Master 2 MIGS (*Mathématiques pour l'Ingénierie, alGorithmique, Statistique*)

Work-Study Program Report

SAUCE Justine

Université de Bourgogne *CHU de Dijon*

Apprenticeship Master : Jonathan COTTENET
Pedagogical Supervisor : Brice NAFETAT

Oral presentation on March the 29th, 2023

Contents

Introduction 1

I CHU Organization and Database 2

1 Research Department at the CHU 2

1.1 Département d'Information Médicale (DIM) 2

1.2 *Service de Biostatistique et d'Information Médicale* (SBIM) 3

2 Medico-administrative PMSI Database 4

2.1 Local perspective 4

2.2 National perspective 4

2.3 Anonymous Linkage 6

II Framework for the Use of PMSI on SAS 7

1 Import and Management of Data on SAS 7

1.1 SAS Tables from PMSI 7

1.1.1 Error Filtering 7

1.1.2 Number of stays and patients 8

1.2 Research on Diagnoses 9

1.3 Analysis of Stay Trajectories 10

2 Statistical Analysis with SAS 12

2.1 Statistical Tests Procedures 12

2.1.1 Tests on Qualitative Variables 12

2.1.2 Tests on Quantitative Variables 13

2.2 The LOGISTIC Procedure 15

2.3 The LIFETEST & PHREG Procedures 15

III Introduction to the Research Project 17

1 Research Objectives 17

1.1 Description of the Overall Context 17

1.2 Methodology Overview 17

Conclusion 18

Appendix 19

Acronyms

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

CCAM *Classification Commune des Actes Médicaux*

CHU *Centre Hospitalier Universitaire*

CIM-10 / ICD-10 *Classification statistique internationale des maladies et des problèmes de santé connexes 10ème révision* / International Classification of Diseases 10th revision

CNIL *Commission Nationale de l'Informatique et des Libertés*

DAS *Diagnostic Associé Significatif*

DP *Diagnostic Principal*

DR *Diagnostic Relié*

DIM *Département d'Information Médicale*

FINESS *Fichier National des Etablissements de Santé Sanitaires et Sociaux.*

GHM *Groupe Homogène de Malades*

HAD *Hospitalisation à Domicile*

MCO *Médecine, Chirurgie, Obstétrique et Odontologie*

NAS *Numéro Administratif de Séjour*

PMSI *Programme de Médicalisation des Systèmes d'Information*

RSA *Résumé de Sortie Anonymisée*

RSS *Résumé de Sortie Standardisé*

RUM *Résumé d'Unité Médicale*

SBIM *Service de Biostatistique et d'Information Médicale*

SNDS *Système National des Données de Santé*

SSR *Soins de Suite ou de Réadaptation*

T2A *Tarification à l'Activité*

UM *Unité Médicale*

Introduction

The work-study program as part of the Master 2 MIGS takes place at the *Université de Bourgogne* and at the *Centre Hospitalier Universitaire* (CHU) of Dijon. The core of this program is to integrate a biostatistical and epidemiological research unit to become familiar with statistical research tools in the medical sector. The apprenticeship will take place in the research unit of the CHU of Dijon alongside statisticians of the *Service de Biostatistique et d'Information Médicale* (SBIM).

The first step was to become familiar with the databases used. In particular, it was necessary to understand how *the Programme de Médicalisation des Systèmes d'Information* (PMSI) works. In a second step, time was devoted to mastering the table management procedures and the statistical tools offered by the SAS command language. Finally, the beginning of a research project will take place once the tools have been mastered.

The objective of this work-study report is to retrace the work done and the skills acquired during this first phase of work-study at the CHU of Dijon. We will start by presenting the statistical units of the CHU and their missions. Then we will detail the essential points of the PMSI for the understanding of the studies carried out on these data. Then, we will try to explain the variables that are essential to studies based on PMSI data, as well as their management on SAS. A part of the report will also be devoted to SAS statistical procedures discovered during these first months of professional experience. Finally, an overview of the research project that will be conducted in the second phase of apprenticeship will be given.

Part I.
CHU Organization and Database

Chapter I.1
Research Department at the CHU

Since 2008, the budget of each hospital depends on the medical activity described in a specific computer program that compiles discharge abstracts of all admissions. This program is called *Programme de Médicalisation des Systèmes d'Information* (PMSI). It collects data on all hospitalizations in France (in all public and private hospitals) in order to better manage the financing of health care institutions and to organize the supply of care.

The global organization for collecting and processing PMSI data in the field of *"Médecine, Chirurgie, Obstétrique et Odontologie"* (MCO) is managed by the *Agence Technique de l'Information sur l'Hospitalisation* (ATIH). Prior to the release of (anonymized) medico-administrative data by the ATIH, the data are collected and produced within each public or private hospital, thanks to the *Département d'Information Médicale* (DIM), dedicated to managing the collection of these data. All these data are then transmitted to the ATIH to build the national PMSI database.

I.1.1 *Département d'Information Médicale* (DIM)

The mission of the DIM (directed by Pr Quantin) is to ensure the coherence between the strategic orientations and the implementation of the medico-economic measures of the CHU Dijon Bourgogne through the exhaustiveness and the quality of the medico-administrative data of the PMSI. It handles the exploitation of this data to optimize the CHU's receipts under the *Tarification à l'Activité* (T2A)[1].

Missions of a DIM Engineer:

- Management of the completeness and quality control of PMSI data in the *Médecine, Chirurgie, Obstétrique et Odontologie* (MCO), *Soins de Suite ou de Réadaptation* (SSR) and *Hospitalisation à Domicile* (HAD) sectors.

- Management of nomenclatures for describing medical activity: diagnoses (International Classification of Diseases 10th revision (ICD-10)) and acts ((CCAM) *Classification Commune des Actes Médicaux*.

[1] The Tarification à l'Activité (T2A) is the unique method of financing private and public health care institutions since 2008. It guarantees that pricing depends on the nature and volume of activities and is no longer based on an expenditure authorization.

I.1.2 *Service de Biostatistique et d'Information Médicale* (SBIM)

The DIM is completed by a research unit, the *Service de Biostatistique et d'Information Médicale* (SBIM, directed by Pr Quantin) whose objective is to analyze the CHU's healthcare offer in relation to its environment, and to ensure clinical or epidemiological research based on SNDS[2] and/or PMSI data. In addition, it participates in various medico-economic studies, especially by comparison of revenues and costs of stays.

All these missions converge towards better patient care: studies of re-hospitalization rates, reduction of mortality, and reduction of time spent in hospital. The objective is to promote the evaluation of care.

Missions of a SBIM Statistician:

- Conducting epidemiological studies[3] at the request of clinicians, using PMSI or SNDS data.
- Medico-economic studies (on diagnostic and therapeutic medical strategies, based on SNDS and/or PMSI data).
- Elaboration of reports and international articles.

[2] The SNDS integrates data from the *Assurance Maladie* (consultations and medication), hospital data (PMSI database) and medical causes of death (Inserm's CépiDC database).

[3] The use of these data is subject to the authorization of the CNIL and possibly to the permission of CESREES in certain cases.

Chapter I.2
Medico-administrative PMSI database

Inspired by the American model of diagnosis-related groups (DRG), the collection of national health administrative data was introduced in France in 1991 and extended to all French health establishments in 1997. This gave rise to the PMSI, a collection of information on health care activity and its billing, but above all, a medico-administrative database covering more than 99% of the population (65 million people).
The core of the PMSI is the systematic collection of administrative data (relating to the patient and the stay) and medical data (diagnoses and procedures, etc.) coded according to imposed classifications.

It includes 4 "fields":

- *"Médecine, Chirurgie, Obstétrique et Odontologie"* (MCO)
- *"Soins de Suite ou de Réadaptation"* (SSR)
- *"Psychiatrie"* in the form of the RIM-Psy (collection of medical information in psychiatry)
- *"Hospitalisation à Domicile"* (HAD)

I.2.1 Local Perspective

We place ourselves in the framework of the PMSI MCO, in this context, the activity is registered in the form of a *Résumé de Sortie Standardisé* (RSS) produced for each hospitalization (stays and sessions). The information contained in the RSS is a standardized and coded abstract of the contents of the patient's medical record. Each stay in a health care institution (clinics, hospitals, public or private) gives rise to the production of an RSS. The RSS number is assigned under the control of a responsible physician, in a unique way. The RSS must reflect the patient's stay as accurately as possible.

The RSS is made up of the RUMs relating to the same stay of a patient in the MCO field, i.e., it includes as many RUMs as there are *Unités Médicales*[4] (UMs) frequented during the patient's stay. For a multi-unit stay, several RUMs are created and grouped under a single RSS number. When the RUMs correspond to distinct stays, separate RSS are assigned.

I.2.2 National Perspective

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). It allows following the hospitalizations of the same patient, regardless of the location (public or private sector). The anonymous linkage is based on the creation of a unique anonymous number for each patient, using a software program that uses three variables: social insurance number, date of birth and gender. 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.

[4] Individualized set of resources providing care to the patient (identified by a specific code).

An RSA is then produced, using the anonymous patient number and the RSS from the grouped RUM-RSS. It will contain slightly more restrictive information. In the RSS one could find information such as a *Numéro Administratif de Séjour* (NAS) or the RSS number, which are identifiers specific to the patient's stay. This kind of information will be removed (or replaced) to ensure anonymity.

The censored information is as follows:

- RSS & NAS (removed)
- Birth Date: replaced by the age calculated on the date of entry (in days for children less than one year old on that date)
- *Numéro d'Unité Médicale* : only the number of RUMs making up the original RSS is given.
- Post code: replaced by a geographical code allocated according to a list agreed at national level.
- Dates of entry and discharge: replaced by the length of stay, the month, and the year of discharge.
- Date of the last menstrual period and date of performance of the procedures: replaced by the delay in days in relation to the date of entry.

Each RSA is classified, using a classification algorithm, in one and only one *Groupe Homogène de Malades* (GHM). The classification of all stays in an institution into GHMs determines the rate of reimbursement by the health insurance scheme, since stays classified in the same group have, by construction, similar resource consumption. The classification is also medical, because its first level of classification is based on medical criteria (functional device or notorious reason for hospitalization)

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 have been related to the principal diagnosis), and *Diagnostic Associé Significatif* (DAS) (all complications and morbidities 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).

Information found in an RSA can be constant, whatever the stay this information will be present in the RSA; or variable, it will not necessarily be present in all patient's stays. Below is a summary of information contained in an RSA.

Information contained in the RSA:

- *Fichier National des Etablissements de Santé Sanitaires et Sociaux* number (FINESS)[5]
- *Numéro d'Index*[6]
- Gender
- Admission date (2019)
- *Groupe Homogène de Malades* (GHM)
- Mode of admission & Mode of discharge
- Provenance & Destination
- *Diagnostic Principal* (DP), *Diagnostic Relié* (DR) & *Diagnostic Associé Significatif* (DAS)
- Medical procedures

[5] Legal entity (public) or Geographical entity (private).

[6] Identifiers of stays in one institution.

I.2.3 Anonymous Linkage

The purpose of linkage is to be able to follow a patient through his hospitalizations. Within our context, this means that we want to be able to identify the same patient in a strictly anonymous way, and to find all his stays in an MCO department of a hospital (private or public). This link will allow us, for example, to study re-hospitalization rates according to certain pathologies, etc.

Anonymous linkage is based on the creation of an anonymous number named *Numéro Anonyme*, for each patient, through a software program that uses three variables: *Numéro de Sécurité Sociale*, date of birth and gender. The anonymous number is characteristic of an individual because the same anonymous number is obtained from the same identification variables (reproducibility).

The software used is provided by the ATIH and is called *Module d'Anonymisation et de Gestion des Informations de Chaine* (MAGIC). It uses a *Fonction d'Occultation des Informations Nominatives*[7] (FOIN) created by Pr Catherine Quantin, used by the *Caisse Nationale d'Assurance Maladie des Travailleurs Salariés*, validated by the *Commission Nationale de l'Informatique et des Libertés* (CNIL).

The anonymous number created is linked to the NAS and a file called ANO-MCO is created. The anonymous number is thus inserted into an ANO-MCO file that contains neither medical nor billing data. This file will contain the Index number and the FINESS number that will allow it to be linked with the RSA.

In order to verify the consistency of the anonymous number created, 7 *Codes retour* are created, indicating anomalies detected in the information at the origin of the linkage key, as well as possible errors when joining the linkage information with the medical summaries. The detailed list of different return codes is provided in Appendix A. The *Codes retour* are set to 0 when the information is correct and set to blank in case of errors/inconsistencies.

Content of the linkage file ANO-MCO:

- FINESS Number
- *Numéro d'Index*
- Month of discharge
- Year of discharge
- *Codes retour*
- *Numéro Anonyme*
- *Numéro de Séjour*

The linkage key and the stay number are linked to the PMSI summaries using the FINESS number and PMSI sequential number information, by year and PMSI field.

There are then two ways of identifying the stays of each patient. The first is to refer to FINESS number and *Numéro d'Index*. The second is to use *Numéro Anonyme* and *Numéro de Séjour*.

[7] To hide nominative information.

Part II
Framework for the Use of PMSI on SAS

Chapter II.1
Import and Management of Data on SAS

The command language used at SBIM is SAS software. It allows a very good management of the databases while giving us the hand on many tools of statistical analysis. This section focuses on the process of managing SAS tables and on describing the most commonly used variables in a PMSI research project and their employment in SAS. Examples of short and simple SAS programs will be given to illustrate the explanations.

II.1.1 SAS Tables from PMSI

A familiarization work with SAS & PMSI database consisted in importing a text file containing the PMSI data of all the patient stays of one year in France. Each variable respects a precise format and a precise position in the text file.

The most interesting part is the import of the variable part of the information, in particular the import of DAS variables and associated medical procedures, which are widely used in research. An example of the code realized, as well as a part of the file explaining the format of each variable for illustration, will be left in the Appendix A.

To work on PMSI data for a research project, there are several SAS tables already built on each year. These tables are generally called "RSA", "ANO", "DAS" and "*Actes*" followed by the corresponding year. The "RSA" table is the table that contains most of the information about the stay. "ANO" is the table for the file ANO-MCO described in section (I.2.3 Anonymous Linkage). The "DAS" and "*Actes*" tables contain the information on the DAS and the procedures associated with each stay.

II.1.1.1 Error Filtering

The first thing to do when working on the PMSI SAS tables is to filter out the anonymous stays and numbers in error.

As seen in section (2.2), when the *Numéro Anonyme* is created, return codes are created which validate or not the coding of the *Numéro Anonyme*. For this number to be valid, all the return codes must be set to 0. In SAS, an exclusion filtering these return codes will be systematically performed to maintain consistency.

In addition, a GHM is dedicated to errors and other unclassifiable stays; all patients in GHMs beginning with 90 will be excluded.

SAS Example:

```
DATA RSA2013;

      SET BN.RSA2013;

      /* Exclusion of GHMs beginning with '90' */
      if substr(GHM,1,2) = '90' then delete;

      /* Only return codes at 0 are kept */
      if retnoss='0' and retdnaiss='0' and retsexe='0' and  retnoadm='0'
         and retfusionah='0'  and retfusionap='0' and retdateref='0'
         and retdnaissvid='0' and retsexevid='0'
      then output;

RUN;
```

The table named 'BN.RSA2013' is the original "RSA" table of all 2013 patient stays, merged with the 2013 "ANO" table containing return codes and *Numéro Anonyme*. The patients and stays detected with errors are excluded from this table which gives the table RSA2013.

II.1.1.2 Number of stays and patients

The number of stays corresponds to the total number of observations in the table. Indeed, the RSA2013 table is constructed in such a way that one row is equivalent to one patient stay. A patient appears as many times in the table as he/she has been admitted to a healthcare institution.

The number of patients included will be the number of distinct *Numéro Anonyme* keys generated associated with the RSAs after exclusion of the detected errors (presented above). One can simply count them with an SQL procedure or sort the table with the `nodupkey` option on the *Numéro Anonyme* and look at its number of observation (number of rows).

SAS Example:

```
PROC SORT DATA=RSA2013
          OUT=Table_of_Patients nodupkey;

      /* NoAno : variable Numéro Anonyme */

      /* Only distinct NoAno's are kept */
      BY NoAno;

RUN;
```

SAS Journal:

```
NOTE: There were 23872551 observations read from the data set WORK.RSA2013.
NOTE: 12152742 observations with duplicate key values were deleted.
NOTE: The data set WORK.TABLE_OF_PATIENTS has 11719809 observations and 61 variables.
```

In this example, we have 11,719,809 patients admitted to MCO over 2013 and 23,872,551 stays.

II.1.2 Research on Diagnoses

For a research project, it is common to import a table containing patients with specific characteristics, and it is essential to be able to filter, especially on the diagnosis of each patient.

One of the specific concerns of the PMSI tables is not to lose any DAS among the patient stays. As explained previously, the DAS represent a part of the variable information of a stay, there may be no DAS associated with the stay as there may be several.

The goal is to retrieve individual information for each patient whose diagnosis (PD, DR, or DAS) is the one sought for the study. This is an essential point in understanding the process of including patients in a study and requires special attention. We cannot simply link the RSA and DAS tables and then carry out the search, because for each stay with more than one DAS, we would only keep the first one.

One way to handle it with SAS is to use the `RETAIN` option and to create indicators of the presence of the pathologies sought.

The search performed below is for patients diagnosed with Heart Failure in 2013. The ICD-10 code for this diagnosis is 'I50' ("Heart failure"). We proceed by creating an indicator whose purpose is to indicate the presence of the pathology and we will then keep only the patients with the pathology.

SAS Example:

```
DATA Inclusions_DAS_2013;

	/* NoIndex : variable Numéro d'Index */

	/* We work on DAS table, which is already sorted by FINESS NoIndex*/
	SET BN.DAS2013;
	BY FINESS NoIndex;

	/* We define a numeric indicator */
	ATTRIB Indicateur_DAS_IC length = 3.;

	/* We use the retain option to keep the value when it switches */
	RETAIN Indicateur_DAS_IC;

	/* We initialize it to 0 */
	if first.NoIndex then Indicateur_DAS_IC =0;

	/* We update the value when we find the pathology */
	if substr(Code,1,3) = 'I50' then Indicateur_DAS_IC=1;

	/* We keep only the last occurrence of the associated stay */
	if last.NoIndex and Indicateur_DAS_IC=1 then output;

RUN;
```

Since more than one DAS can be registered for the same stay, the DAS table contains several rows for the same stay. It is therefore necessary to keep only one stay per patient, so that it can be merged more easily with the RSA to obtain the information associated with the stay and the patient.

Following is the merge with a filtration on the Diagnostic Principal (DP) and Diagnostic Relié (DR).

SAS Example:

```
DATA Inclusions_2013 (drop = Indicateur_DAS_IC /* No longer useful to us */);
       /* We merge on FINESS NoIndex, tables are already sorted */
       MERGE RSA2013 Inclusions_DAS_2013;

       BY FINESS NoIndex;

       /* We keep the patients with the pathology in DP or DR or DAS */
       if substr(DP,1,3)  = 'I50' or substr(DR,1,3)  = 'I50' or
          Indicateur_DAS_IC=1

       then output;
RUN;
```

The Inclusions_2013 table lists all stays where patients were diagnosed with heart failure.

II.1.3 Analysis of Stay Trajectories

After the transition from RSS to RSA, we lose direct information on the dates of entry and discharge as seen in section 1.2. To situate the stays in time, to identify their duration and to be able to study the delays before a possible re-hospitalization, we have the variable *Numéro de Séjour* and *Durée de Séjour*.

The variable *Numéro de Séjour* in the ANO-MCO file indicates the beginning of the anonymous discharge abstract (RSA). The beginning of a stay in an MCO unit is therefore identified via the variable *Numéro de Séjour*.

The variable *Durée de Séjour*, present in the RSAs, corresponds to the number of nights spent in the hospital (the length of the stay). It is calculated automatically during anonymization, as the difference between discharge date and entry date.

The end of the stay is identified by adding the *Durée de Séjour* to the *Numéro de Séjour*. Once the start and end of a stay have been correctly identified, we can then look at the time between hospitalizations. The interval between the end of a first stay and the beginning of another one corresponds to this delay and is counted in days.

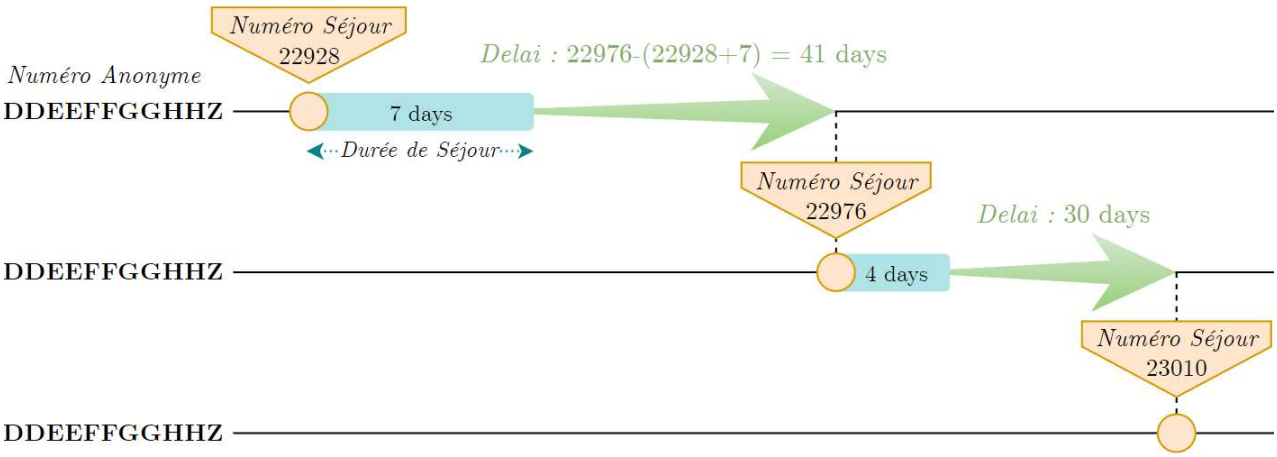

Figure 1 : Rehospitalization scheme

In research, it could be used to study re-hospitalizations for example. We often find studies on re-hospitalization within 30 days, with a delay value between 1 and 30 days.

One way to do this is to sort the stays chronologically and calculate the difference between each. In general, the first hospitalization of each patient is identified and a delay with the first re-hospitalization is calculated.

SAS Example:

```
/* Sorting procedures */

	PROC SORT DATA=Inclusions_2013;
		/* We sort by Numéro Anonyme and Numéro de Séjour */
		BY NoAno NoSejour;
	RUN;

	PROC SORT DATA=Inclusions_2013 OUT=Patients_2013
		/* We rename the variable to distinguish it from other hopsitalizations & we
		keep only the first stay per patients. */
		(rename=(NoSejour=NoSejourIndex DureeSejour=DureeSejIndex)) nodupkey;
		BY NoAno;
	RUN;

	PROC SORT DATA=RSA2013;
		/* We sort by Numéro Anonyme and Numéro de Séjour */
		BY NoAno NoSejour;
	RUN;

/* Merge of the 2 tables (Patients & All patient stays on the same year */

	DATA Rehospit_2013;
		MERGE Patients_2013(in=a) RSA2013;
		BY NoAno;
		/* We keep only the stays following the first hospitalization for the patients
		included*/
		if a and NoSejour > NoSejourIndex then output;
	RUN;

/* Creating the variable Délai */

	DATA Rehospit_2013;
		SET Rehospit_2013;

		ATTRIB Delai length = 3.;

		/* (NoSejourIndex + DureeSejIndex) corresponds to the end of the first stay
		    NoSejour corresponds to the beginning of the following stays */
		Delai = NoSejour - (NoSejourIndex + DureeSejIndex);

		/* If the stay took place before the reference stay it is deleted */
		if Delai < 0 then delete;

	RUN;
```

The "Rehospit_2013" table contains all the hospitalizations that took place after the index hospitalization[8], as well as the delay between the end of the reference stay and the beginning of the stay concerned. We can then construct an indicator that will reflect the presence or absence of a rehospitalization within 30 days. In some studies, this variable will be the variable to be explained.

[8] It corresponds to the first hospitalization where the pathology was diagnosed.

Chapter II.2
Statistical Analysis with SAS

In epidemiology, when using medical data, it is often necessary to describe an event or phenomenon that is itself influenced by the occurrence of other events or phenomena called exposure factors. We seek to construct a statistical analysis model that highlights an association between a qualitative variable (called the outcome or response variable or variable to be explained or dependent variable) and variables that may be qualitative or quantitative (called explanatory or independent variables). The outcome is the occurrence or non-occurrence of the event studied - pathology, re-hospitalization, etc. - and the explanatory variables are factors likely to influence the occurrence of the event.

There are several multivariate analysis models commonly used for this purpose: logistic regression, Poisson regression, Cox model, etc.

SAS is a popular language for its statistical analysis procedures. The purpose of this section is to give an overview of the one used in the apprenticeship program and of how it works.

II.2.1 Statistical Tests Procedures

The construction of an analysis model involves several steps. Firstly, it is necessary to study each of our variables and the possible existence of a linear relationship between them.

It is also necessary to analyze the links between each of the explanatory variables and the outcome: we will introduce into our model the explanatory variables for which the association with the variable to be explained is sufficiently strong without being too strict.

Within the framework of work-study, we usually study a categorical response variable, and we commonly use indicators of pathology presence. We have some additional quantitative information such as age or the length of stay. We are therefore looking to perform tests between two qualitative variables or between one qualitative variable and one quantitative variable[9].

For this purpose, SAS provides all the classical test procedures such as the Chi-square ($\chi 2$) test, Student's t-test, etc. We will see their syntax in this section.

II.2.1.1 Tests on Qualitative Variables

An example of a comparison between variables is to detect whether certain factors may be significantly associated with a disease. These factors can be qualitative variables such as sex, or the presence of other pathologies coded by 0 or 1. In order to compare those kinds of variables, we usually use a Chi-square (χ^2) test, or, in the case of a small sample size, a Fisher test.

[9] In the case of two quantitative variables, we could use the Pearson correlation coefficient by using the CORR procedure.

i) Chi-Square (χ^2) Test

To run the Chi-square (χ^2) test in SAS we use the `PROC FREQ` with `chisq` option, specified in the `TABLES` statement. `PROC FREQ` computes several chi-squared tests for each two-dimensional table.

SAS Command :

```
PROC FREQ DATA=Table;
      TABLES v1*v2 / chisq;
RUN;
```

The output statistic called Chi-square corresponds to the Pearson Chi-square test. This is the one we will rely on.

ii) Fisher's Exact Test

When we work with small samples, the χ^2 test won't be adapted, in this case, the SAS output mentions it and a Fisher's exact test will be more accurate than the chi-square test. The Fisher's exact test does not depend on any distributional assumptions and is therefore appropriate even for small samples. We can run a Fisher's exact test by using `fisher` option in the `FREQ` procedure.

SAS Command:

```
PROC FREQ DATA=Table;
      TABLES v1*v2 / fisher;
RUN;
```

Note: For the Chi-Square and Fisher's Exact test, the null hypothesis is rejected in case of small p-values, which confirms the hypothesis of a link between the variables.

II.2.1.2 Tests on Quantitative Variables

When comparing a quantitative variable with a qualitative variable, many tests exist depending on the sample and the number of modalities of the qualitative variable.

For parametric tests we'll work with the Student's t-test or the analysis of variance (ANOVA), it depends on the qualitative variable. If the qualitative variable has 2 modalities ('YES/NO', or '0/1' for example) we will use a basic Student's t test, otherwise if it has more than 2 modalities we will have to proceed to an analysis of variance (ANOVA).

In the case where parametric tests cannot be used, we'll work with the non-parametric Wilcoxon/Mann-Whitney test for a qualitative variable with 2 modalities and Kruskal-Wallis for more than 2 modalities.

i) The TTEST Procedure

To run a Student's t-Test on independent variable, we use the TTEST procedure, the CLASS statement lists the categorical variables and VAR statement the numerical variables.

SAS Command :

```
PROC TTEST DATA=Table;

      CLASS CatVar; /* Categorical variable */
      VAR NumVar; /* Numeric variable */

RUN;
```

There are different formulas for the test statistic and degrees of freedom, based on whether we assume that the two groups have equal variances. The TTEST procedure performs both Student's t-tests in the case of equality of variances.

In the situation where the p-value is below our chosen significance level, we reject the null hypothesis, and conclude that the variables being compared are significantly differently distributed.

Note: The Student's *t*-Test can only compare the means for two (and only two) groups. It cannot make comparisons among more than two groups. If you wish to compare the means across more than two groups, you will likely want to run an ANOVA.

i) The NPAR1WAY Procedure

The NPAR1WAY procedure computes a one-way ANOVA test fort each score that can be specified. Among the tests available in this procedure are the anova option which requires the standard analysis of variance and the wilcoxon option which requires the Wilcoxon/Mann-Whitney or Kruskal-Wallis analysis.

SAS Command: (ANOVA)

```
PROC NPAR1WAY DATA=Table anova;

   CLASS CatVar;
   VAR NumVar;

RUN;
```

SAS Command: (Wilcoxon)

```
PROC NPAR1WAY DATA=Table wilcoxon /* plots=(wilcoxonboxplot) */;

   CLASS CatVar;
   VAR NumVar;

RUN;
```

II.2.2 The LOGISTIC Procedure

A logistic regression model estimates the strength of association between a qualitative outcome and explanatory variables that may be qualitative or quantitative. Once the explanatory variables are chosen (either by univariate analysis or via the literature), we can perform a logistic regression by using the LOGISTIC procedure on SAS.

SAS Command:

```
PROC LOGISTIC DATA=Table;

        CLASS CatVar1 (ref='0') / param=ref;
        MODEL ResponseVar (event='1') = NumVar1 CatVar1;

RUN;
```

In this example, we construct a logistic model to explain the variable `ResponseVar` using a numeric variable, e.g., age, and a qualitative variable, e.g., gender, coded 0 or 1.

We added `(ref='0') / param=ref` to the class statement. This tells SAS that for the variable `CatVar1`, the desired reference class is 0 (we could also use class 1 as the reference class), and then tells SAS that we want to use the reference coding scheme in the parameter estimates.

Once the model has been built, there are several possible strategies for arriving at a final model that must carry the maximum amount of information while having a limited number of variables to facilitate interpretation. This selection can be made while building the model, by specifying it in the MODEL statement. To do this, we add the option `selection=backward` (or `forward`).

II.2.3 The LIFETEST & PHREG Procedures

In epidemiology, when studying rehospitalization for example, it can be interesting to consider the time when the event occurs and not only an indicator of rehospitalization. Let's remember the construction of our indicator, it is based on the variable *Delai*, which represents the time elapsed between the end of the first stay and the beginning of the second. This variable can be used in a survival analysis.

Survival analysis is a type of statistical method used to study the occurrence and timing of events. It models the factors that influence the time of occurrence of an event. This section presents the procedures and describes the coding needed in SAS to model survival data by two methods.

i) The LIFETEST Procedure

The first step in survival analysis is often to examine overall survival using nonparametric methods such as Kaplan-Meier. The SAS command below shows how to obtain a table and a graph of the Kaplan-Meier estimator of the survival function from PROC LIFETEST.

SAS Command:

```
ODS GRAPHICS ON;

PROC LIFETEST DATA=Table
              /* plots=survival(cb) */;

       TIME Time*Event(0);
       STRATA CatVar;

RUN;

ODS GRAPHICS OFF;
```

The `TIME` statement requires at least the event time variable. Since many observations in our data are right-censored, we must also specify a censoring variable, such as an indicator, and the numeric code that identifies a censored observation, which is done with `Event(0)`. All numbers in parentheses are treated as censoring indicators, which implies that all excluded numbers between brackets are treated as indicators that the event has occurred.

When a categorical variable is specified in the `STRATA` statement, SAS produces graphs of the survival function stratified by the grouping

By default, the SAS output will produce the table of Kaplan-Meier estimates of the survival function and the graph of Kaplan-Meier estimates. This will allow to check the conditions of application of the Cox model including the proportional risk assumption.

ii) The PHREG procedure

Nonparametric methods do not model the hazard rate or estimate the magnitude of the effects of covariates; thus regression models are often used in survival data analysis. While with nonparametric methods we usually study the survival function, with regression methods we examine the hazard function. The hazard function for a given time interval gives the probability that the event will occur in that interval, given that the event has not occurred up to that point.

A very popular regression technique for survival analysis is Cox proportional hazards regression, which is used to relate multiple risk factors, considered simultaneously, to survival time.

With the `PROC PHREG` whose command is given below, we build a simple Cox model, where we determine the effects of a categorical predictor, `CatVar` (e.g. sexe), and a quantitative predictor, `NumVar` (e.g. age), on the hazard rate.

SAS Command:

```
PROC PHREG DATA=Table;

       CLASS Event (ref = '0') CatVar (ref = '0') / param=ref;
       MODEL Time*Event(0)= NumVar CatVar;

RUN;
```

Part III
Introduction to the research project

Chapter III.1
Research Objectives

After a period of familiarization with the PMSI database, the epidemiological framework and SAS procedures, a research proposal within the framework of the alternation was initiated. This research project will focus on the study of factors associated with rehospitalization of patients following a first hospitalization for prostate surgery based on PMSI data.

Research Project: Study of factors associated with rehospitalization of patients following a first hospitalization for prostate surgery using PMSI data.

III.1.1 Description of the overall context

Studies on the risk of rehospitalization are common at an international level, but the reasons for rehospitalization remain poorly understood. Many studies in the United States have shown that rehospitalizations have a deleterious effect on patient well-being and result in a significant increase 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 rehospitalizations that could be prevented by better coordination between hospitals and primary care.

The purpose of this project is to try to clarify these murky areas by studying the impact of factors that may be associated with rehospitalisation, to better understand the situation and better identify the determinants, in the context of prostate surgery.

III.1.2 Methodology Overview

The research project will evaluate factors associated with rehospitalization for each of the selected conditions for prostate surgery. We are interested in rehospitalizations of patients initially admitted for prostate surgery and diagnosed with one of the diseases, whether they are present in DP, DR, or DAS. For patients older than 18 years of age hospitalized for one of the conditions Index hospitalizations will be defined as the initial hospitalization in 2012, 2013, or 2014 for the condition under consideration, after excluding patients already hospitalized for the condition under consideration the previous year.

We will study the frequency of their rehospitalizations between discharge from the index hospitalization and readmission within 30 days. We will use a multilevel logistic regression to study the factors associated with rehospitalization.

The individual factors will be obtained from the PMSI summaries; diagnoses and procedures of the index hospitalization in order to construct the Charlson Comorbidities Index[10] to adjust for comorbidities or to collect elements on the severity of the pathology, as well as age and length of stay.

[10] Charlson Comorbidities Index assesses comorbidity level by considering both the number and severity of 17 pre-defined comorbid conditions.

Conclusion

During these first 5 months at the CHU of Dijon, I had the opportunity to familiarize with the SAS language and with the way the PMSI hospital database works. I had enough time to integrate these new elements, both in terms of PMSI and SAS. Despite some more difficult aspects to assimilate, I am now at ease, thanks to the wise advice and availability of my training supervisor Mr. Cottenet. I had the support of Mr. Nafetat's courses in SAS, and the courses of Mr. Cardot and Mrs. Truntzer in Statistical Methodology, which allowed me to understand more easily both the syntax of SAS commands and the statistical procedures proposed by this language.

From a personal point of view, the objectives of this first phase of the work-study program have been achieved. Having no knowledge of SAS before this second year of the Master's program, it was a stressful factor during the first few weeks at the CHU to have to work entirely on this language. However, being able to discover SAS step by step, in class as well as at the CHU, made me feel very comfortable and I am grateful to have been given the opportunity to apply each of the newly discovered methods to medical data.

Another key point of this first phase was to familiarize with the PMSI. Its functioning is complex and understanding how to use the data, whether it be for a search for rehospitalizations or even finding diagnoses, required a great deal of attention and many attempts before being fully understood.

These first months at the CHU of Dijon taught me a lot, both on a practical aspect with the knowledge of SAS and on a more personal aspect. Actually, the discovery of the PMSI and its functioning opened my eyes to the importance of the quality and the accuracy of the data used for biostatistical research. Both the DIM and the SBIM attach a major importance to this. Moreover, the research projects carried out at SBIM are conducted with great insight and a lot of concern in the department and I hope to do the same in the coming months, now having the necessary skills to launch my study project.

Appendix

Importation of PMSI data (2013)

To provide an overview of the import process, below is a part of the Excel file describing the format and location of the variables in the text file. The constant part stops at position 223. The variable « *Type d'autorisation à portée globale valide n°nb* » is variable as is the DAS part.

Fichier des Résumés de Sortie Anonymes (RSA) 2013 (format 220)

Libellé	Taille	Début	Fin
Numéro FINESS	9	1	9
Numéro de version du format du RSA	3	10	12
N° d'index du RSA	10	13	22
Numéro de version du format du "RSS-groupé"	3	23	25
…	…	…	…
N° séquentiel du RUM ayant fourni le DP	2	201	202
Diagnostic principal (DP)	6	203	208
Diagnostic relié (DR)	6	209	214
Nombre de diagnostics associés dans ce RSA	4	215	218
Nombre de zones d'actes dans ce RSA	5	219	223
Type d'autorisation à portée globale valide n° 1	2	224	225
…	…		
Type d'autorisation à portée globale valide n° Nb_AutPGV	2		
… …	…	…	…
DA n° 1 du RUM n° 1	6		
…	…	…	…
DA n° Nb_DA_R_1 du RUM n° 1	6		
…	…	…	…
DA n° 1 du RUM n° NbRUM	6		
…	…	…	…
DA n° Nb_DA_R_NbRUM du RUM n° NbRUM	6		

Table 1 : Extract from the Excel file describing the format and location of the variables

Note: Some parts have been excluded from the table, the purpose being to give an idea of the formats.

As for the previous table, only pieces of the SAS code will be given, in order to give a brief overview considering the total length of the SAS code. For a better understanding of the code process, each step has been commented, the constant part and the variable part have also been indicated.

SAS Program :

```
DATA RSA2013 DAS2013 Actes2013;

        /* We read the source file */

        INFILE "E:\SBIM\Justine\RSA 2013\rsa13.txt.000";


        /*********************************/
        /* -Constant information import- */
        /*********************************/


        /* Each variable is defined in adequacy with the stated formats */

        ATTRIB FINESS       format = $9.;
        ATTRIB NoRSA        length =  3.;
        ATTRIB RSA          length =  7.;

        ...

        ATTRIB DP           format = $5.;
        ATTRIB DR           format = $5.;
        ATTRIB Nb_DAS       length =  4.;
        ATTRIB Nb_Actes     length =  4.;


        /* We give the locations of each variable */

        INPUT   FINESS              $      1-9
                NoRSA                      10-12
                NoIndex                    13-22

                ...

                DP              $      203-208
                DR              $      209-214
                Nb_DAS                 215-218
                Nb_Actes               219-223      @;


        /* We label the variables according to the French nomenclature */

        LABEL   FINESS              = "Numéro FINESS "
                NoRSA               = "Numéro de version du format du RSA"
                NoIndex             = "Numéro d'Index"

                ...

                DP              = "Diagnostic Principal"
                DR              = "Diagnostic Relié"
                Nb_DAS          = "Nombre de Diagnostics Associés Significatifs"
                Nb_Actes        = "Nombre d'Actes";


        /* We output the collected information in the RSA table */

        OUTPUT RSA2013;
```

```
        /*********************************/
        /* -Variable information import- */
        /*********************************/


        /* We start by importing the variable part on DAS */

        /* The 'start' variable allows us to position ourselves at a specific position
        in the file */

        start = 224 + Nb_AutPGV*2 + Nb_RDTH*7 + Nb_RUM*58;


        /* We get the information for the number 'Nb_DAS' of DAS, using a loop */

        do i = 1 to (Nb_DAS);


               /* The information is collected from the position given by 'start' */

               INPUT @(start)  Code        $6.   @;

               LABEL Code = "Diagnostic Associé Significatif";

               /* We output the collected information in the DAS table */

               OUTPUT DAS2013;

               /* We increment the 'start' variable to go to the next DAS information */

               start= start + 6;
        end;


        /* We import the variable part about Actes */

        /* We proceed in the same way for medical procedures */

        start = 224 + Nb_AutPGV*2 + Nb_RDTH*7 + Nb_RUM*58 + Nb_DAS*6;

        do i = 1 to (Nb_Actes);

               /* We will specify the position for each information using the variable
               'start' */

               INPUT @(start)             Delai    3.
                       @(start)+3         Code    $7.

                       ...

                       @(start)+19        Nb       $2.         @;

               LABEL Delai = "Délai depuis la date d'entrée"
                     Code = "Code CCAM"

                       ...

                      Nb = "Nombre de réalisations de l'acte";

               OUTPUT Actes2013;
               start = start + 22;

        end;

        /* We drop the variables that are no longer needed */

        DROP start i;

RUN;
```