methods that can be used to confirm or refute these impressions, as we'll see in section 4.4.3.

4.4.2 Proportional-hazards

The Cox PH model does not only assume log-linearity. Explicit in its name, the proportional-hazards assumption is also assumed by the model. In the following subsections, we'll look at where this hypothesis comes from, what it implies and how to assess it.

Consistent with the design of a Cox PH model, the hazard ratio of two subjects $i, j \in \{1, \dots, n\}$ with covariate vectors X_i and X_j is given by,

$$\mathrm{HR}_{ij}(t) = \frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp(X_i^T\beta)}{\lambda_0(t)\exp(X_j^T\beta)} = \frac{\exp(X_i^T\beta)}{\exp(X_j^T\beta)} = \exp((X_i - X_j)^T\beta) = \exp(\beta_{ij}) \tag{4.8}$$

This relation tells us that the hazard ratio does not depend on time t, indicating that instantaneous risks remain proportional over time. The model can only be valid if reality is consistent with this construction, which is by no means automatic in reality and must be checked. Among the many methods used in epidemiology for such purposes, we chose to implement two of them, one dedicated to quantitative variables, the other to qualitative variables. These methods are quite intuitive and were chosen because they best reflect and translate the assumptions.

4.4.2.1 Graphic validation method

The first method for checking PH assumption (for categorical covariates), and the one most commonly used in epidemiology, is to plot $\log(-\log(\hat{S}(t))$ curves stratified by covariate level to see if they seem parallel (D. Commenges and H. Jacqmin-Gadda, 2015, p.140 [6]).

To clarify the mathematics behind this, we consider the case of a single binary-coded covariate X_k. By definition, the hazard of a subject for whom $X_k = 1$ and of a subject for whom $X_k = 0$ are linked by the following relationship:

$$\lambda_1(t) = \exp(\beta)\lambda_0(t) \tag{4.9}$$

Since the survival function can be written as (4.3), we can express (4.9) in the following form,

$$S_1(t) = S_0(t)^{\exp(\beta)}$$

Equivalently,

$$\log(S_1(t)) = \exp(\beta)\log(S_0(t))$$
$$-\log(S_1(t)) = -\exp(\beta)\log(S_0(t))$$

This brings us back to the cumaltive hazard function.

$$A_1(t) = \exp(\beta)A_0(t)$$
$$\log(A_1(t)) = \log(A_0(t)) + \beta$$

Thus, following Andersen et al. (P.K. Andersen et al., 1993, p.539-542 [1]) discussion, one may plot the estimate $\hat{A}_0(t, \hat{\beta})$ and $\hat{A}_1(t, \hat{\beta})$ versus t (or $\log(t)$). Kaplan-Meier curves can be used for this, see Appendix B.1.2. Under the proportional hazards model, these curves should be approximately parallel, the constant vertical distance between $\log \hat{A}_1(t, \hat{\beta})$ and $\log \hat{A}_0(t, \hat{\beta})$ being approximately $\hat{\beta}$. This method can be extended to variables with more than two modalities (P.K. Andersen et al., 1993, p.540 [1]), the curves must be parallel in pairs. Kaplan-Meier curves can be plotted using the LIFETEST procedure, and in particular the '`lls`' option is used to plot loglogs curves, corresponding to the estimate $\hat{A}_0(t, \hat{\beta})$ and $\hat{A}_1(t, \hat{\beta})$ versus $\log(t)$.

```
ODS GRAPHICS ON;
    PROC LIFETEST DATA=Patients plots=lls;
        TIME Delay*RH30(0);
        STRATA CR;
    RUN;
ODS GRPHICS OFF;
```

Example. Loglogs curves were plotted for the different modalities of the Cancer Rank (CR) covariate, and for a 3-level categorization of the Charlson Index (CCI).

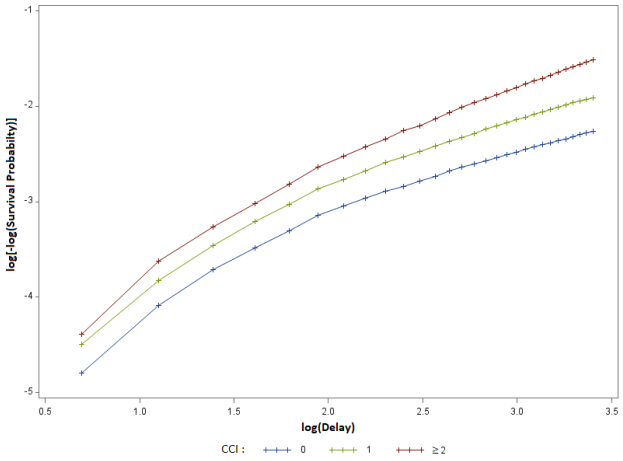

Figure 4.3: Log of negative survivor log estimated for 3-level CCI covariate

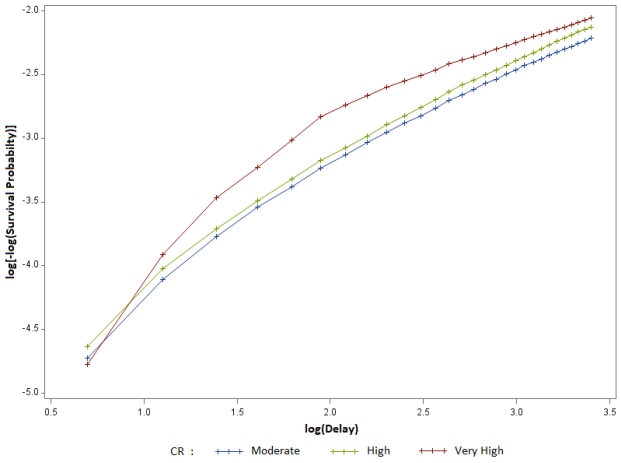

Figure 4.4: Log of negative survivor log estimated for CR covariate

We note that the 3-level categorization of the CCI covariate respects the PH assumption, with the curves running parallel 2 to 2, see Figure 4.3. However, for the CR covariate, see Figure 4.4, the situation is more delicate, with a crossover and a lack of consistency for the Very High level. In addition, the curves for the moderate and high levels are very close, which may reflect an insufficient difference between the risk associated with these two levels. The coding of this covariate will need to be reworked before it can be included in a Cox model.

4.4.2.2 Interaction with time

To test hypotheses of proportionality of risk for quantitative covariates, a practical solution is to introduce an interaction between a continuous function of time and the covariate to be tested.

If we include in the model an interaction term between time (or a function of time such as the logarithm) and the variable we wish to test, this will introduce a parameter β_{k_t} associated with a time-dependent variable: $X_k \times \log(t)$. Considering this function of time, the hazard ratio for a unit deviation of X_k increases linearly with time. Thus, a test of nullity of the coefficient β_{k_t} is equivalent to a test of hazard proportionality for X_k (D. Commenges and H. Jacqmin-Gadda, 2015, p.142-143 [6]).

It is easy to do this with SAS by including it directly in the PHREG procedure by adding an interaction term as follows.

```
PROC PHREG DATA=Patients;
    MODEL Delay*RH30(0) = Age AgeT;
    AgeT = Age*log(Delay);
    TEST AgeT = 0;
RUN;
```

In the SAS code, TEST statement is used to request a test on the nullity of the coefficient associated with the parameter `AgeT`. If the null hypothesis is rejected, we conclude that the coefficient is significantly different from zero. Therefore, the interaction is significant.

Using age as an example, the interaction with time tested according to the syntax presented above gives us a p-value of less than 0.0001 for the covariate Age (`Age`), for the interaction between age and the logarithm of time (`AgeT`), as well as for the requested null coefficient test. This tells us that for age, the hazard ratio for a one-year difference increases linearly with time, which does not support the proportional hazard assumptions.

4.4.3 Martingale residues

In the foregoing, we proposed methods widely used in epidemiology, which seem to be the most transparent. However, there are other more powerful methods based on residuals. Several types of residual have been proposed and discussed, but residuals are much trickier to use for assessing the fit of a Cox PH model than they are for a linear model. This is one of the reasons why Schoenfeld and Cox-Snell residuals won't be presented here, but martingale residues will be briefly introduced. The other reason is that martingale residues can be used to check both loglinearity and PH assumptions. In addition, methods are available in SAS.

Martingale residuals are derived from the theory of counting processes. Our study can be seen as a simple counting process: the process takes the value 0 until a certain time T when the patient is rehospitalized, at which point the process takes the value 1. A test using martingale residuals has been developed (D.Y. Lin et al. 1993 [14]) to test the PH hypothesis globally (D. Commenges and H. Jacqmin-Gadda, 2015, p.142;152 [6]). Martingale residuals are also useful for checking loglinearity (by verifying the functional form of the variable) (D. Commenges and H. Jacqmin-Gadda, 2015, p.149 [6]).

The graphs and tests based on martingale residues are implemented directly in the PHREG procedure, in the ASSESS statement. The '`resample`' option of ASSESS gives a test of PH based on a Kolmogorov-type supremum test. The null hypothesis is that the covariate respects hazard proportionality. An example of code to test the proportional hazards assumption[2] for the age covariate is given below (D. Commenges and H. Jacqmin-Gadda, 2015, p.337-338 [6]).

```
PROC PHREG DATA=Patients;
    MODEL Delay*RH30(0) = Age;
    ASSESS VAR=(Age) PH / resample;
RUN;
```

We emphasize that these tools are time-consuming and only give a general idea which do not allow us to assert that the coding is adequate in case of non-rejection of the supremum test.

Example. Here we show two examples of ASSESS statement applications, for the Age and LOS covariates.

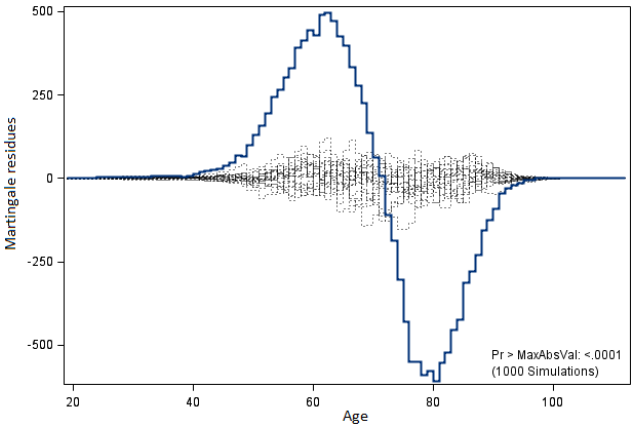

Figure 4.5: ASSESS PH output of Age covariate

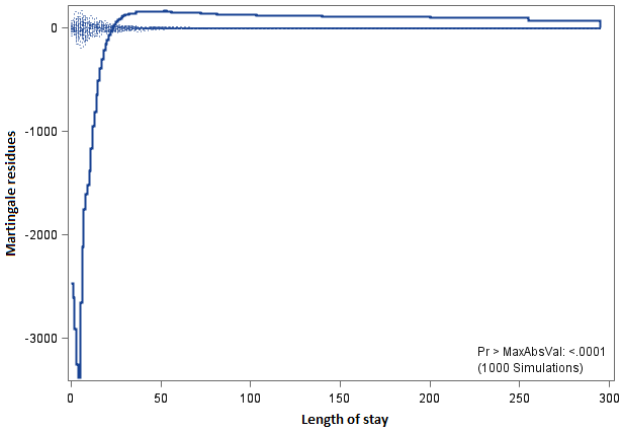

Figure 4.6: ASSESS PH output of LOS covariate

Both graphs on Figure 4.5 and 4.6, obtained via ASSESS for Age and LOS covariates, point in the same direction: reality is far from the simulations. The shape gives us a visual indication that the PH hypothesis has been violated, and the supremum test at the bottom right is significant, confirming the violation. Considering the previous results on PH assumptions, we can assert that these two covariates will have to be transformed in order to be included in a Cox PH model. The new challenge at this stage is to deal with covariates that don't support the model's assumptions. The following section presents possible solutions to this problem.

[2] To test loglinearity the code is similar, just remove the '`PH`' option, in this case the null hypothesis is that the functional form of the covariate has been correctly specified.

4.5 What if assumptions underlying the model are violated ?

Once we have identified the variables that do not meet the assumptions of the model, we must find a way to address them.

4.5.1 Quantitative covariates

Generally, it's preferable to keep quantitative covariates in continuous form whenever possible, since we have more information, interpretation is straightforward, and we use a single degree of freedom for testing. However, if the hypotheses are not met, we have to transform the variable.

There are several options to consider, the most common of which are the following:

- **Introducing a continuous function of the covariate:** Depending on its evolution over time, the variable can be transformed into an appropriate functional form, e.g. squared or logarithmic. It preserves all the information, in the case of predictions studies it is very interesting to obtain more precise models. In the context of association studies, one will tend not to introduce a covariate function, because it complicates a lot the interpretation of the covariates.
- **Categorization:** The idea is to divide the continuous covariate into several sub-levels, with each interval corresponding to a level. It allows researchers to avoid strong assumptions about the continuous covariates. This approach is often favored by non-statisticians because it has the advantage of providing results that are easier to interpret. However, categorization raises several issues such as the number of cutpoints, where to place them and the loss of information.
- **Modeling interaction with time:** If we wish to model hazard ratios as a function of time, we will introduce interaction with time into the model. However, care must be taken to keep the model reasonably simple and easy to interpret. Logarithmic or quadratic relationships can be modeled, but more complex models will be much harder to interpret. Additionally, different hazard ratios will have to be given for different time periods for the covariate that does not respect the model's assumptions (D. Commenges and H. Jacqmin-Gadda, 2015, p.148 [6]).

The interpretation of hazard ratios and the search for associations are at the heart of this study, therefore we gave priority to the categorization of quantitative covariates. Which raises another important query; **How to define cutpoints?**. In the literature, the most common way to define cutpoints is to consider deciles or quartiles. This method will be studied to get a first idea of the variation of the covariate regarding the outcome. In a second step, to allow the best segmentation, one can study the trends of the covariate relatively to the outcome. Each time the rehospitalization rate changes, a new category can be considered.

Note: After adjusting the functional form, univariate analyses should be performed once again to confirm that the covariate still provides enough information to be studied.

4.5.2 Qualitative covariates

Qualitative variables generally cause fewer problems than quantitative ones, not least because loglinearity assumptions can be avoided by introducing **dummy variables**. However, when they do not satisfy the proportional-hazards assumption, we can try to combine certain modalities. This was the case in our study for medical procedure covariate; see Appendix B.2.2.

Ultimately, in the case of a binary variable that does not meet the hypothesis, **stratification** remains the easiest option to implement. When the effect of a qualitative covariate is found to vary over time, it may be interesting to stratify based on this covariate (D. Commenges and H. Jacqmin-Gadda, 2015, p.147 [6]). But this is only useful when the covariate is not of great interest in our study and is qualitative with few categories. Indeed, we cannot stratify with respect to this covariate and introduce it into the model. Thus, if we choose to stratify with respect to a certain covariate, we will not be able to quantify its effect on the event.

In order to make a clean selection of covariates while taking into account the assumptions of the Cox model, we will have to combine all these methods to find the form that best fits the model.

Chapter 5

Covariates correlation

Once all the covariates have been checked to the model assumptions, it is still necessary to ensure that these covariates are not correlated which each other. One of the main purposes of regression analysis is to independently assess the effect of each covariate. The idea is that we can change the value of one covariate and not the others. However, when covariates are correlated, it indicates that changes in one covariate are associated with shifts in another one. The stronger the correlation, the more difficult it is to change one variable without changing another.

Correlated covariates causes the following two basic types of problems:

- It reduces the precision of the estimated coefficients, which weakens the statistical power of our model.
- The coefficient estimates can swing wildly based on which other covariates are in the model. The coefficients become very sensitive to small changes in the model.

However, these issues affect only covariates that are correlated. The solution is to remove some of the highly correlated covariates. We will not introduce them simultaneously, instead we'll run several models; one model based on the first covariate and one based on the second.

In statistics, one way to measure these associations is to use correlation coefficients. They provide a quantitative measure of both the direction and strength of this tendency to vary with each other. In this study, we'll be comparing categorical variables, so we'll essentially be using Cramér's V to measure this.

5.1 Cramér's V correlation coefficient

Cramér's V coefficient is a measure of association derived from better-known Pearson's chi-square, involving the differences between observed and expected frequencies. It is designed to vary between -1 and 1 for 2×2 tables and between 0 and 1 for larger tables. Cramér's V is computed as,

$$V = \sqrt{\frac{\frac{1}{n}\sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}}{\min(N-1, P-1)}} \quad \text{for } N \times P \text{ tables}$$

where n_{ij} is the observed frequency in table cell (i, j) and e_{ij} is the expected frequency for table cell (i, j).

In our study we worked with a majority of binary variables (due to Cox model assumptions), by doing so, we study 2×2 tables. Note that in this context, Cramér's V can be simplified as follows,

$$V = \frac{(n_{11}n_{22} - n_{12}n_{21})}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}$$

Once we've calculated our scores, we need to define a threshold above which we can't neglect the correlation between covariates.

5.2 Association thresholds

In the literature, the rule of thumb to interpret the size of a correlation varies according to the study. In Sociology, the correlation thresholds given in Table 5.1 are the most often applied, a study in Psychology (J.F. Hemphill, 2003 [12]) was carried out on which thresholds must be chosen, the conclusions of which lead in the same direction. The literature remains unclear concerning the thresholds used in Medical research.

Size of Correlation	**Interpretation**
.70 to 1.0 ($-.70$ to -1.0)	Very strong positive (negative) correlation
.40 to .70 ($-.40$ to $-.70$)	High positive (negative) correlation
.20 to .40 ($-.20$ to $-.40$)	Moderate positive (negative) correlation
.00 to .20 (.00 to $-.20$)	Negligible correlation

Table 5.1: Rules of thumb about correlation coefficient size

When J. Cohen (J. Cohen, 1988, p.478 [5]) discussed effect sizes in the context of multiple regression and correlation analysis (MRC), the following thresholds were introduced: under 0.15 is considered low, 0.15 to 0.35 is moderate and above 0.35 is considered high. Cramér's V can be considered as a standardized effect size because they indicate the strength of the relationship between variables using unitless values that fall within a range of -1 to +1. Since the principles of f^2 discussed by Cohen and Cramér's V are similar, we will consider a threshold of 0.15 for neglecting the correlation in our study. Beyond this threshold, we will consider that the covariates cannot be included in the model at the same time. If necessary, it can be slightly relaxed to 0.20.

To simplify the process, we implemented a macro that creates a correlation matrix containing Cramér's V scores (of the same design as Pearson's). The macro is accompanied by a heatmap, which legend is adapted to the tolerance threshold defined above. It can be consulted in Appendix B.3.

5.3 Application

In our study, we quickly realized that the Elixhauser index was highly correlated (≥ 0.20) with all covariates except age, see Figure 5.1, making it impossible to use it in a model containing covariates other than Age.

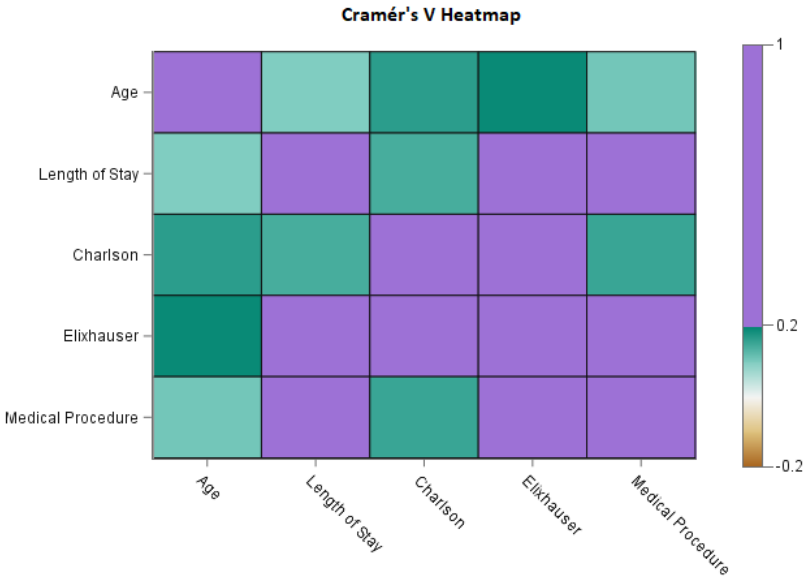

Figure 5.1: Heatmap for Age, LOS, CCI , ELX , and MP

Since we wanted to work on the basis of Elixhauser, we proposed to perform a classification of comorbidity groups. However, in doing so, we introduce a new covariate for each group. Therefore, we need to eliminate potential comorbidity groups that are correlated with the other covariates and with each other. We used our macro to do this. We constructed a matrix of the correlation between the two covariates chosen for our model - Age and LOS - and the 31 comorbidity groups, and obtained the heatmap presented below on Figure 5.2.

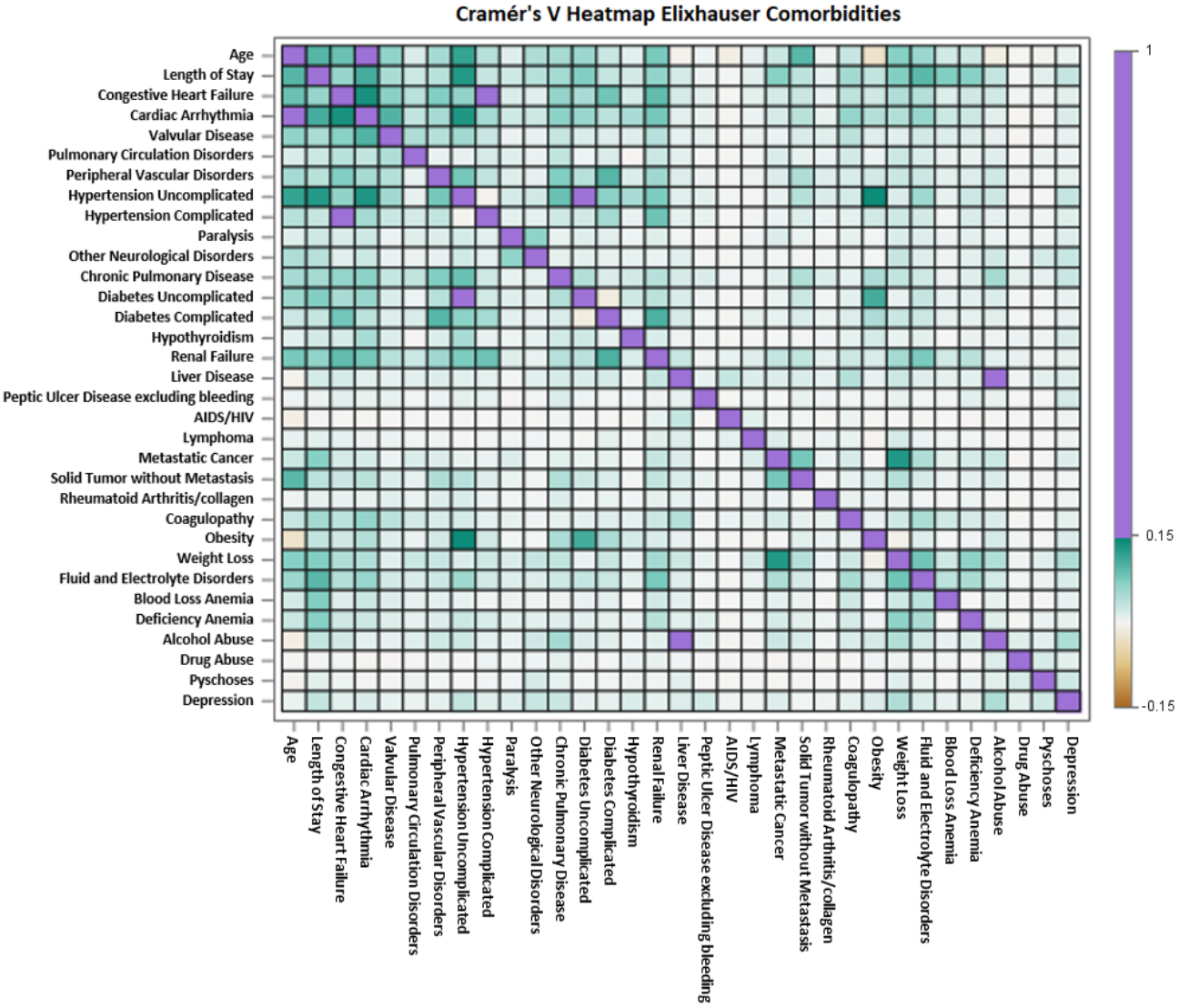

Figure 5.2: Heatmap for Age, LOS and 31 Elixhauser comorbidity groups

As the group of comorbidities associated with cardiac arrhythmia is correlated with age, we removed it. We also had to choose between certain comorbidity groups that were correlated with each other, basing this choice on an initial classification containing all the variables. When two variables were correlated, we kept the one with the highest weight in the classification.

We carried out a new classification, excluding the correlated variables, and reproduced a new heatmap, in order to obtain 10 uncorrelated comorbidity groups providing the most information. We obtained the heatmap shown in figure B.3, see Appendix B.3.2. One of the final Cox PH models was built on the basis of these covariates.

Chapter 6

Interaction between covariates

6.1 Motivation

An interaction effect occurs when the effect of one covariate depends on the value of another. This type of effect makes the model more complex, but if the real world behaves in this way, it's essential to incorporate it into our model. Failure to include interaction terms in a model will result in assessing only the main effects without taking into account the impact of interaction.

As an example, we consider a Cox PH model adjusted to explain the risk of rehospitalization within 30 days by age (X_{Age}), Length of stay (X_{LOS}) and Charlson comorbidity index (X_{CCI}). This Cox PH model assumes that the hazard for the ith individual ($i = 1, \ldots, n$) is as follows,

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 X_{i\text{Age}} + \beta_2 X_{i\text{LOS}} + \beta_3 X_{i\text{CCI}})$$

Thus, the model asserts that LOS acts linearly on $\log(\lambda_i(t))$ and that the coefficient of linearity is independent of the type/number of pathologies (CCI score) that the patient may have. However, one can imagine that the effect of length of stay on the patient's condition depends on the patient's comorbidities.

One possible model with interaction between length of stay and Charlson score is obtained by replacing the covariate LOS by two covariates,

$$X_{\text{CCI}_{\leq 4}} = \begin{cases} X_{\text{CCI}} & \text{if LOS is 4 days or less} \\ 0 & \text{otherwise} \end{cases} \qquad X_{\text{CCI}_{>4}} = \begin{cases} X_{\text{CCI}} & \text{if LOS is more than 4 days} \\ 0 & \text{otherwise} \end{cases}$$

which is done by adding a crossover term, CCI $\times$ LOS, as a covariate in the model code.

6.2 The role of interactions in our study

In this study, we consider interactions for the following two purposes: we want to see whether they improve the model, and whether the effects are similar in the subgroups studied (D. Commenges and H. Jacqmin-Gadda, 2015, p.58 [6]). For example, if we introduce the CCI $\times$ LOS interaction into a model, we'll keep it provided that, either it proves to improve the model, or the effects of CCI are different according to LOS. As the aim is to observe main effects, in this study we will not attempt to interpret interactions, but simply introduce them to consider their possible impact on the HR of main covariates.

Another important point to note is that, since we are referring to the literature, which hardly ever mentions interaction effects in models (in the case of association search), and given our mathematical approach, we focused on interaction terms that were significant and that we felt were truly relevant. Although it may not yet be clear how to account for interactions and their effects on a model, the next chapter is devoted to testing the fit of the model we've built. In particular, the following chapter contains an example illustrating the difference in fit when the interaction term, CCI $\times$ LOS, is excluded versus when it is included.

Chapter 7

Goodness of fit

The question raised in this chapter is the following: **How can we test the reliability of our model?**

Prior to Cox's proportional-hazards model, logistic regression was commonly used to analyze survival data. Goodness of fit in the case of logistic regression was tested by the Hosmer-Lemeshow test (D. Commenges and H. Jacqmin-Gadda, 2015, p.77 [6]), which assesses whether or not the observed event rates match expected event rates in subgroups of the model population. To test the overall adequacy of the model in the case of a Cox regression, we would like to work with the same type of test as the Hosmer-Lemeshow test.

The most naïve approach to apply the Hosmer-Lemeshow goodness-of-fit test to survival data would be to simply use logistic regression with the event/censoring indicator as the binary outcome and then use the Hosmer-Lemeshow test. The concern with using logistic regression is that the time to the event/censoring is ignored completely. The results may be similar if the times to events are the same on average as the times to censoring (or if the event is rare and has a short time of occurrence) but this is often not the case. Since logistic regression has a different interest, it may not make sense to test the goodness-of-fit of the model using the Hosmer-Lemeshow test.

The test that seem to be an equivalent to the Hosmer-Lemeshow goodness-of-fit test for logistic regression is the Grønnesby and Borgan test.

7.1 Grønnesby and Borgan's test

The Grønnesby and Borgan test is based on martingale residuals which represent the difference between the number of observed events and the model based estimate of the expected number of events (J.K. Grønnesby and O. Borgan, 1996 [11]).

The estimated martingale residual for subject i at time t for the Cox PH model is defined as,

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u) \exp(X_i^T \hat{\beta}) \, \mathrm{d}\hat{\Lambda}_0(u) \quad i = 1, \ldots, n$$

where $\Lambda_0(t)$ is the baseline cumulative intensity process, estimate by the Breslow estimator as follow,

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\mathrm{d}N_\cdot(u)}{\sum_{l=1}^n Y_i(u) \exp(X_i^T \hat{\beta})}$$

where $N_\cdot(t) = \sum_{i=1}^n N_i(t)$.

Consequently,

$$\hat{M}_i = N_i - \int_0^\infty Y_i(u) \exp(X_i^T \hat{\beta}) \, \mathrm{d}\hat{\Lambda}_0(u) \tag{7.1}$$

$$= \text{observed}_i - \text{expected}_i \tag{7.2}$$

The observations are divided into G groups according to their estimated risk score $\hat{r}_i = X_i^T \hat{\beta}$, an approach similar to the one used by Hosmer and Lemeshow for logistic regression. The sum of the martingale residuals is computed within each group. If the model assumptions hold, this sum should be close to zero.

Following May and Hosmer (S. May and D.W. Hosmer, 1998 [16], P.K. Andersen et al., 1993 [1]), we use the counting process formulation of the model. We assume that an event can occur only once and that the process is subject only to right censoring. Under the counting process approach, the observed number of events in each group is approximately a Poisson variate with parameter estimated - under the hypothesis that the fitted model is correct - by the model-based estimate of the expected number of events.

If the estimated expected number of events is large, then the following standardized statistic,

$$Z_g = \frac{(\text{observed}_g - \text{expected}_g)}{\sqrt{\text{expected}_g}} \qquad g = 1, \ldots, G$$

should be approximately distributed as a $\mathcal{N}(0, 1)$ variate.

Thus a p-value computed from standard normal distribution can be used to assess agreement between observed and estimated expected within each group. The expected number of events for each group can be calculated by subtracting the martingale residual for group g from the observed number of events for group g.

Therefore,

$$Z_g = \frac{\hat{M}_g}{N_g - \hat{M}_g} \qquad g = 1, \ldots, G$$

We used May and Hosmer's method as the computation of the test statistic, since in the form introduced by Grønnesby and Borgan it requires calculating the covariance matrix which is somewhat tedious. This test remains a simple and effective way to evaluate the viability of the model. However, care must be taken with the choice of groups.

7.2 Risk score groups

The choice of groups is a widely debated subject, but in general, groups based on risk score deciles were chosen. In our study, this choice was not the most appropriate and required further reflection. Indeed, the asymptotic properties of all survival estimators depend on the existence of a sufficiently large expected number of events in the g-th decile. May and Hosmer (S. May and D.W. Hosmer, 2004, [17]) show that the Grønnesby and Borgan test becomes too liberal when the number of events per decile is too low. Unfortunately this was the case in our study, we didn't have enough events by deciles.

To ensure a sufficient number of events per decile, we therefore considered several different approaches on the recommendations of May and Hosmer and Demler et al. (O.V. Demler et al., 2015, [10]). First, we used the collapsing strategy of Demler et al. starting with 10 deciles and collapsing the smaller deciles with their nearest neighbors. This strategy makes it possible to use all the data while ensuring estimator convergence.

Still based on risk scores, a second approach aimed at constructing groups with equal numbers of events. Using May and Hosmer's recommendations, we also constructed 10 groups, each containing the same number of events. Since we have observed a total of around 30,000 events, it means that group 0 will contain around 3,000 rehospitalizations for patients with a low score, and group 9 will contain around 3,000 rehospitalizations for patients with a high score.

We use these two group strategies to ensure that the test is powerful enough to detect a poor fit. Both methods provided similar results. However, the method using equal numbers of events per group seemed to detect deviations more effectively. The implementing code is left in Appendix B.4.

7.3 Application

The question of model fit arises in a general way, but in this example it makes more sense to illustrate the difference between fit with and without significant interaction terms.

Let's consider the model constructed from the covariates Age, Length of stay (LOS) and Charlson comorbidity index (CCI). Without taking interaction terms into account, we found that only 70% of the p-values obtained per group are greater than 0.05, see Table 7.1.

Group	Observed	Expected	Martingale Residuals	Z	p-value
0	3049	3157.0	-108.0	-1.92215	0.0546
1	3050	3144.5	-94.5	-1.68522	0.0919
2	3050	3000.5	49.5	0.90367	0.3662
3	3050	3058.5	-8.5	-0.15370	0.8778
4	3050	3031.5	18.5	0.33600	0.7369
5	3050	3042.5	7.5	0.13597	0.8918
6	3050	2926.0	124.0	2.29237	**0.0219**
7	3050	2943.5	106.5	1.96299	**0.0497**
8	3050	2965.5	84.5	1.55170	0.1207
9	3041	3214.0	-173.0	-3.05157	**0.0022**

Table 7.1: Test results of the model without interaction terms

However, we would like to see if by considering the interaction terms we are able to avoid rejecting the null hypothesis for at least one or two more groups. When we add interaction terms between CCI and LOS, we obtained the following results.

Group	Observed	Expected	Martingale Residuals	Z	p-value
0	3049	3100.5	-51.5	-0.92489	0.3550
1	3050	3088.5	-38.5	-0.69277	0.4885
2	3050	2972.5	77.5	1.42148	0.1552
3	3050	3082.0	-32.0	-0.57641	0.5643
4	3050	3060.0	-10.0	-0.18078	0.8565
5	3050	3111.5	-61.5	-1.10253	0.2702
6	3050	2979.0	71.0	1.30084	0.1933
7	3050	2988.0	62.0	1.13423	0.2567
8	3050	3007.5	42.5	0.77497	0.4383
9	3041	3094.5	-53.5	-0.96174	0.3362

Table 7.2: Test results of the model with interaction terms

The Z statistics and p-values in Table 7.2 support agreement between the observed and estimated expected number of rehospitalizations, since we didn't reject null hypothesis in each group, confirming that the Cox PH model with interaction terms seems to fit well the data.

Chapter 8

Generalized linear mixed models

In a classical model, errors are assumed to be independent and identically distributed according to a normal distribution. However, this is not the case when data are structured at multiple levels, typically when certain individuals share a common environment that is likely to influence the outcome of interest. Examples include students in a school, employees in a company, patients in a hospital, etc. **Multilevel models** are designed to address questions raised by such data. They allow to detect such heterogeneity, to quantify it, and/or simply to obtain unbiased estimates of the impact of some individual variables, the latter being the focus in this study.

Analysis of discrete data taking into account cluster correlation effects can be performed using **Generalized Linear Mixed Models (GLMMs)**. These are built on the same principle as Linear Mixed Models (LMMs) and the same root as Generalized Linear Models (GLMs).

8.1 Fixed and random effects

Let's start with a quick review of what is considered a fixed and a random effect. The definition of fixed and random effects is a matter of debate in the literature. There are several possible definitions of fixed and random effects and we will present here the ones that seem to be the most coherent with the project and the simplest to understand and apply.

- **Fixed effect (deterministic process):** When a covariate has a fixed effect, the data come from all possible levels of qualitative covariate or from a quantitative covariate. The aim is to draw conclusions about the levels of the qualitative covariate or the relationship between the quantitative covariate and the outcome. As an example, let's compare the rehospitalization risks of patients who have undergone 3 different types of medical surgical procedures. The type of medical procedure is a fixed effect (all three types have been sampled) and we wish to draw conclusions about the effects of these three specific types of medical procedures.
- **Random effect (stochastic process):** Random effect variables are also called random factors because they are only categorical variables. A random effect occurs when individual observations are naturally grouped into larger clusters. These are usually grouping factors whose influence we want to control in the model, but whose specific effect on the outcome we are not interested in. Take, for instance, the purpose of this study: rehospitalization of patients hospitalized due to prostate surgery in France (between 2012 and 2014). Patients from the same geographic code may have some correlation with each other because they share the same socio-environmental conditions. Although we are not interested in the specific effect of each geographic code, we can include random effects at the geographical level to account for different sources of variability.

Given the binary nature of our outcome, it is a generalized linear mixed model (with a logit link) that seem suitable. Conditional on random effects, the outcome follows a generalized linear model and the random effects are included in the set of covariates.

8.2 General structure of the model

8.2.1 From GLMs to GLMMS

Denote y the outcome and $X = (1|X_1|\dots|X_p)$ the design matrix of p covariates, a generalized linear model is defined by:

- the outcome distribution must belong to the exponential family.
- a combination of covariates X and parameters β.
- a link function connecting $\mu = \mathbb{E}(y)$ to the linear predictor.

Let y_i be the outcome observation for patient i, $i = 1, \dots, n$. The expectation of a basic GLM was,

$$\mathbb{E}(y_i) = g^{-1}(X_i^T \beta).$$

Following McCulloch (C.E. McCulloch, 2005, p.220-221 [18]), random effects are incorporated by enlarging the model as follows,

$$\mathbb{E}(y_i|b) = g^{-1}(X_i^T \beta + Z_i b)$$

where $g(\bullet)$ is the link function, X_i is the ith row vector of fixed effects covariates of dimension p, β is the p-vector of parameter for fixed effects; to that specification we add Z_i, which is the ith row vector fo random effects of dimension $q \leq p$ and b a q-vector of identically and independently distributed random effects.

As with GLMs, when constructing a GLMM we are interested in odds ratios and their confidence intervals. Yet their significance is more nuanced in the presence of mixed effects. In classical logistic regression, odds ratios are the odds ratios expected if all other covariates are fixed. The same is true for mixed effects logistic models, with the addition that holding everything else fixed includes holding the random effect fixed. In other words, the odds ratio here is the conditional odds ratio for a person with constant age and length of stay, as well as for a person with either the same geographical code, or geographical codes with identical random effects.

While this may make sense, when there is high variability between geographical codes, the relative impact of fixed effects may be small. In this case, it is useful to examine the effects at different levels of the random effects, or to obtain the average fixed effects by marginalizing the random effects.

We'll look at how to get this information using the GLIMMIX procedure in section 8.3. But first, let's specify these differences a little more by pointing out the construction of a GLMM with a logit link.

8.2.2 Multilevel logistic regression

In the case of a binary outcome, a logit link is introduced, and considering that the distribution of random effects is modeled by a Gaussian distribution, $\mathcal{N}(0, \sigma_b^2)$, the model can be expressed as follows (D. Commenges and H. Jacqmin-Gadda, 2015, p.180 [6]).

$$\text{logit}(P(y_i = 1|b)) = X_i^T \beta + Z_i^T b$$

And so,

$$P(y_i = 1|b) = \frac{\exp(X_i^T \beta + Z_i^T b)}{1 + \exp(X_i^T \beta + Z_i^T b)} \tag{8.1}$$

Given this model[1] specification, the expectation of y can be written as follows.

$$\begin{aligned}\mathbb{E}(y_i) &= \mathbb{E}\left[\mathbb{E}(y_i|b)\right] \\ &= \mathbb{E}\left[\mathbb{E}g^{-1}(X_i^T\beta + Z_i^T b)\right]\end{aligned}$$

Having a log link, $g(\mu) = \log(\mu)$ and $g^{-1}(x) = \exp(x)$, thus we have,

$$\mathbb{E}(y_i) = \mathbb{E}\left[\mathbb{E}(\exp(X_i^T\beta + Z_i^T b))\right] \tag{8.2}$$
$$= \exp(X_i^T\beta)\mathbb{E}\left[\exp(Z_i^T b)\right] \tag{8.3}$$

In general, relation (8.2) couldn't be more simplified, but further explanations are given in McCulloch (C.E. McCulloch, 2005, p.223 [18]). This equation illustrates mathematically the differences explained in the previous section: where in logistic regression the expectation of the outcome is explained by $\exp(X_i^T\beta)$, here variations are added.

8.3 GLIMMIX Procedure

SAS offers the GLIMMIX procedure for fitting generalized linear mixed models (GLMMs). The GLIMMIX procedure enables to specify a generalized linear mixed model and perform confirmatory inference in such models. The syntax is somewhat similar to that of the PHREG procedure and includes the CLASS, MODEL and RANDOM statements, which are the main features used in this study.

The procedure syntax is as follows.

```
PROC GLIMMIX DATA=Patients;
    CLASS CCI (ref='0');
    MODEL RH30 (event='1') = Age LOS CCI Rural Public
                          / dist=binary link=logit oddsratio solution;
    RANDOM intercept / subject=CodeGeo;
RUN;
```

We specify the option '`dist=binary`' in the MODEL instruction, in order to specify the logit link function. By default, for a binary outcome, the logit link is chosen by the GLIMMIX procedure. The '`oddsratio`' option will also be specified to return the odds ratio produced by the model for each covariate. The RANDOM instruction states that the linear predictor contains an intercept term that varies randomly at the level of the cluster effect. In other words, a random intercept is drawn separately and independently for each cluster in the study. The '`solution`' option is used to display fixed-effects parameter estimates (their construction is detailed in the dedicated SAS help - GLIMMIX MODEL Statement).

SAS output gives us information about the model, class level and number of observations, dimensions, optimization, iteration and convergence status, tests of fixed effects and additional information on the estimation, we won't cover the details of this here, but the final message in the journal, is important[2]:

```
Convergence criterion (PCONV=1.11022E-8) satisfied.
```

It indicates that the iterative algorithm of the estimation process was able to settle on an answer.

Note: For a model containing random effects, the GLIMMIX procedure, by default, estimates the parameters by applying pseudo-likelihood techniques as in Breslow and Clayton (N. E. Breslow and D. G. Clayton, 1993 [2]).

[1] Model 8.1 with $q = 1$ and $Z_{ij} = 1$ leads back to the special case of logistic regression.
[2] This information is also present in the PHREG procedure.

Part III

Conclusion of the research project

Chapter 9

Results

9.1 Descriptive analysis of individual and environmental factors

The characteristics of our study population are reported in Table 9.1 and Table 9.2, which also shows the distribution of each factor between rehospitalized and non-rehospitalized patients. A total of 270473 patients who underwent prostate surgery between January 2012 and November 2014 met the inclusion criteria, 11.27% of them, i.e. 30490 patients, were rehospitalized within 30 days. All factors, both environmental and individual, were found to be statistically associated with the risk of 30-day rehospitalization.

9.1.1 Individual factors

The results of the univariate tests on individual factors are summarized in the Table 9.1, page 31. Because the quantitative variables did not meet the assumptions of the Cox PH model, they were coded as categorical variables. As mentioned above, the Elixhauser index was correlated with the other covariates. We therefore chose to study only 10 of the Elixhauser comorbidity groups as specified in section 3.2. The Elixhauser comorbidity groups selected (via a classification tree), the results of the associated univariate analyses and the verification of model assumptions are left in Appendix C.

9.1.1.1 Overall population

The median age of patients was 69 years and the median length of stay was 4 days. In most cases (78.03%), the surgical procedure was classified as "Anesthesia" or "Surgery". Cancers classified as "High Rank" were diagnosed less frequently in overall patients (26.39%) than cancers classified as "Very High Rank" (36.86%) or "Moderate Rank" (36.75%). The comorbidity, Charlson and Elixhauser indices were adjusted by excluding pathologies identified at inclusion, in order not to bias the study. The majority of patients had no comorbidity with a Charlson comorbidity score of 0 (78.27%), the rest being split into patients with a score exactly equal to 1 (12.61%) and patients with a score of 2 or more (9.12%). Regarding Elixhauser Index, the majority of patients in the study had less than 2 comorbidities (81.00%).

9.1.1.2 Comparison between rehospitalized and non-rehospitalized patients

Among the 270473 patients included, 11.27% were rehospitalized within 30 days. These rehospitalized patients were significantly ($p<0.0001$) older (median age of 71, 34% of patients over 75 years) than non-rehospitalized patients (median age of 68, 24.87% of patients over 75 years). The length of stay was also significantly ($p<0.0001$) higher for patients rehospitalized (59,58% of patients with a LOS over 4 days) compared to patients non-rehospitalized (43,85% of patients with a LOS over 4 days). Concerning medical procedures, rehospitalized patients received significantly less technical procedure during the initial hospitalization than non-rehospitalized patients (12,19% vs 23,21%, $p<0.0001$).

Within rehospitalized patients, there are notable differences in cancer rank, especially within the high and very high rank levels. In fact, rehospitalized patients were less diagnosed with benign or indeterminate tumors have (24.17%) than non-rehospitalized patients (26.67%). Conversely, rehospitalized patients were more likely to be diagnosed with malignant tumors (39.21%) than non rehospitalized patients (36.56%). Although this variable is statistically significant, from a clinical point of view, this difference is not marked. This suggests that it might lack sufficient discriminatory power against other factors, potentially making it less suitable for inclusion in a model. Concerning comorbidity scores, rehopitalized patients had significantly higher score than non-rehospitalized patients, especially for a score over 2. Indeed, regarding the CCI, the rate of a score greater than 2 was twice as high in rehospitalised patients (15.95%) as in non-rehospitalised patients (8.25%). The difference was less for the Elixhauser score, but it was still significant (27.90% for rehospitalized patients vs 17.87%, for non-rehospitalized patients).

	Overall Population	**No Rehospitalization**	**Rehospitalization**	***p*-value**
Number (%)	270473	239983 (88.73)	30490 (11.27)	
Individual Factors				
Age (years)				
Mean (std)	69.27 ± 9.25	69.03 ± 9.15	71.03 ± 9.82	< 0.0001
Median (IQR)	69 (63-76)	68 (63-75)	71 (64-78)	
≤ 75 years	200413 (74.10)	180300 (75.13)	20113 (65.97)	< 0.0001
> 75 years	70060 (25.90)	59683 (24.87)	10377 (34.03)	
Length of stay (days)				
Mean (std)	4.89 ± 5.05	4.69 ± 4.71	6.50 ± 6.20	< 0.0001
Median (IQR)	4 (2-7)	4 (2-6)	5 (4-8)	
≤ 4 days	147075 (54.38)	134752 (56.15)	12323 (40.42)	< 0.0001
> 4 days	123398 (45.62)	105231 (43.85)	18167 (59.58)	
Medical Procedure				< 0.0001
Technical	59429 (21.97)	55712 (23.21)	3717 (12.19)	
Anesthesia/Surgery	211044 (78.03)	184271 (76.79)	26773 (87.81)	
Cancer Rank				< 0.0001
Moderate rank[(1)]	99406 (36.75)	88242 (36.77)	11164 (36.62)	
High rank[(2)]	71379 (26.39)	64011 (26.67)	7368 (24.17)	
Very high rank[(3)]	99688 (36.86)	87730 (36.56)	11958 (39.21)	
Charlson Comorbidity Index[(4)]				< 0.0001
0	211706 (78.27)	190767 (79.49)	20939 (68.67)	
1	34110 (12.61)	29420 (12.26)	4690 (15.38)	
≥ 2	24657 (9.12)	19796 (8.25)	4861 (15.95)	
Elixhauser Index[(5)]				< 0.0001
< 2	219093 (81.00)	197110 (82.13)	21983 (72.10)	
≥ 2	51380 (19.00)	42873 (17.87)	8507 (27.90)	

[(1)]Benign Hyperplasia or Low Grade Dysplasia of the Prostate, [(2)]In situ, Benign or Unpredictable tumors, [(3)]Malignant tumors, [(4)]Charlson comorbidity index excl. inclusion conditions, [(5)]Elixhauser index excl. inclusion conditions.

Table 9.1: Individual factors associated with 30-day rehospitalization

9.1.2 Environmental factors

The results of the distribution and univariate tests on the environmental factors are summarized in table 9.2, page 32. The quantitative variable French Deprivation index (FDep) was considered in a quantitative and a categorical form.

	Overall Population	**No Rehospitalization**	**Rehospitalization**	***p*-value**
Number (%)	270473	239983 (88.73)	30490 (11.27)	
Environmental Factors				
Urban/Rural status				< 0.0001
Urban	138188 (55.55)	122869 (55.73)	15319 (54.17)	
Rural	110575 (44.45)	97615 (44.27)	12960 (45.83)	
French Deprivation index				< 0.0001
Mean (std)	-0.1406 ± 1.4061	-0.1552 ± 1.4124	-0.0271 ± 1.3524	
Median (IQR)	0.04 (-0.85-0.76)	0.03 (-0.87-0.75)	0.10 (-0.71-0.86)	
$< P_{20}^{(1)}$	51006 (19.75)	45898 (20.06)	5108 (17.38)	< 0.0001
$[P_{20}; P_{40}[$	52393 (20.29)	46614 (20.37)	5779 (19.66)	0.0055
$[P_{40}; P_{60}[$	51363 (19.99)	45688 (19.96)	5948 (20.23)	0.1988
$[P_{60}; P_{80}[$	51684 (20.01)	45603 (19.92)	6081 (20.69)	0.0023
$\geq P_{80}$	51557 (19.96)	45077 (19.69)	6480 (22.04)	< 0.0001
Private/Public status				< 0.0001
Private	198884 (73.53)	178091 (74.63)	19793 (64.92)	
Public	71589 (26.47)	60892 (25.37)	10697 (35.08)	

(1) P_{20}, P_{40}, P_{60} and P_{80} are the first (-1.1098), second (-0.2290), third (0.2946) and fourth (0.9379) quintiles.

Table 9.2: Environmental factors associated with 30-day rehospitalization

9.1.2.1 Overall population

In terms of environmental factors, the distribution between urban and rural residence is fairly even, with more patients living in urban zones (55.55%) than in rural zones (44.45%). The deprivation index was slightly below 0 (-0.14) on average, but the median remained well centered at 0 (0.04), indicating a fairly good distribution between economically favored and deprived areas. The majority of hospitalizations for the surgical procedure were carried out in private hospitals (73.53%).

9.1.2.2 Comparison between rehospitalized and non-rehospitalized patients

The proportion of rehospitalized patients living in rural areas is slightly higher (45.83%) compared to those not rehospitalized (44.27%) at this point, the difference is statistical but not clinically significant. Regarding the deprivation index, rehospitalized patients have higher values: a mean of -0.03 and a median of 0.10 for patients rehospitalized within 30 days, whereas the mean is -0.16 and the median is 0.03 for non-rehospitalized patients. A similar pattern is seen in the quintile analysis. In the lowest quintile (below P_{20}), rehospitalized patients constitute a smaller proportion (17.38%) than non-rehospitalized patients (20.05%). Conversely, in the highest quintile ($\geq P_{80}$), more patients experience rehospitalization (22.04%) compared to non-rehospitalization (19.69%).

We will return to these findings later in the section dedicated to the analysis of environmental factors. Finally, a clear pattern emerges indicating that rehospitalized patients were significantly more often admitted for prostate surgery in the public sector (35.08%) compared to non-rehospitalized patients (25.37%).

9.2 Multivariate analysis with individual factors

For the analysis of individual factors, Cox PH models were built, with prior verification of the assumptions of loglinearity and proportional-hazards, as well as the absence of inter-covariate correlation.

9.2.1 Selected covariates

After an in-depth study of correlation and interaction using only the methods presented in Part II, several models were considered. Several issues arose, the first of which was cancer rank, which did not provide sufficient information and did not fit the Cox PH model. We decided not to include it in the model after several attempts at categorization and stratification proved inconclusive. The second relates to the medical procedure; the type of surgery undergone was strongly correlated with length of stay. Although length of stay is important in our study, we opted to endeavor fitting models based on medical procedures rather than length of stay. Unfortunately, these models showed significantly lower reliability in the Grønnesby and Borgan tests. We therefore did not retain a model with the medical procedure.

Finally, two multivariate Cox PH models based on individual factors were selected. Since both comorbidity indices cannot be included in the same model (even considering only some of the Elixhauser comorbidity groups, due to correlations), we considered one model based on the Charlson comorbidity index and one based on Elixhauser groups. Grønnesby and Borgan's tests were carried out, after which the inclusion of certain interactions showed to provide more reliable models.

The first model contains the covariates Age, Length of stay and Charlson comorbidity index, plus an interaction term between Charlson comorbidity index and LOS. This interaction was discussed in Chapter 6.

The second model contains the covariates Age, Length of stay, and the following 10 Elixhauser comorbidity groups; Renal Failure, Solid Tumor without Metastasis, Metastatic Cancer, Congestive Heart Failure, Hypertension, Uncomplicated, Fluid and Electrolyte Disorders, Valvular Disease, Chronic Pulmonary Disease, Blood Loss Anemia and Coagulopathy. We also introduced two interaction terms, one between the Renal Failure group and Age and the other between the Solid Tumor without Metastasis group and Age.

9.2.2 Association of individual factors with the risk of rehospitalization

Table 9.3 shows the analysis of individual factors associated with the risk of rehospitalization by Cox's proportional hazards with hazard ratio (HR) and 95% confidence interval (CI), the p-value of covariates is also given.

In Model 1, Age was associated with the risk of rehospitalization, with patient less aged 75 years or less as the reference (HR=1), patients aged over 75 have a 31.4% higher risk of rehospitalization. Length of stay also proved to be a discriminating factor, with a length of stay of less than 4 days as the reference, patients with a length of stay of more than 4 days were at greater risk, with the model indicating a 68.0% higher risk of rehospitalization than the reference. The Charlson comorbidity index also proved to be significantly associated with the risk of rehospitalisation. Taking patients at level 0 as a reference, we find that those at level 1 already have a 26.6% greater risk of rehospitalisation, and when we go beyond level 2 the risk rises to 76.2%.

Age and LOS appear to have the same effect in Model 2, which is based on independent Elixhauser groups of comorbidities. The comorbidity groups that appear to be most associated with the risk of rehospitalisation are Renal Failure (60.4%), Metastatic Cancer (50.1%) and Solid Tumour without Metastasis (60.2%), which increase the risk of rehospitalisation by more than 50%.

	Model 1[1]			Model 2[2]		
	HR	**95% CI**	***p*-value**	**HR**	**95% CI**	***p*-value**
Age (years)						
$\leq$ 75 years	ref	ref	ref	ref	ref	ref
$>$ 75 years	1.314	1.282-1.346	$<$0.0001	1.313	1.281-1.345	$<$0.0001
Length of stay (days)						
$\leq$ 4 days	ref	ref	ref	ref	ref	ref
$>$ 4 days	1.680	1.641-1.719	$<$0.0001	1.672	1.633-1.711	$<$0.0001
Charlson Comorbidity Index[3]						
0	ref	ref	ref	-	-	-
1	1.266	1.226-1.307	$<$0.0001	-	-	-
$\geq$ 2	1.762	1.706-1.819	$<$0.0001	-	-	-
Elixhauser comorbidity groups[4]						
Renal Failure	-	-	-	1.604	1.509-1.706	$<$0.0001
Solid Tumor without Metastasis	-	-	-	1.501	1.430-1.575	$<$0.0001
Metastatic Cancer	-	-	-	1.602	1.485-1.729	$<$0.0001
Congestive Heart Failure	-	-	-	1.395	1.312-1.484	$<$0.0001
Hypertension Uncomplicated	-	-	-	1.032	1.006-1.058	0.0138
Fluid and Electrolyte Disorders	-	-	-	1.190	1.105-1.282	$<$0.0001
Valvular Disease	-	-	-	1.441	1.335-1.555	$<$0.0001
Chronic Pulmonary Disease	-	-	-	1.146	1.087-1.208	$<$0.0001
Blood Loss Anemia	-	-	-	1.399	1.257-1.558	$<$0.0001
Coagulopathy	-	-	-	1.470	1.322-1.634	$<$0.0001

[1]Adjustment with consideration of CCI $\times$ LOS interaction terms (HRs given in Appendix C.4), [2]Adjustment with consideration of Renal Failure $\times$ Age and Solid Tumor without Metastasis $\times$ Age interaction terms (HRs given in Appendix C.4), [3]Charlson comorbidity index excl. inclusion conditions, [4]The 10 most relevant comorbidity groups.

Table 9.3: Multivariate models predicting the risk of 30-day rehospitalization by Cox's proportional-hazards

9.3 Multilevel analysis with environmental factors

9.3.1 Selected covariates

The GLMMs models were built using the same covariates (and interaction terms) as the Cox models, only environmental factors were added. As the models built using GLMMs gave similar results to those obtained without considering environmental factors, only the results from the models based on the Charlson comorbidity index are presented in this section (the tables based on Elixhauser comorbidity groups are left in Appendix C.5). When it comes to asssumptions, only loglinearity is required for GLMMs. For qualitative variables, the use of the `CLASS` statement has alleviated any problems. For quantitative variables, only the French deprivation index is concerns, the results and their interpretation are much clearer with quintile splitting, so this is the form we retained. New correlations between covariates were found, involving Urban/Rural household status and FDep. We therefore built two separate models, Model 3 containing the covariates Age, Length of stay, Charlson Comorbidity Index, plus the interaction term between Charlson score and length of stay, FDep and Private/Public status. Model 4 contains the same covariates, but replacing the deprivation index, FDep, by Urban/Rural status.

9.3.2 Association of individual and environmental factors with the risk of rehospitalization

The results obtained for both models are presented in Table 9.4, page 35.The results of the analyses of the individual and environmental factors associated with the risk of rehospitalization are presented as odds ratio (OR), 95% confidence interval (CI) of the odds ratio, and p-value indicating the significance of each covariate in the risk of 30-day rehospitalization.

	Model 3[(1)]			**Model 4**[(1)]		
	OR	**95% CI**	***p*-value**	**OR**	**95% CI**	***p*-value**
Age (years)						
$\leq$ 75 years	ref	ref	ref	ref	ref	ref
$>$ 75 years	1.348	1.312-1.384	$<$0.0001	1.352	1.316-1.390	$<$0.0001
Length of stay (days)						
$\leq$ 4 days	ref	ref	ref	ref	ref	ref
$>$ 4 days	1.642	1.582-1.705	$<$0.0001	1.626	1.564-1.690	$<$0.0001
Charlson Comorbidity Index[(2)]						
0	ref	ref	ref	ref	ref	ref
1	1.283	1.238-1.329	$<$0.0001	1.277	1.232-1.324	$<$0.0001
$\geq$ 2	1.837	1.768-1.909	$<$0.0001	1.826	1.755-1.899	$<$0.0001
Urban/Rural status						
Urban	-	-	-	ref	ref	ref
Rural	-	-	-	1.040	1.007-1.074	0.0159
French Deprivation index						
$< P_{20}^{(3)}$	0.889	0.844-0.937	$<$ 0.0001	-	-	-
$[P_{20}; P_{40}[$	0.965	0.917-1.016	0.1721	-	-	-
$[P_{40}; P_{60}[$	ref	ref	ref	-	-	-
$[P_{60}; P_{80}[$	1.002	0.954-1.052	0.9464	-	-	-
$\geq P_{80}$	1.048	0.999-1.099	0.0574	-	-	-
Private/Public status						
Private	ref	ref	ref	ref	ref	ref
Public	1.480	1.440-1.520	$<$0.0001	1.483	1.443-1.525	$<$0.0001

[(1)]Adjustment with consideration of CCI $\times$ LOS interaction terms, [(2)]Charlson comorbidity index excl. inclusion conditions, [(3)]P_{20}, P_{40}, P_{60} and P_{80} are the first (-1.1098), second (-0.2290), third (0.2946) and fourth (0.9379) quintiles.

Table 9.4: Multivariate models predicting the risk of 30-day rehospitalization by multilevel logistic regression (Charlson comorbidity index version)

The results point in the same direction, but with an even higher assessed risk (82.6-83.7%) for patients with a Charlson score greater than 2, than the risk assessed using the Cox PH model.

With regard to environmental factors, measuring their effect was not the main issue; we simply wanted to take account of their impact on the analyses; nevertheless, we found that the derivation index shows that patients living in a deprived area have a lower risk (11.1%) of rehospitalization than patients living in areas considered to be average socio-economic zones. In addition, patients who underwent surgery in a public hospital had a 48% higher risk of 30-day rehospitalization than patients who underwent surgery in a private hospital.

Chapter 10

Conclusion and perspectives

The aim of the project was to study the impact of individual and environmental factors on the risk of 30-day rehospitalization, taking into account not only age and length of stay, but also a number of individual clinical factors, as well as socio-economic factors.

10.1 Conclusion

Drawing from the outcomes of these analyses, we can begin to identify the key factors that contribute significantly to the risk of 30-day rehospitalization.

Looking at age, it is reasonable to observe that older patients have an increased risk of 30-day rehospitalization. This observation is supported by the results of the study, which show a positive correlation for patients aged 75 years and older.

One of the main factors considered was length of stay, a variable that provided a significant amount of information in the models but was complex to interpret. As with age, it seemed likely that the longer the length of stay, the stronger the impact of the surgery on the patient and the higher the risk of 30-day rehospitalization. This study highlighted this by showing that a length of stay over 4 days was positively associated with the risk of rehospitalization.

The influence of comorbidity indices is noteworthy, particularly as the severity of certain medical conditions emerges as a significant risk factor, as is the case with the Charlson index. The analyses revealed a robust positive correlation between a Charlson score of 2 or higher and the likelihood of 30-day rehospitalization.

In addition, by evaluating the Elixhauser comorbidity groups independently, we were able to identify comorbidity groups that had a stronger impact on the risk of rehospitalization than others. Among these, 3 groups - Renal Failure, Metastatic Cancer and Solid Tumor without Metastasis - showed strong positive associations with the risk of rehospitalization.

The consideration of environmental factors confirmed the previously found results, attributing a stronger impact to patients with a Charlson score of 2 or higher. The FDep findings suggest a trend towards non-rehospitalization in deprived areas and increased rehospitalization in advantaged areas based on the data set. This may be influenced by factors such as the prevalence of patients undergoing prostate surgery in private hospitals, which may be less accessible in deprived areas. They also raise another concern, as the private or public status of the hospital where the surgery was done, seems to play a role in the risk of rehospitalization within 30 days. This difference between public and private status can be due to a number of factors, from higher costs in private hospitals to the fact that public hospitals tend to receive more serious cases.

10.2 Strength and limitation

We worked almost exclusively with PMSI databases for this study. We also used the INSEE open access database to reconstitute environmental factors, but no additional patient information was available. Working on the PMSI has a lot of strengths: we can observe a large number of patients, and we have access to all public and private hospitalizations, which means that we can identify rehospitalizations of the same patient. We are able to retrieve a large amount of patient characteristics and the main causes of hospitalization.

However, we have very limited information about other factors that may affect a patient's health, such as smoking or alcoholism that are underestimated in the PMSI database since they do not have a direct impact on the patient care. In addition, we only observe hospital admissions, so we don't know each patient's care pathway, whether that includes out-of-hospital care including treatments.

For environmental factors, we only have an aggregated code, so reconstitution is very limited, with missing data. Our ability to take any environmental impact into account is therefore limited.

Finally, the study is a first line of research carried out without the active participation of a clinician, possibly implying shortcomings in terms of clinical factors, such as the cancer ranks and types of procedure undergone, which could possibly have been retained in the study.

10.3 Possible future research line

First of all, as explained previously, GLMMs built with a logit link function (i.e., multilevel logistic regression) can be considered for this type of study, but their weakness is that they don't take into account right censoring. It could therefore be interesting to work on models that take into account both censoring and random effects, for example with so-called frailty models (Commenges and Jacqmin-Gadda, 2015, p.247 [6]). Frailty models are Cox proportional hazard models with mixed effects; the term frailty model is used to denote a survival regression model (typically a Cox proportional-hazards model) that incorporates random effects.

The second point is to consider out-of-hospital data and the ambulatory aspect. The aim would be to include the type of care path taken by patients, between their discharge from hospital (n) and their eventual readmission ($n+1$), to the study. In this way, we could try to jointly evaluate hospital and primary care factors, in order to identify the determinants that are relevant to both sectors. These kind of data are contained in the *Système national inter-régimes de l'assurance maladie* (SNIIRAM). In addition to PMSI data, SNIIRAM includes data on the consumption of ambulatory care (i.e. all reimbursed services, with detailed coding of the service); on the consumption of hospital care, in particular hospital ambulatory activity, as well as drugs and medical supplies billed "in addition" to fixed-rate charges; and also on the pathologies treated.

Finally, even if the type of medical procedure was not retained in this study, it still seems that a reconsideration should be made. Especially by working on the categorization of the variable according to the opinion of a clinician, based on his knowledge of the most frequent and/or most burdensome/impactful surgeries for patients. As the database corresponds to CCAM codes, it is difficult to propose other categorizations without the opinion of a clinician.

Ultimately, the combination of these points could lead to a better understanding of the individual and socio-economic factors associated with patient re-hospitalization. This would enable to guide better patient follow-up according to the specific patient characteristics that proved to be discriminating in this study, but also from a more socio-economic point of view.

Part IV

Challenges encountered during the work-study program

Chapter 11

First step into the study

11.1 Lack of clinical knowledge

The first steps in the study were the trickiest, because with an extremely mathematical eye, the factors to be considered in such a medical study seemed unclear. The literature review shed light on a number of issues, but most prostate studies were carried out using data from dedicated centers, and designed to study a specific procedure (e.g. Robot-assisted radical prostatectomy (RARP) [20], Transurethral resection of the prostate (TURP) [21]). The literature on the use of PMSI data to study the risk of rehospitalization after prostate surgery remains relatively poor. In addition, the proposal included important factors to consider, but was not intended to be an exhaustive list. Specifically, cancer rank and medical procedures were variables beyond those specified in the protocol. We chose to include these in the study because they seemed relevant. Although the univariate results were significant, in multivariate analysis things got more complicated (e.g., correlation, PH assumption not respected, significant results for Grønnesby and Borgan).

11.2 Model building

Another major challenge of this study was to cope with the conditions for applying the model, whether in terms of correlations or assumptions of a Cox PH model.

11.2.1 Correlations

As previously mentioned, inter-covariate correlations can introduce substantial bias to the study. Notably, the Elixhauser index exhibited correlations with all covariates. A solution emerged during the initial project presentation, a member of the jury, Hervé Cardot, suggested the utilization of a classification system to identify the most influential factors. While classification was initially impractical because we had to select the information to be reconstructed, a workaround emerged to address the correlation problems associated with the Elixhauser index. The idea was to work with the underlying comorbidity groups instead of the index itself. In essence, considering the 31 groups independently, and then define which of them are the most relevant; this is where classification comes into its own. To streamline the process without sacrificing information, we opted to focus on 10 uncorrelated groups. The objective was twofold: addressing the incorporation of the Elixhauser index, a frequently cited index in literature, and conducting a correlation study involving 10×10 variables.

11.2.2 Assumptions

The assumptions required to apply a Cox PH model are essential, and their verification is mandatory. There are many different methods in the literature, and we have chosen to consider two for each assumption and type of variable. This choice to consider several methods is first of all due to their power - some methods are not sensitive enough - and the fact that the method based on martingale residuals offered by PHREG procedure is extremely time-consuming. Indeed, we worked with 270473 patients, implying that running a procedure with the ASSESS statement only on age covariate took 47 hours. So we had to demonstrate very good time management with the entire department, which is by no means easy. To overcome this situation, and to be able to move forward despite the fact that the results have not yet been obtained, we spent a lot of time implementing other methods and mathematically proving their relevance.

Aside from the concerns over the methods, there was one variable that particularly bugged us, this variable was age. The results presented in section 4.4.2.2 and section 4.4.3 showed us that including age as such wasn't conceivable since it assumes neither loglinearity nor hazards proportionality. Given the nature of the study, we chose to categorize it. A first study was carried out on loglinearity, dividing the variable using the Figure 4.1 (cutpoints version): < 65, 65 to 74 and 75 or older. As can be seen in Figure 11.1, this categorisation did not meet the loglinearity assumption. After a few trials, we found a suitable categorization according to rehospitalization trends, we defined the following intervals (trends version): < 70, 70 to 79 and 80 or older.

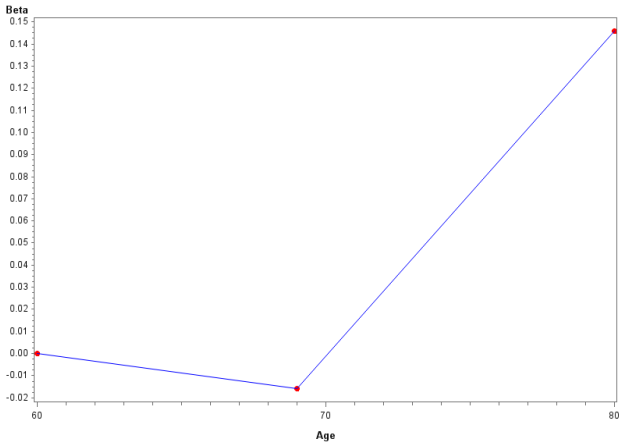

Figure 11.1: Beta estimates against 3-level age covariate (cutpoints version).

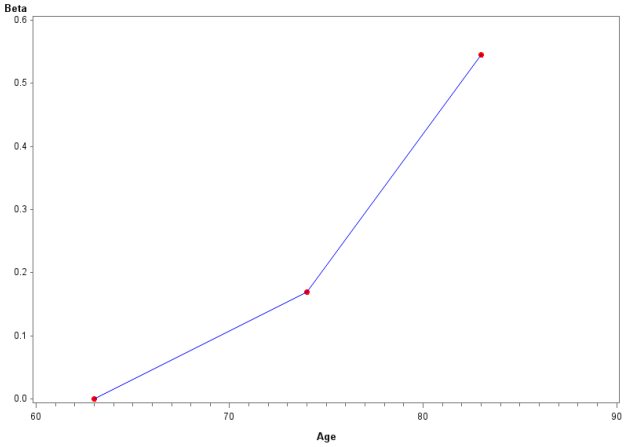

Figure 11.2: Beta estimates against 3-level age covariate (trends version).

Compared with Figures 4.1, and 11.1, we can clearly observe that the 3-level (trends version) categorization (Figure 11.2) seems better suited to comply with loglinearity assumption. All that remained was to check hazards proportionality. We realized that the categorization previously found was inadequate since the 3-level categorization curves represented in Figure 11.3 are found to intersect, representing a violation of the PH assumption. This finding made our previous work nugatory. Attempts were made to define another three-level categorisation, but none of them agreed with the two hypotheses. We decided to consider a 2-level categorization using the median, which was also unsuitable. As a last solution, we studied the trends in rehospitalization in greater depth. We noticed that the trend changed at the 75-year threshold, which is why we chose this cut.

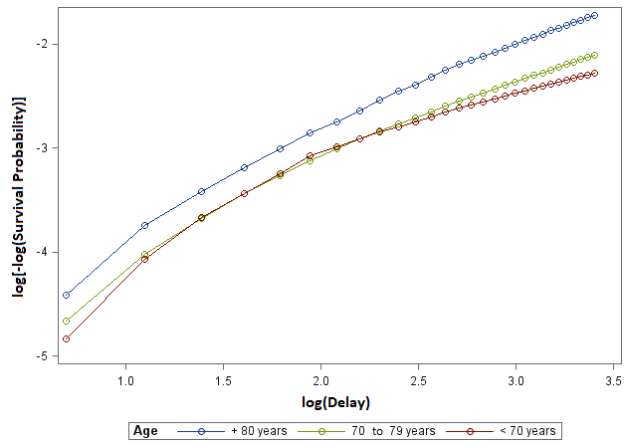

Figure 11.3: Log of negative survivor log estimated for 3-level categorization of Age covariate

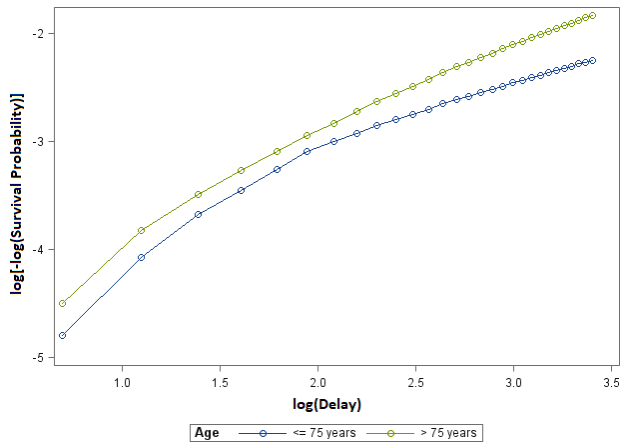

Figure 11.4: Log of negative survivor log estimated for 2-level categorization of Age covariate

The curves for 2-level covariate age, in Figure 11.4, look roughly parallel, and by working with a binary covariate log-linearity is no longer an issue; therefore this choice seems to be the right one. To ensure this, we introduced an interaction between time and this covariate, the details of which are given in Appendix D.1. This work was particularly tedious, but necessary if we wanted to include age covariate in a Cox PH model. Finally, the 2-level categorization preserves the significance of the covariate while simplifying the interpretation of the hazard ratio. We have therefore succeeded in finding the best compromise.

Chapter 12

Methods not proposed by SAS

12.1 Cramér's V Heatmap

It's easy enough to obtain Cramér's V with SAS, we just have to run the FREQ procedure with the '`chisq`' option as mentioned for assumption testing with Pearson's chi-square test statistic. Indeed, Cramér's V is one of the outputs of this procedure. However, the output of this procedure is difficult to manage and does not allow you to print only the Cramér's V. SAS output prints test results for several different statistics, so when you need to test the correlation between each covariate 2 by 2 for more than a dozen covariates, it becomes extremely tedious to go through all the output. As mentionned in the section 5, some time was spent creating a macro to build a Camér's V matrix on the same principle as the Pearson Correlation matrix. We wanted to keep the variable names, their labels and obtain a symmetrical matrix, while at the same time having a code that could run fairly efficiently given the large number of patients included in the study. The macro was carefully designed to retrieve only the '`chisq`' output from the FREQ procedure and store it in an array. These arrays were then merged successively.

Given that our academic background was limited to a simple introduction to SAS, this part of creating the macro took a little time to set up, but was definitely worth it. Furthermore, the combined use of IML procedure to build a heatmap makes it possible to visualize all correlations extremely efficiently. Regarding the IML procedure, the code remains relatively simple, we've just adapted the procedure's functions. This challenge helped us be more efficient and will be useful in future studies.

12.2 Grønnesby and Borgan

In epidemiological research, it's very common to use Hosmer-Lemeshow when fitting a logistic regression. This is a crucial step in ensuring that the model is appropriate and especially high quality. However, when using Cox PH model, the ground is extremely unclear. Most Pubmed publications, articles and open-access courses on Cox regression check the viability of the model based exclusively on the assumptions of its construction. Studies carried out on SAS focus almost entirely on the ASSESS tool of the PHREG procedure, which ultimately doesn't give us as much information on the adequacy of the model. References to a method similar to Hosmer-Lemeshow are scarce, and it was only through in-depth exploration of the Hosmer-Lemeshow approach and identification of a corresponding adaptation for Cox regression that Grønnesby and Borgan's method emerged as a viable implementation. This feasibility was facilitated by the guidance provided by May and Hosmer (May and Hosmer, 1998 [16]).

It's hard to imagine working without such a tool, as it's useful for confirming covariate selection, considering possible interactions, and validating correct hazard ratio construction. Without this tool, building a Cox model would have been a fuzzy process. That's why it was so rewarding to implement and study this method. Even if its implementation was relatively easy, thanks in particular to the fact that martingale residuals were obtained via the '`resmart`' option in the PHREG procedure, the real struggle was to find papers attesting to its reliability and about the number of risk score groups to be set up. Therefore, we have chosen to focus on the first paper published by Gronnesby and Borgan, and the papers published by May and Hosmer (May and Hosmer, 2004 [17]).

Appendices

Appendix A

A.1 Inclusion CCAM and ICD-10 codes

A.1.1 Prostate surgery procedures considered for inclusion

The CCAM codes mentioned in the protocol for the inclusion of patients in the study are classified[1] as follows:

- **Anesthesia Procedures:**
  - **JGFA015:** Urethrocystoscopic resection of prostatic hypertrophy
- **Surgical procedures:**
  - **JGFA005:** Transvesical adenomectomy of the prostate, by laparotomy
  - **JGFA006:** Total vesiculoprostatectomy by laparotomy
  - **JGFA007:** Retrovesical or transvesical removal of the utricle of the prostate by laparotomy
  - **JGFA009:** Retropubic or transcapsular adenomectomy of the prostate, by laparotomy
  - **JGFA011:** Total vesiculoprostatectomy, perineal approach
  - **JGFA014:** Palliative prostate resection [Urethral recalibration], by urethrocystoscopy
  - **JGFA016:** Urethrocystoscopic resection or marsupialization of a prostate collection or urethral diverticulum
  - **JGNJ900:** Rectal destruction of prostate lesions using high-intensity focused ultrasound
  - **JGFC001:** Total vesiculoprostatectomy by coelioscopy
  - **JGFE023:** Non-laser urethrocystoscopic resection of prostatic hypertrophy
  - **JGFE365:** Laser resection of an enlarged prostate using urethrocystoscopy
  - **JGNE171:** Destruction of prostate hypertrophy by laser [photovaporization], urethrocystoscopy
- **Technical medical procedures:**
  - **JGJB001:** Finger-guided transrectal or transperineal evacuation of prostate collections
  - **JGND002:** Prostate Cryotherapy
  - **JGHB001:** Transrectal or transperineal puncture-cytoaspiration of the prostate gland
  - **JGNE003:** Destruction of prostatic hypertrophy by radiofrequency urethrocystoscopy with ultrasound guidance
  - **JGNL001:** Prostate brachytherapy with permanent insertion of iodine-125
  - **JGNJ001:** Destruction of prostate hypertrophy by microwaves [Thermotherapy of the prostate].
  - **JGHB002:** Finger-guided transperineal prostate biopsy
  - **JGHD001:** Finger-guided transrectal prostate biopsy
  - **JGHJ001:** Transrectal ultrasound-guided prostate biopsy
  - **JGHJ002:** Transperineal ultrasound-guided prostate biopsy

[1] According to the official CCAM website: `https://www.ameli.fr/accueil-de-la-ccam/index.php`.

A.1.2 Prostate cancer pathologies considered for inclusion

Concerning pathologies, the ICD-10 codes used to identify the cancers concerned by the inclusion are as follows.

- **Very high rank:**
  - **C61:** Malignant Prostate Tumor
- **High rank:**
  - **D07.5:** Carcinoma in situ Prostate (High-grade dysplasia)
  - **D29.1:** Benign Prostate Tumor
  - **D40.0:** Unpredictable (or unknown evolution) Prostate Tumor
- **Moderate rank:**
  - **N40:** Benign Prostatic Hyperplasia (or Hypertrophy)
  - **N42.3:** Low Grade Prostatic Dysplasia

All pathologies beginning with the above codes will be considered for inclusion in the study.

A.1.3 Specific stays to be excluded

The GHM codes for stays for therapeutic purposes/iterative stays are listed below:

- CMD 27 / CMD 28 (organ transplants and sessions)
- **11K02:** Renal dialysis
- **17M05:** Chemotherapy for acute leukemia
- **17M06:** Chemotherapy
- **17K04:** Irradiation
- **17K05:** Prostate brachytherapy
- **17K06:** Other brachytherapy or internal irradiation
- **17K08:** Brachytherapy for all locations, excluding iodine seeds
- **17K09:** Internal irradiation
- **23M09:** Chemotherapy for non-tumor diseases
- **02C05 / 02C12:** Stays for cataract surgery

A.2 Hospitalization chaining explanation scheme

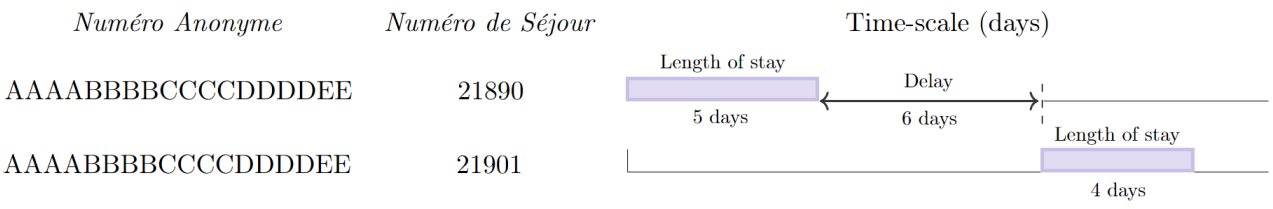

Figure A.1: Schematic diagram of hospitalization chaining

Taking the example of an artificial patient identified as AAAABBBBCCCCDDDDEE, we add the *Numéro de Séjour*, 21890, to the length of stay, in this case 5 days, thus the stay ends in 21895. The following stay begins at 21901, so we have a delay of 21901 minus 21895 days, or 6 days, meaning the patient was rehospitalized within 30 days.

A.3 ICD-10 codes for comorbidity indices

For both indices, we chose to exclude the pathologies considered to be included: C61, D07.5, D29.1, D40.0, N40 and N42.3, to prevent bias in the models. We based their construction on the article published by H. Quan et al. (H. Quan et al., 2005 [22]).

A.3.1 Charlson comorbidity index

Comorbidities	**Weights**	**ICD-10**
Myocardial infarction	1	I21.x, I22.x, I25.2
Congestive heart failure	1	I09.9,I11.0, I13.0, I13.2, I25.5, I42.0, I42.5-I42.9, I43.x, I50.x, P29.0
Peripheral vascular disease	1	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Cerebrovascular disease	1	G45.x, G46.x, H34.0, I60.x–I69.x
Dementia	1	F00.x–F03.x, F05.1, G30.x, G31.1
Chronic pulmonary disease	1	I27.8, I27.9, J40.x–J47.x, J60.x–J67.x, J68.4, J70.1, J70.3
Rheumatic disease	1	M05.x, M06.x, M31.5, M32.x–M34.x, M35.1, M35.3, M36.0
Peptic ulcer disease	1	K25.x–K28.x
Mild liver disease	1	B18.x, K70.0–K70.3, K70.9, K71.3–K71.5, K71.7, K73.x, K74.x, K76.0, K76.2–K76.4, K76.8, K76.9, Z94.4
Diabetes without chronic complication	2	E10.0, E10.l, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, El2.6, E12.8, El2.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes with chronic complication	2	E10.2–E10.5, E10.7, E11.2, E11.5, E11.7, E12.2–E12.5, E12.7, E13.2–E13.5, E13.7, E14.2–E14.5, E14.7
Hemiplegia or paraplegia	2	G04.1, G11.4,G80.1, G80.2, G81.x, G82.x, G83.0–G83.4, G83.9
Renal disease	2	I12.0, I13.1, N03.2–N03.7, N05.2–N05.7, N18.x, N19.x, N25.0, Z49.0–Z49.2, Z94.0, Z99.2
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	2	C00.x–C26.x, C30.x–C34.x, C37.x–C41.x, C43.x, C45.x–C58.x, C60.x–C76.x, C81.x–C85.x, C88.x, C90.x–C97.x
Moderate or severe liver disease	3	I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7
Metastatic solid tumor	6	C77.x–C80.x
AIDS/HIV	6	B20.x–B22.x, B24.x

Table A.1: ICD-10 Coding algorithms for Charlson comorbidity index

A.3.2 Elixhauser index

Comorbidities	ICD-10
Congestive heart failure	I09.9, I11.0, I13.0, I13.2,I25.5, I42.0, 142.5-I42.9, I43.x, I50.x, P29.0
Cardiac arrhythmias	I44.1-I44.3, I45.6, I45.9, I47.x-I49.x, ROO.O, ROO.1, ROO.8, T82.1, Z45.0, Z95.0
Valvular disease	A52.0, I05.x-I08.x, I09.1, I09.8, I34.x-I39.x, Q23.O-Q23.3, Z95.2, Z95.4
Pulmonary circulation disorders	I26.x, I27.x, I28.0, I28.8,I28.9
Peripheral vascular disorders	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8,K55.9, Z95.8, Z95.9
Hypertension, uncomplicated	I10.x
Hypertension, complicated	I11.x-I13.x, I15.x
Paralysis	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0-G83.4, G83.9
Other neurological disorders	G10.x-G 13.x, G20.x-G22.x, G25.4, G25.5, G31.2, G31.8,G31.9, G32.x, G35.x-G37.x, G40.x, G41.x, G93.1, G93.4, R47.0, R56.x
Chronic pulmonary disease	I27.8, 127.9, J40.x-J47.x, J60.x-J67.x, J68.4, J70.1, J70.3
Diabetes, uncomplicated	E10.0, E10.1, E10.9,E11.0, E11.1, E11.9, E12.0, E12.1, E12.9, E13.0, E13.1, E13.9, E14.0, E14.1, E14.9
Diabetes, complicated	E10.2-E10.8,E11.2-E11.8, E12.2-E12.8, E13.2-E13.8, E14.2-E14.8
Hypothyroidism	E00.x-E03.x, E89.0
Renal failure	I12.0, I13.1, N18.x, NI9.x, N25.0, Z49.0-Z49.2, Z94.0, Z99.2
Liver disease	B18.x, I85.x, I86.4, I98.2, K70.x, K71.1, K71.3-K71.5, K71.7, K72.x-K74.x, K76.0, K76.2-K76.9. Z94.4
Peptic ulcer disease excluding bleeding	K25.7, K25.9, K26.7, K26.9, K27.7, K27.9, K28.7, K28.9
AIDS/HIV	B20.x–B22.x, B24.x
Lymphoma	C81.x-C85.x, C88.x, C96.x, C90.0, C90.2
Metastatic cancer	C77.x-C80.x
Solid tumor without metastasis	C00.x-C26.x, C30.x-C34.x, C37.x-C41.x, C43.x, C45.x-C58.x,C60.x-C76.x, C97.x
Rheumatoid arthritis, collagen vascular diseases	L94.0, L94.1, L94.3, M05.x, M06.x, M08.x, M12.0, M12.3, M30.x,M31.0-M31.3,M32.x-M35.x, M45.x, M46.1, M46.8, M46.9
Coagulopathy	D65-D68.x, D69.1,D69.3-D69.6
Obesity	E66.x
Weight loss	E40.x-E46.x, R63.4, R64
Fluid and electrolytedisorders	E22.2, E86.x, E87.x
Blood loss anemia	D50.0
Deficiency anemia	D50.8, D50.9, D51.x-D53.x
Alcohol abuse	F10, E52, G62.1, I42.6, K29.2, K70.0, K70.3, K70.9, T51.x, Z50.2, Z71.4, Z72.1
Drug abuse	F11.x-F16.x, F18.x, F19.x, Z71.5. Z72.2
Psychoses	F20.x, F22.x-F25.x, F28.x, F29.x, F30.2, F31.2, F31.5
Depression	F20.4, F31.3-F31.5, F32.x, F33.x, F34.1, F41.2, F43.2

Table A.2: ICD-10 Coding algorithms for Elixhauser index

Appendix B

B.1 Survival analysis basis

This appendix section is designed for those with limited knowledge of survival analysis. It quickly covers the key points of survival analysis, so that the reader has all the elements needed to understand the mathematical tools and arguments used in Part II.

B.1.1 Survival data

The term **survival data** is used when there is a delay before an event. The analysis of this type of data is called survival analysis and corresponds to the study of the time to occurrence of the event of interest, called time-to-event. In our study, the event of interest was the first 30-day rehospitalization of the patient after an admission for prostate surgery. One of the main characteristics of survival analysis is the difficulty of completely observing all the time-to-events. For example, in our case, the event being studied is a 30-day rehospitalization; for patients who are not rehospitalized within 30 days, the event date is not observed. This type of observation is called right censoring.

In order to analyze the influence of covariates / factors on the survival of the patients, regression models are used. A logistic regression model can be used to study the association between covariates and the risk of event occurrence. In this case, the qualitative binary outcome is whether or not the event occurs within 30 days. This simple approach results in a loss of information and completely ignores the phenomenon of right censoring. This is why, in the presence of right-censoring, or of patients lost to follow-up, we use a Cox PH model.

There are three main principles to understand for this study:

- **Delay of interest:** Here, the delay is based on the patient's own date of origin (other than birth), i.e. when he was admitted for prostate surgery. This will give us a relevant delay before the event of interest.
- **The event of interest:** Our study is based on the patient's first rehospitalization (defined by the protocol), so the definition of this event is clear and occurs at a precise point in time.
- **End date:** This is the date after which the patient's information will no longer be taken into account. Here, it is defined as 30 days.

The survival time is usually represented by a random variable $T > 0$. Here, we assume that T is continuous on $\mathbb{R}^+$. The probability distribution of this variable is characterized by the following two functions.

- *Probability density function:*

$$f(t) = \lim_{\delta t \to 0^+} \frac{P(t \leq T < t + \delta t)}{\delta t}$$

  We assume that the limit exists for all t, so the probability of occurrence of the event for δt small is equal to $f(t)\delta t$.

- *Distribution function:*

$$F(t) = P(T \leq t)$$

The most common approach is to estimate the *survival function*, since it represents the probability of undergoing the event beyond t, i.e. of not having been rehospitalized at t. It is characterized as follows,

$$S(t) = P(T > t).$$

B.1.2 Kaplan-Meier

The **Kaplan-Meier estimator** (Kaplan et Meier 1958) is the non-parametric maximum likelihood estimator of the survival function. It is used to estimate and plot the survival function $S(\bullet)$ from a sample of patients with times (= survival time) that can be right-censored. It is often the first step in carrying out the survival analysis, as it is the simplest approach.

The estimator is defined as the fraction of observations who survived for a certain amount of time under the same circumstances and is given by the following expression:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{B.1}$$

where t_i is the time at which at least one event occurred, with $t_1 < t_2 < \cdots < t_i < \cdots < t_k$; d_i is the number of events that occurred at time t_i; n_i is the number of individuals known to have survived to time t_i, this is the number of observations at risk at time t_i.

By definition we have $\hat{S}(0) = 0$. The function $\hat{S}(t)$ is a decreasing step function, constant between two consecutive event times, continuous on the right, with a jump at each observed event time.

The logic behind expression (B.1) is quite intuitive: calculating the probability of not yet having experienced the event at t_i is equivalent to calculating the probability of not having experienced the event in t_{i-1} and the probability of not having experienced it in t_i, knowing that the event has not occurred until t_{i-1}.

Kaplan-Meier curves are very popular in survival analysis, making it possible to assess relatively straightforwardly whether one group is more at risk than another.

B.2 More about Cox assumptions

Code extracts and modifications used to check and adjust for the assumptions of a Cox model are given in this section. The code associated with the graph for checking the age covariate is given in the first subsection. Next, an example of modality grouping with the medical procedure (MP) covariate is given to illustrate the use of loglogs curves and the graphical results.

B.2.1 Checking loglinearity

As explained in section 4.4.1.2, we wanted to graphically represent the beta slope by interpolating with a dozen points. We started by determining the deciles of our age variable, then created a `Age Dec` variable which will be constructed by assigning one level per decile. We then built a Cox PH model on this single variable, using the CLASS statement to obtain one coefficient per level, as follows.

```
PROC PHREG DATA=Patients;
        CLASS Age_Dec (ref='0') / param=ref;
        MODEL Delay*RH30(0)= Age_Dec;
RUN;
```

This displays the coefficients associated with each level, and we now need to get the midpoint values, which will be done using MEANS procedure and the '`median`' option, as shown opposite.

```
PROC MEANS DATA=Patients median;
        VAR Age;
        CLASS Age_Dec;
RUN;
```

Then we created a table containing the estimate coefficients and their respective midpoints, so that they can be easily represented graphically. Here we're doing it by hand, as we're only dealing with ten or so points, but it's possible to extract them all using the OUTPUT statment.

```
DATA AgeGraph;
        INPUT Age Beta;
        CARDS;
                55          0
                59          -0.03933
                62          0.02123
                65          0.02688
                67          0.06253
                69          0.0970
                72          0.13726
                75          0.19942
                79          0.37610
                85          0.62485
        ;
RUN;
```

Finally, we plotted the whole using the GPLOT procedure and a few graphical options.

```
PROC GPLOT DATA=AgeGraph;
        SYMBOL interpol=join ci=blue value=dot height=1 cv=red;
        PLOT Beta*Age /frame;
RUN;
QUIT;
```

B.2.2 Medical procedure covariate and PH assumptions

Initially, the covariate concerning the type of medical procedure undergone at index hopsitalization was divided into three categories in accordance with the CCAM classification. Upon examination, unfortunately, we found that this categorization did not satisfy the PH assumption. We therefore chose to group together two levels of this covariate, since considering another categorization was impossible given our level of clinical knowledge.

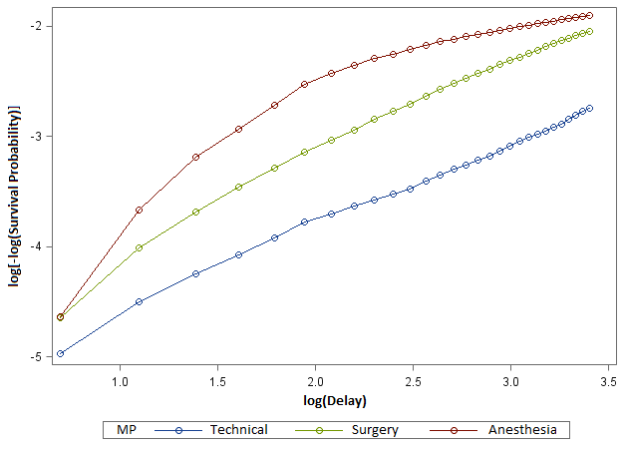

Figure B.1: Log of negative survivor log estimated for 3-level of MP covariate

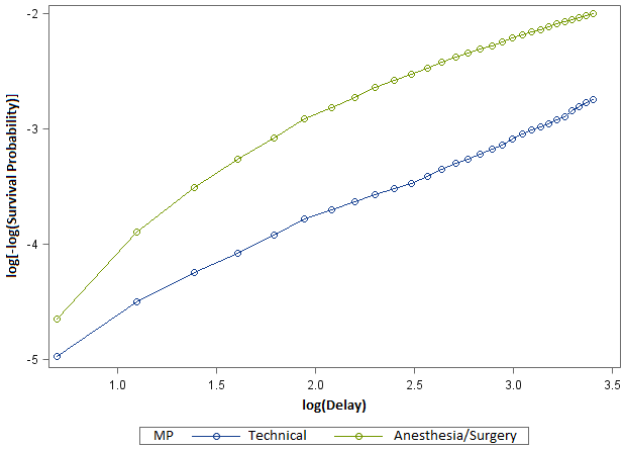

Figure B.2: Log of negative survivor log estimated for 2-level of MP covariate

As can be seen from the figure B.1, the level corresponding to the medical anesthesia procedure does not behave proportionally to the surgical procedure, hence we chose to group it with the closest modality, i.e. surgery.

B.3 Cramér's V

As mentioned in Chapter 5, some time was spent developing a macro that generates a Camér's V matrix on the same principle as Pearson's correlation matrix. This macro is supplied with a heatmap whose legend is adapted to the 0.15 tolerance threshold. The code is given below.

B.3.1 Macro implementation

```
%macro cramv_matrix(table, varlist, lablist, n, CRAMV_Matrix);
        /* Matrix initialization */
        DATA &CRAMV_Matrix;
                ATTRIB Variable length = $10.; /* format suitable for our study */
                ATTRIB Label length = $39.;
                /* Retrieve information of parameter variables */
                %do i=1 %to &n;
                        %let v1 = %scan(&varlist,&i.,'-');
                        %let l1 = %scan(&lablist,&i.,'-');
                        Variable = "&v1"; Label = "&l1"; output;
                %end;
        RUN;
        /* Variables are crossed to obtain their Cramér's V correlation coefficient */
        %do i=1 %to &n;
                %let v1 = %scan(&varlist,&i.,'-');
                %do j=1 %to &n;
                        %let v2 = %scan(&varlist,&j.,'-');
                        PROC FREQ DATA=&table noprint;
                                TABLES &v1*&v2 / chisq;
                                OUTPUT OUT = CRAMV_&j.
                                        (keep = _CRAMV_ rename=(_CRAMV_=&v1))
                                        chisq;
                        RUN;
                %end;
                /* This gives us the column of correlation associated with &v1 */
                DATA CRAMV_&v1;
                        SET
                                %do j=1 %to &n;
                                        CRAMV_&j.
                                %end;
                        ;
                RUN;
                /* We then merge with the previous columns obtained */
                DATA &CRAMV_Matrix;
                        MERGE &CRAMV_Matrix CRAMV_&v1;
                RUN;
                /* Clean up the WORK library as you go along. */
                PROC DATASETS lib=work noprint;
                        DELETE CRAMV_&v1
                                %do j=1 %to &n;
                                        CRAMV_&j.
                                %end;
                        ;
                        RUN;
                QUIT;
        %end;
        /* Labels automatically assigned by SAS are removed for more clarity */
        PROC DATASETS lib=work noprint;
                MODIFY &CRAMV_Matrix;
                ATTRIB _all_ label=' ';
                RUN;
        QUIT;
%mend cramv_matrix;
```

B.3.2 Application with comorbidities

The heatpmap on Figure 5.2, presented in section 5.3, was obtained via the previous macro. As explained, we had to repeat the classification and correlation checking process until we had 10 groups of uncorrelated comorbidities. In order to present the use of the IML procedure and how we generated the heatmaps, the code used and the results are shown below.

```
/* Macro-variables declaration */
%let varlist = Age-LOTS-ELX_GRP_14-ELX_GRP_20-ELX_GRP_19-ELX_GRP_1-ELX_GRP_6-ELX_GRP_25-ELX_
GRP_3-ELX_GRP_10-ELX_GRP_26-ELX_GRP_22;

%let lablist = Age-Length of Stay-Renal Failure-Solid Tumor without Metastasis-Metastatic Ca
ncer-Congestive Heart Failure-Hypertension Uncomplicated-Fluid andElectrolyte Disorders-Valv
ular Disease-Chronic Pulmonary Disease--Blood Loss Anemia-Coagulopathy;

%let n = 12;
%let CRAMV_Matrix = mCRAMV_ELX;
%let table = Patients;

/* Macro call */
%cramv_matrix(&table, &varlist, &lablist, &n, &CRAMV_Matrix);

/* Using iml to visualize correlations */
PROC IML;
        USE &CRAMV_Matrix;
           read all var "Variable" into ColNames;  /* get names of variables */
           read all var "Label" into Labels;       /* get labels of variables */
           read all var (ColNames) into mCramVar;  /* matrix of Cramer's V */
        CLOSE &CRAMV_Matrix;
        Colors = palette('BRBG', 5);
        /* Heatmap */
        CALL HeatmapCont(mCramVar) xvalues = Labels  yvalues = Labels
           colorramp = Colors  range = {-0.15, 0.15} /* tolerance threshold */
           title = "Cramer's V Heatmap Elixhauser Comorbidities";
QUIT;
```

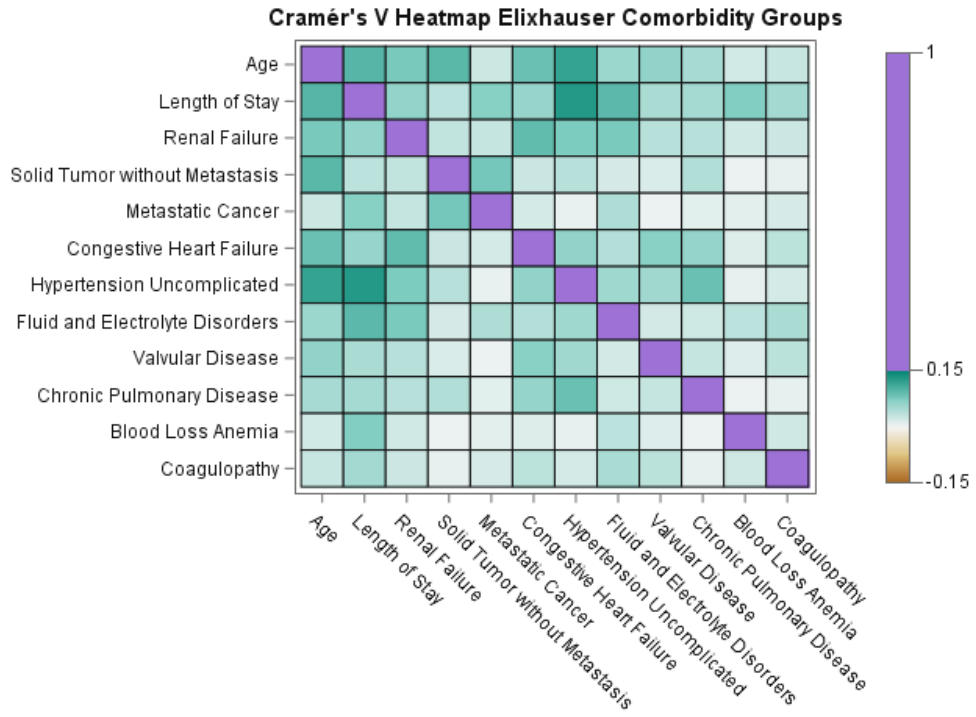

Figure B.3: Heatmap for Age, LOS and 10 Elixhauser comorbidity groups

We can see that none of the covariates exceed the defined correlation threshold.

B.4 Grønnesby and Borgan implementation

The 10 groups were clustered using two different methods. In both cases, the risk is obtained by adding an OUTPUT statment with 'xbeta' option to the PHREG procedure. We got martingale residuals with the 'resmart' option, so that we can calculate the number of expected events, as follows.

```
/* Generation of martingale residuals and risk score */
PROC PHREG DATA=Patients;
        CLASS Age (ref='0') LOTS (ref='0') CCI (ref='0');
        MODEL Delay*RH30(0) = Age LOTS CCI CCI*LOTS / rl;
        OUTPUT out=GB_Model_CCI resmart=resmart xbeta=risk;
RUN;

/* Expected number of events calculation */
DATA GB_Model_CCI;
        SET GB_Model_CCI;
        ATTRIB expected length = 3.;
        expected = RH30 - resmart;
RUN;
```

Once this has been done, we sorted the table by risk score, from the lowest to the highest, using SORT procedure. G groups are then created, in two different ways. One approach is to divide our data set into 10 groups and then consolidate the groups with the lowest number of events. The other approach is to simply divide the data set by the number of events, as soon as we have 10% in one group, we move on to the next group. Since these methods are very specific to the number of data events and their distribution, the code is as well, so we don't show this part.

Once the groups are created, the interesting part is the computation of the statistics, which is executed as follows[1].

```
/* Calculation of the expected and observed number of events per group */
DATA GB_Model_CCI (keep = G g_observed g_expected);
        SET GB_Model_CCI;
        BY G;
        ATTRIB g_observed g_expected length = 3.;
        RETAIN g_observed 0 g_expected 0 ;

        if first.G then do; g_observed = RH30; g_expected = expected; end;
        else do; g_observed + RH30;        g_expected + expected; end;

        if last.G then output;
RUN;

/* Calculation of the Z score per group */
DATA GB_Model_CCI;;
        SET GB_Model_CCI;
        g_deviation = g_observed - g_expected;
        z = g_deviation / sqrt(g_expected);
RUN;
```

In this way we recover only the observed and expected numbers per group, then we calculate a statistic per group on the same principle as Hosmer and Lemeshow.

Note: The p-values associated with the Z-statistics were calculated using the R language.

[1] The G variable corresponds to the variable indicating the assigned group, level-coded from 0 to 9.

Appendix C

C.1 Classification: Elixhauser comorbidity groups

		Learning	
Covariate	**Label**	**Relative**	**Importance**
ELX GRP 14	Renal Failure	1.0000	11.9015
ELX GRP 20	Solid Tumor without Metastasis	0.9353	11.1318
ELX GRP 19	Metastatic Cancer	0.7195	8.5637
ELX GRP 1	Congestive Heart Failure	0.7130	8.4861
ELX GRP 6	Hypertension Uncomplicated	0.5307	6.3162
ELX GRP 25	Fluid and Electrolyte Disorders	0.3702	4.4059
ELX GRP 3	Valvular Disease	0.3657	4.3520
ELX GRP 10	Chronic Pulmonary Disease	0.2998	3.5683
ELX GRP 26	Blood Loss Anemia	0.2642	3.1449
ELX GRP 22	Coagulopathy	0.2468	2.9371

Table C.1: Classification tree: the 10 most relevant Elixhauser comorbidity groups

C.2 Univariate analysis: Elixhauser comorbidity groups

	Overall Population	**No Rehospitalization**	**Rehospitalization**	***p*-value**
Elixhauser comorbidity groups[(1)]				
Renal Failure	4777 (1.77)	3662 (1.53)	1115 (3.66)	<0.0001
Solid Tumor without Metastasis	9469 (3.50)	7631 (3.18)	1838 (6.03)	<0.0001
Metastatic Cancer	2989 (1.11)	2286 (0.95)	703 (2.31)	<0.0001
Congestive Heart Failure	5448 (2.01)	4329 (1.80)	1119 (3.67)	<0.0001
Hypertension Uncomplicated	71103 (26.29)	61970 (25.82)	9133 (29.95)	<0.0001
Fluid and Electrolyte Disorders	3804 (1.41)	3067 (1.28)	737 (2.42)	<0.0001
Valvular Disease	3413 (1.26)	2714 (1.13)	699 (2.29)	<0.0001
Chronic Pulmonary Disease	9861 (3.65)	8373 (3.49)	1488 (4.88)	<0.0001
Blood Loss Anemia	1630 (0.60)	1288 (0.54)	342 (1.12)	<0.0001
Coagulopathy	1599 (0.59)	1247 (0.52)	352 (1.15)	<0.0001

[(1)]The 10 most relevant comorbidity groups.

Table C.2: Elixhauser comorbidity groups associated with rehospitalization within 30 days

C.3 PH Assumption: Elixhauser comorbidity groups

Once the study of correlations and classification has been carried out, we need to check that the covariates meet the assumptions of a Cox model. In this case, we're working with binary variables, so only the proportional-hazards assumption needs to be verified. To do this, we plotted the 10 loglogs curves associated with each group. As can be seen in the figures below, all groups met the PH assumption.

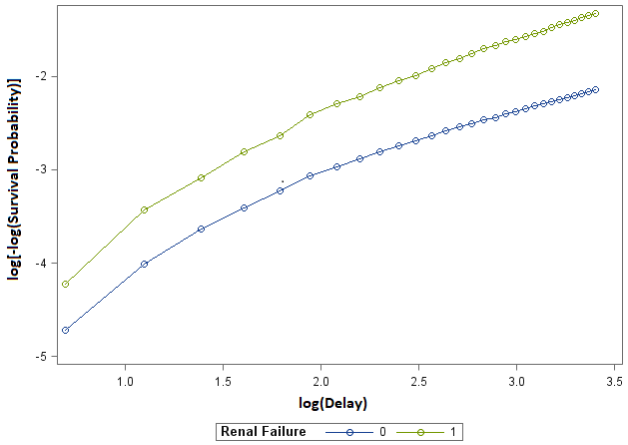

Figure C.1: Log of negative survivor log estimated for Renal Failure group

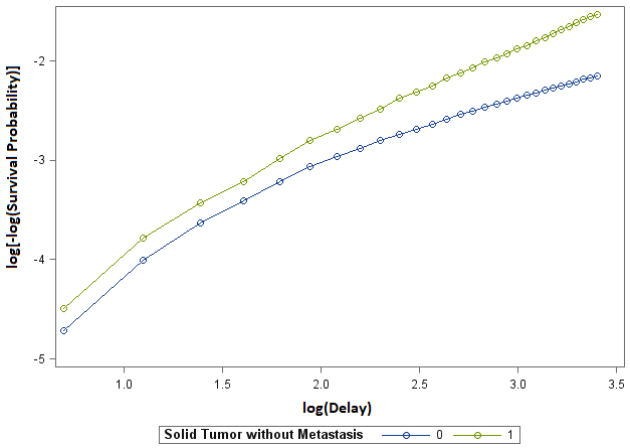

Figure C.2: Log of negative survivor log estimated for Solid Tumor without Metastasis group

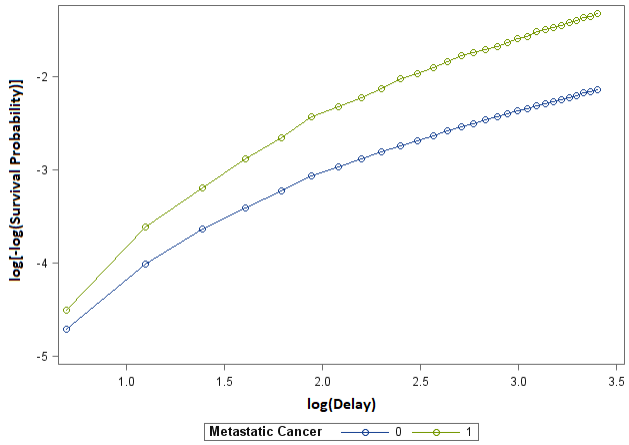

Figure C.3: Log of negative survivor log estimated for Metastasic Cancer group

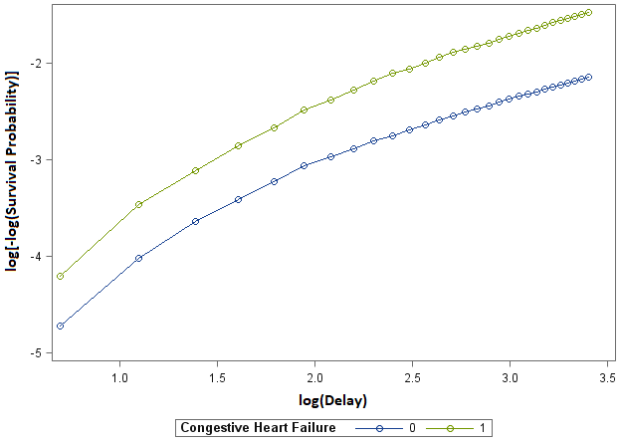

Figure C.4: Log of negative survivor log estimated for Congestive Heart Failure group

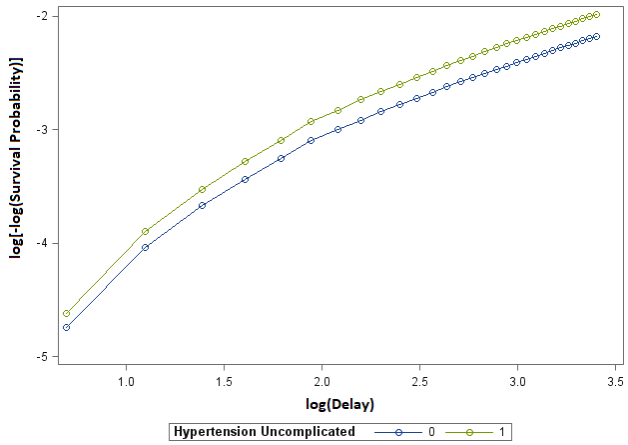

Figure C.5: Log of negative survivor log estimated for Hypertension Uncomplicated group

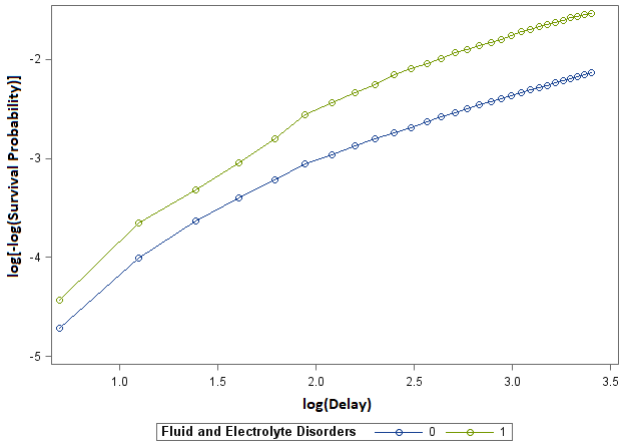

Figure C.6: Log of negative survivor log estimated for Fluid and Electrolyte Disorders group

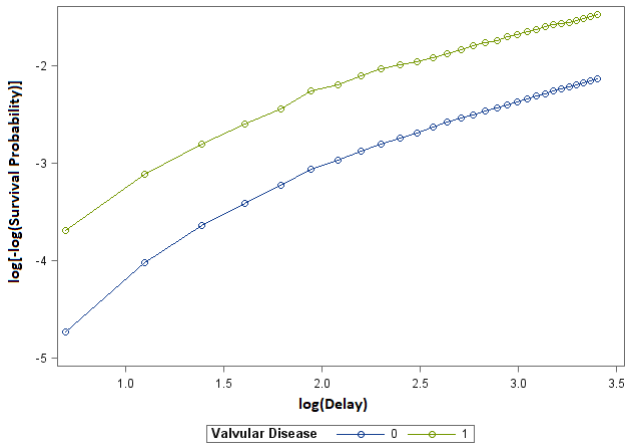

Figure C.7: Log of negative survivor log estimated for Valvular Disease group

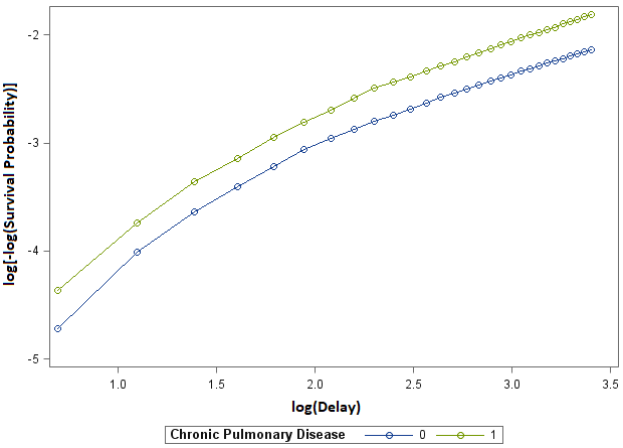

Figure C.8: Log of negative survivor log estimated for Chronic Pulmonary Disease group

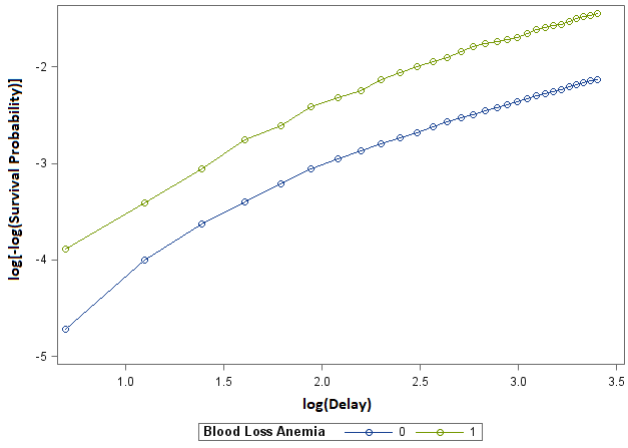

Figure C.9: Log of negative survivor log estimated for Blood Loss Anemia group

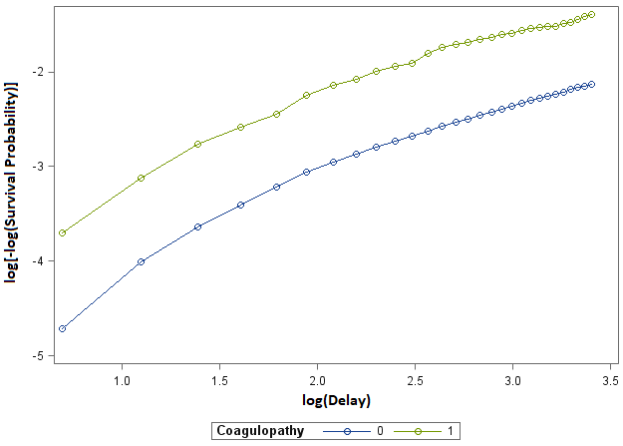

Figure C.10: Log of negative survivor log estimated for Coagulopathy group

C.4 Multivariate analysis: Interaction terms

Incorporating the interaction terms into a trained Cox model allows us to obtain hazard ratios of the crossed terms and thus differentiate the effect of one covariate conditional on another. In this appendix subsection, we present the results of the interactions terms (Table C.3) in both Model 1 and Model 2, since they allow us to note something interesting in the effects of the covariates.

	Model 1[(1)]			**Model 2**[(2)]		
	HR	**95% CI**	***p*-value**	**HR**	**95% CI**	***p*-value**
CCI × LOS						
$1 \times \leq 4$ days	1.322	1.252-1.396		-	-	
$1 \times > 4$ days	1.233	1.185-1.282	0.0426	-	-	-
$\geq 2 \times \leq 4$ days	1.975	1.868-2.088		-	-	
$\geq 2 \times > 4$ days	1.671	1.608-1.737	<0.0001	-	-	-
ELX GRP 14 × Age						
$1 \times \leq 75$ years	-	-		1.894	1.733-1.207	
$1 \times > 75$ years	-	-	-	1.417	1.304-1.539	<0.0001
ELX GRP 20 × Age						
$1 \times \leq 75$ years	-	-		1.192	1.112-1.279	
$1 \times > 75$ years	-	-	-	1.032	1.006-1.058	<0.0001

[(1)]Model 1: Age, LOS, CCI, CCI × LOS, [(2)]Model 2: Age, LOS, ELX GRP 14, ELX GRP 20, ELX GRP 19, ELX GRP 1, ELX GRP 6, ELX GRP 25, ELX GRP 3, ELX GRP 10, ELX GRP 26, ELX GRP 22, ELX GRP 14 × Age, ELX GRP 20 × Age. The labels of the comorbidity groups are given in the table C.1, page 52.

Table C.3: Cox's proportional-hazards of interaction terms

As a reminder, in the results presented in section 9.2.2, a patient with a CCI of 2 or more had a significantly higher risk of 30-day rehospitalization (76.2%). In addition, a length of stay of more than 4 days was also associated with a higher risk of 30-day readmission (68.0%). However, here we can see that the two covariates seem to behave slightly differently when crossed, as we can see that patients with a CCI of at least 2 and a length of stay greater than 4 days have a slightly lower risk (67.1% versus 97.5%) than those with a CCI greater than 2 and a short length of stay (less than or equal to 4 days).

Similar behavior was observed in the Elixhauser comorbidity groups according to patient age. Very pronounced for patients diagnosed with renal failure, who have a higher risk (89.4% versus 41.7%) of being rehospitalized if they are younger ($\leq$ 75 years). Less pronounced for patients with solid tumor without metastasis, but with a higher risk (19.2% versus 3.2%) for patients aged 75 years and younger.

C.5 Multilevel analysis: GLMMs with Elixhauser comorbidity groups

	Model 5[(1)]			Model 6[(1)]		
	OR	**95% CI**	***p*-value**	**OR**	**95% CI**	***p*-value**
Age (years)						
$\leq$ 75 years	ref	ref	ref	ref	ref	ref
$>$ 75 years	1.340	1.304-1.377	<0.0001	1.347	1.309-1.380	<0.0001
Length of stay (days)						
$\leq$ 4 days	ref	ref	ref	ref	ref	ref
$>$ 4 days	1.708	1.665-1.753	<0.0001	1.695	1.656-1.742	<0.0001
Elixhauser comorbidity groups[(2)]						
Renal Failure	1.629	1.515-1.752	<0.0001	1.587	1.473-1.711	<0.0001
Solid Tumor without Metastasis	1.590	1.505-1.680	<0.0001	1.581	1.494-1.674	<0.0001
Metastatic Cancer	1.633	1.491-1.787	<0.0001	1.615	1.471-1.773	<0.0001
Congestive Heart Failure	1.377	1.283-1.479	<0.0001	1.376	1.282-1.476	<0.0001
Hypertension Uncomplicated	1.075	1.045-1.105	<0.0001	1.067	1.038-1.097	<0.0001
Fluid and Electrolyte Disorders	1.266	1.160-1.381	<0.0001	1.252	1.150-1.363	<0.0001
Valvular Disease	1.502	1.375-1.642	<0.0001	1.512	1.386-1.650	<0.0001
Chronic Pulmonary Disease	1.134	1.069-1.204	<0.0001	1.138	1.074-1.207	<0.0001
Blood Loss Anemia	1.525	1.346-1.729	<0.0001	1.524	1.347-1.650	<0.0001
Coagulopathy	1.499	1.456-1.538	<0.0001	1.493	1.320-1.688	<0.0001
Urban/Rural status						
Urban	-	-	-	ref	ref	ref
Rural	-	-	-	1.036	1.004-1.070	0.0290
French Deprivation index						
$< P_{20}^{(3)}$	0.889	0.844-0.936	$<$ 0.0001	-	-	-
$[P_{20}; P_{40}[$	0.966	0.918-1.016	0.1830	-	-	-
$[P_{40}; P_{60}[$	ref	ref	ref	-	-	-
$[P_{60}; P_{80}[$	1.004	0.957-1.054	0.8569	-	-	-
$\geq P_{80}$	1.053	1.004-1.105	0.0337	-	-	-
Private/Public status						
Private	ref	ref	ref	ref	ref	ref
Public	1.496	1.456-1.538	<0.0001	1.489	1.449-1.529	<0.0001

[(1)] Adjustment without consideration of Renal Failure $\times$ Age and Solid Tumor without Metastasis $\times$ Age interaction terms, [(2)] The 10 most relevant groups, [(3)] P_{20}, P_{40}, P_{60} et P_{80} are the first (-1.1098), second (-0.2290), third (0.2946), fourth (0.9379) quintiles.

Table C.4: Multivariate models predicting the risk of 30-day rehospitalization by multilevel logistic regression (Elixhauser comorbidity groups version)

Appendix D

D.1 Interaction between age and logarithm of time

The 2-level categorization defined for age ($\leq$ 75 years or $>$ 75 years, see section 11.2.2) seemed to be a very good choice. We wanted to confirm the visuals with a statistical test, so we introduced an interaction between time and this form of age covariate. Following the same principle as in the previous section 4.4.2.2, we divided the variable age into two groups. The first variable we created, `Age1`, corresponds to the continuous age of patients aged 75 or under, and was set to 0 for those over 75. The second variable, `Age2`, corresponds to the continuous age of patients over 75, and was set to 0 for those aged 75 or under. If the coefficients are not equal, this would imply a change in behavior over time, and therefore a violation of the PH hypothesis.

```
PROC PHREG DATA=Patients;
    MODEL Delay*RH30(0) = Age1 Age2 Age1T Age2T;
    Age1T = Age1*log(Delay);
    Age2T = Age2*log(Delay);
    TEST Age1T = Age2T;
RUN;
```

Linear hypothesis results

	Label	**Wald Chi-2**	**Pr > Chi-2**
Age1T = Age2T	Test 1	1.4458	0.2292

Table D.1: PHREG output for interaction between age and logarithm of time

The interaction is no longer being significant, with a p-value equal to 0.2292 for the test of equality of coefficients. Thus we can consider this a solid choice.

References

[1] P.K. Andersen, O. Borgan, R.D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer New York, 1993. DOI: https://doi.org/10.1007/978-1-4612-4348-9.

[2] N.E. Breslow and D.G. Clayton. "Approximate Inference in Generalized Linear Mixed Models". In: *Journal of the American Statistical Association* 88.421 (1993), pp. 9–25. DOI: https://doi.org/10.1080/01621459.1993.10594284.

[3] K. Carey. "Measuring the hospital length of stay/readmission cost trade-off under a bundled payment mechanism." In: *Health Econ.* 24.7 (July 2015), pp. 790–802. DOI: https://doi.org/10.1002/hec.3061.

[4] K. Carey and T. Stefos. "The cost of hospital readmissions: evidence from the Veterans Administration." In: *Health Care Manag Sci.* 19.3 (Sept. 2016), pp. 241–8. DOI: https://doi.org/10.1007/s10729-014-9316-9.

[5] J. Cohen. "CHAPTER 10 - Set Correlation and Multivariate Method". In: *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)* Routledge, 1988, pp. 467–531. DOI: https://doi.org/10.4324/9780203771587.

[6] D. Commenges and H. Jacqmin-Gadda. *Modèles biostatistiques pour l'épidémiologie*. de Boeck, 2015. ISBN: 9782807300262. URL: https://inria.hal.science/hal-01580144.

[7] D.R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (Jan. 1972), pp. 187–220. DOI: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

[8] D.R. Cox. "Partial likelihood". In: *Biometrika* 62.2 (Aug. 1975), pp. 269–276. DOI: https://doi.org/10.1093/biomet/62.2.269.

[9] T. Deborde, E. Chatignoux, C. Quintin, N. Beltzer, FF. Hamers, and A. Rogel. "Breast cancer screening programme participation and socioeconomic deprivation in France." In: *Prev Med.* 115 (Oct. 2018), pp. 53–60. DOI: https://doi.org/10.1016/j.ypmed.2018.08.006.

[10] O.V. Demler, N.P. Paynter, and N.R. Cook. "Tests of calibration and goodness-of-fit in the survival setting." In: *Stat Med.* 34.10 (May 2015), pp. 1659–80. DOI: https://doi.org/10.1002/sim.6428.

[11] J.K. Grønnesby and O. Borgan. "A method for checking regression models in survival analysis based on the risk score." In: *Lifetime Data Anal.* 2.4 (Dec. 1996), pp. 315–28. DOI: https://doi.org/10.1007/BF00127305.

[12] J.F. Hemphill. "Interpreting the magnitudes of correlation coefficients." In: *Am Psychol.* 58.1 (Jan. 2003), pp. 78–9. DOI: https://doi.org/10.1037/0003-066x.58.1.78.

[13] D.Y. Lin. "On the Breslow estimator". In: *Lifetime data analysis* 13 (Jan. 2008), pp. 471–80. DOI: https://doi.org/10.1007/s10985-007-9048-y.

[14] D.Y. Lin, L.J. Wei, and Z. Ying. "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals". In: *Biometrika* 80.3 (Sept. 1993), pp. 557–572. DOI: https://doi.org/10.2307/2337177.

[15] H. Marquaille, G. Clément, X. Lenne, FR. Pruvot, S. Truant, D. Theis, and M. El Amrani. "Predictive factors for utilization of a low-volume center in pancreatic surgery: A nationwide study." In: *J Visc Surg.* 158.2 (Apr. 2021), pp. 125–132. DOI: https://doi.org/10.1016/j.jviscsurg.2020.06.004.

[16] S. May and D.W. Hosmer. "A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model." In: *Lifetime Data Anal.* 4.2 (June 1998), pp. 109–20. DOI: https://doi.org/10.1023/a:1009612305785.

[17] S. May and D.W. Hosmer. "A cautionary note on the use of the Grønnesby and Borgan goodness-of-fit test for the Cox proportional hazards model." In: *Lifetime Data Anal.* 10.3 (Sept. 2004), pp. 283–91. DOI: https://doi.org/10.1023/b:lida.0000036393.29224.1d.

[18] C.E. McCulloch. "CHAPTER 8 - Generalized Linear Mixed Models (GLMMs)". In: *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. 2000, pp. 220–246. ISBN: 9780471193647. DOI: https://doi.org/10.1002/0471722073.

[19] H.B. Mehta, SD. Sura, D. Adhikari, CR. Andersen, SB. Williams, AJ. Senagore, YF. Kuo, and Goodwin JS. "Adapting the Elixhauser comorbidity index for cancer patients." In: *Cancer.* 124.9 (May 2018), pp. 2018–2025. DOI: https://doi.org/10.1002/cncr.31269.

[20] M. Moschini, G. Gandaglia, N. Fossati, P. Dell'Oglio, V. Cucchiara, S. Luzzago, E. Zaffuto, N. Suardi, R. Damiano, SF. Shariat, and F. Montorsi. "Incidence and Predictors of 30-Day Readmission After Robot-Assisted Radical Prostatectomy." In: *Clin Genitourin Cancer.* 15.1 (Feb. 2017), pp. 67–71. DOI: https://doi.org/10.1016/j.clgc.2016.06.002.

[21] F. Palmisano, L. Boeri, M. Fontana, A. Gallioli, E. De Lorenzis, S. Zanetti, G. Sampogna, M. Spinelli, G. Albo, F. Longo, F. Gadda, P. Dell'Orto, and E. Montanari. "Incidence and predictors of readmission within 30 days of transurethral resection of the prostate: A single center European experience". In: *Scientific Reports* 8 (Apr. 2018). DOI: https://doi.org/10.1038/s41598-018-25069-5.

[22] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, JC. Luthi, L. Saunders, C. Beck, T. Feasby, and W. Ghali. "Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data". In: *Medical care* 43 (Dec. 2005), pp. 1130–9. DOI: https://doi.org/10.1097/01.mlr.0000182534.19832.83.

[23] A. Salmivalli, O. Ettala, P.J. Boström, and V. Kytö. "Mortality after surgery for benign prostate hyperplasia: a nationwide cohort study." In: *World J Urol.* 40.7 (July 2022), pp. 1785–1791. DOI: https://doi.org/10.1007/s00345-022-03999-0.

[24] MT. Sharabiani, P. Aylin, and A. Bottle. "Systematic review of comorbidity indices for administrative data." In: *Med Care.* 50.12 (Dec. 2012), pp. 1109–18. DOI: https://doi.org/10.1097/MLR.0b013e31825f64d0.