



Rapport de stage

Stage au sein de l'unité de recherche clinique de
l'Est Parisien (URC-Est) : programmation de
l'analyse statistique d'une recherche clinique

Maitresse de stage : Alexandra ROUSSEAU

Année de formation : 2023-2024

Université et la filière : IUT de Paris – Rives de Seine,
BUT2 Science des données
parcours : Exploration et modélisation statistiques

Service : URC-Est

Période : 15/04/2024 - 26/07/2024

Remerciements

Tout d'abord, je tiens à remercier Alexandra ROUSSEAU, responsable de l'équipe biométrie et ma maîtresse de stage, qui m'a donné l'opportunité d'effectuer mon stage en tant que biostatisticienne à l'URC-Est. Je la remercie également pour son accompagnement tout au long de mon stage, sa patience dans la relecture de mon rapport de stage et la préparation de ma soutenance de stage.

Ensuite, je remercie également Sophie TUFFET, biostatisticienne, pour sa formation à l'utilisation d'une macro SAS, ainsi que pour la préparation de ma soutenance de stage, et pour son aide lorsque j'avais des questions.

Puis, je remercie aussi Soufiane et Mariam, biostatisticiens qui partagent mon bureau et Dominique, data manager, pour leur aide lorsque j'avais des questions.

Sommaire

Remerciements	2
Sommaire	3
Introduction.....	4
1. Les Unités de Recherche Clinique (URC)	5
1.1 Les URCs	5
1.2 Unité de Recherche Clinique de l'Est Parisien(l'URC-EST).....	6
1.2.1 Equipe biométrie	7
2. Le protocole.....	8
2.1 Le protocole de recherche	8
2.2 Le protocole "GABRIEL "	9
2.2.1 : Le schéma d'étude.....	9
2.2.2 : Déroulement de la recherche.....	10
2.2.3 : Avancement de l'étude	10
2.3 Le protocole de recherche pour les biostatisticiens	11
3. La base de données	11
3.1 Gestion des données	11
3.2 les données du protocole GABRIEL	12
4. Les missions	14
4.1 programmation SAS : statistique descriptive	15
4.2 Programmation SAS : Création de variables.....	17
4.2.1 : Le critère d'évaluation principal.....	18
4.2.2 : Les critères d'évaluation secondaire	19
5. Les difficultés rencontrées	21
6. Conclusion	22
Bibliographie.....	23
Tables des illustrations	23
Démarche portfolio	24

Introduction

J'ai effectué mon stage de 2^e année de BUT science des données à l'Unité de Recherche Clinique de l'Est Parisien(URC-EST), basée à l'hôpital Saint-Antoine. L'unité dépend de la Direction de la Recherche Clinique, de l'Innovation de l'AP-HP et est dirigée par le Pr Tabassome Simon. J'ai décidé d'effectuer mon stage à URC-EST car je souhaitais découvrir le domaine de la santé, un domaine passionnant qui permet de contribuer au bien-être collectif. Mon stage se déroule entre le 15/04/2024 au 26/07/2024. La mission principale que la responsable m'a confiée est la réalisation de l'analyses statistiques d'un protocole de recherche clinique portant sur la comparaison de deux traitements après la transplantation rénale en utilisant le logiciel SAS. Donc la mission est en lien avec mes compétences acquises pendant ces 2 ans en BUT science des données.

1. Les Unités de Recherche Clinique (URC)

1.1 Les URCs

La Direction de la Recherche Clinique, de l'Innovation (DRCI) de l'AP-HP, rattachée à la Direction de l'organisation médicale et des relations avec les universités, est chargée de piloter les projets de recherche développés par l'AP-HP et de suivre l'ensemble des activités de recherche se déroulant au sein de l'institution.

La création des Unités de Recherche Clinique (URC) par la DRCI de l'APHP d'Ile de France s'explique par un nombre important de protocoles de recherche gérés et suivis par l'AP-HP, pour permettre la gestion déconcentrée de certaines de ses missions auprès des Groupes Hospitalo-universitaires (GHU).

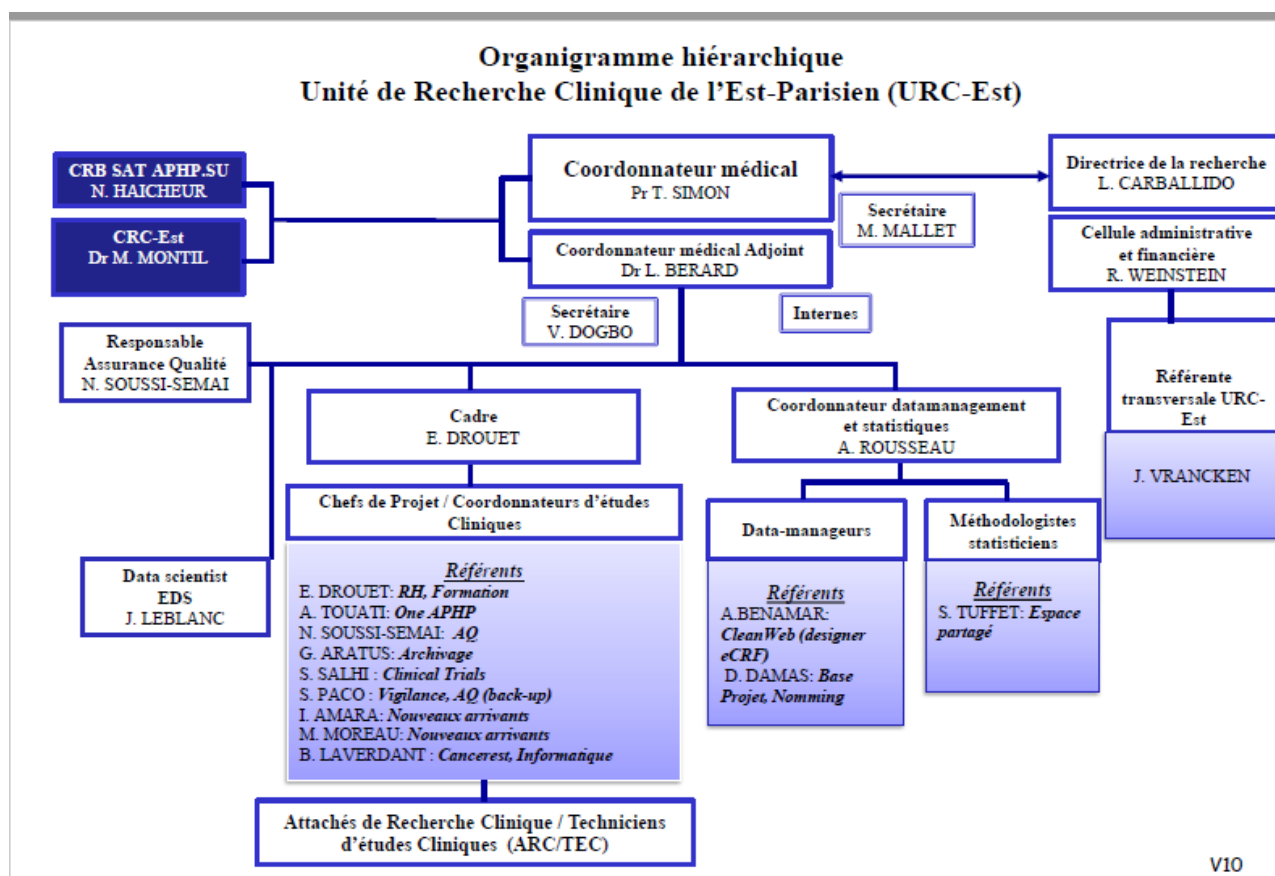
Les missions principales confiées aux unités de recherche clinique sont nombreuses :

- Aider à la conception des programmes de recherche et assurer leur accompagnement (méthodologie, biostatistique, conformité, faisabilité, budgets prévisionnels, réponses aux appels d'offres).
- Assurer le suivi et le contrôle de la qualité des recherches.
- Gérer les bases de données des projets.
- Participer à la coordination des activités de recherche du groupe hospitalier.

1.2 Unité de Recherche Clinique de l'Est Parisien(l'URC-EST)

L'URC-EST située à l'hôpital Saint-Antoine est l'une des 13 URCs de l'APHP.

L'URC-Est a été créée en 2002 pour aider au développement et à la professionnalisation de la recherche clinique dans le GH (Hôpital Tenon, Trousseau, Saint-Antoine, Rothschild et CHNO des Quinze-Vingts). Autrement dit, l'Unité de Recherche Clinique de l'Est Parisien (URC-Est) a été créée pour apporter aux cliniciens qui en font la demande une aide pour les recherches cliniques qu'ils souhaitent entreprendre. L'URC-Est compte aujourd'hui environ 130 personnes. Ce développement s'est accompagné d'une structuration avec la mise en place d'une équipe méthodologique, d'une équipe administrative et financière, d'une équipe secrétariat, d'une équipe logistique et gestion des essais, d'une équipe Assurance Qualité et d'une équipe biométrie.



V10

SCHEMA 1 : ORGANIGRAMME DE L'URC-EST

L'URC-Est participe à l'élaboration de projets de recherche et à leur mise en œuvre. Elle est notamment en charge du data management et de l'analyse statistique des essais.

L'URC-Est remplit également une importante mission universitaire en participant notamment à l'encadrement et à la formation d'internes et de stagiaires issus de formation universitaire (Master, DIU,...) ou privée.

1.2.1 Equipe biométrie

Mon stage se déroule au sein de l'équipe Biométrie de l'URC-EST, où j'ai travaillé en tant que stagiaire biostatisticienne. L'équipe biométrie est composée de la responsable d'équipe, méthodologiste statisticienne, de 5 Data managers dont un alternant, de 5 biostatisticien.nes et 3 stagiaires biostatisticien.nes et en lien avec l'entrepôt de données de santé de l'APHP, 2 data scientifiques et 2 stagiaires.

Je suis notamment en contact avec la coordonnatrice data management et statistique (équipe biométrie) qui est également ma maîtresse de stage, la méthodologiste-biostatisticienne, ainsi que deux autres biostatisticien.nes qu'on est dans un même bureau. Ce sont des personnes à qui je peux poser des questions lorsque je rencontre des difficultés.

2. Le protocole

2.1 Le protocole de recherche

Le protocole de recherche est un document daté, approuvé par le promoteur et le médecin qui porte le projet décrivant le ou les objectifs, la conception, la méthode, les aspects statistiques qui seront rédigés par les biostatisticiens et l'organisation de la recherche. Le protocole se déroule de manière logique, comme une histoire avec un commencement, une partie centrale. C'est-à-dire qui commence par justifier l'étude compte tenu des connaissances actuelles, la formulation des hypothèses de la recherche soit la question scientifique, en suivant avec la description des connaissances relatives à la pathologie concernée. Ensuite, on trouve dans le protocole le résumé des expérimentations pré-cliniques et des essais cliniques concernés, la description de la population à étudier et la justification de son choix, la description succincte du ou des produit(s) soit les médicaments ou acte(s) soit les chirurgies expérimentales. Enfin, il y a aussi le résumé des bénéfices et des risques prévisibles et connus pour les personnes se prêtant à la recherche. Le médecin définit un seul objectif principal qui est aussi la question principale posée par la recherche et quelques objectifs secondaires qui sont des questions secondaires ou complémentaires. Pour répondre à ces questions, il faut ainsi définir un seul critère principal pour l'objectif principal et quelques critères secondaires pour les objectifs secondaires.

A partir du critère principal et des effets attendu des traitements sur le critère : le nombre de sujets nécessaire pour répondre à la question est calculé.

2.2 Le protocole "GABRIEL"

Le protocole de recherche sur lequel j'ai travaillé pendant mon stage intitulé "GABRIEL" qui est l'acronyme du titre complet de protocole en anglais : "Graft Acute kidney injury: vitamin B3 to facilitate renal Recovery In the Early Life of a transplant" . Le chercheur cherche à étudier si la vitamine B3 soit le nicotinamide (NAM) permet d'améliorer la reprise précoce de fonction du greffon rénal après la transplantation.

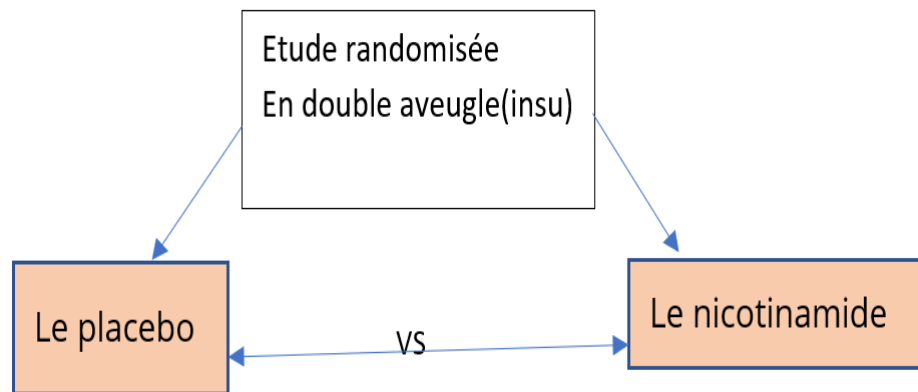
D'après le protocole, la justification scientifique de cette étude est que : " la reprise retardée de fonction (RRF) est fréquente après une transplantation rénale et compromet la survie du greffon. Elle est principalement due à la nécrose tubulaire aiguë (NTA) causée par des épisodes ischémiques. A ce jour, il n'existe pas de traitement spécifique pour la NTA. Le nicotinamide (NAM) a été identifié comme un traitement potentiel pour prévenir ou accélérer la récupération de la NTA. Une étude chez 41 patients ayant une chirurgie cardiaque à haut risque rénal a montré l'efficacité avec une diminution de 35% de l'insuffisance rénale aiguë (IRA). La sécurité du NAM est bien étudiée et établie. L'étude GABRIEL vise à évaluer le potentiel thérapeutique du NAM dans le contexte de la RRF. "

2.2.1 : Le schéma d'étude

Ce protocole est à un essai comparatif randomisé en 2 groupes parallèles, qui est la seule méthode permettant d'avoir des groupes comparables.

Ces 2 groupes sont les patients qui ont pris le traitement placebo et les patients qui ont pris le traitement nicotinamide (NAM).

Donc le but est de comparer les patients qui ont pris le traitement placebo avec ceux qui ont pris le nicotinamide (NAM). Autrement dit le traitement qui ne contient pas d'excipient (le placebo) contre le traitement qui contient l'excipient comme le nicotinamide (NAM).



SCHEMA 2 : SCHEMA DE RANDOMISATION

2.2.2 : Déroulement de la recherche

Tout d'abord, on commence par inclure les patients, s'ils remplissent les critères d'inclusion. Après inclusion et randomisation, les patients commencent une visite J0, c'est-à-dire la visite pré-greffe. Ensuite, après la transplantation, il aura un premier suivi, qui correspond au premier matin post-opératoire (POD1). Et il aura un deuxième jour de suivi (POD2). Puis, il aura un suivi de 7^e jour post-opératoire (POD7). Enfin, on attend 3 mois après la transplantation chez les patients pour effectuer une visite(M3). Donc, au total on compte 5 visites prévues pour les participants.

2.2.3 : Avancement de l'étude

A ce jour, les inclusions sont terminées ainsi que les suivis. Des contrôles sont encore en cours par les attachés de recherche cliniques et le data manager de l'étude. La base de données n'est donc pas gelée, ce qui signifie que les biostatisticiens ne peuvent pas encore commencer les analyses statistiques, mais ils peuvent commencer à programmer l'analyse. Le plan d'analyse statistique est en cours de rédaction.

2.3 Le protocole de recherche pour les biostatisticiens

En tant que stagiaire biostatisticienne, il faut commencer par comprendre le protocole d'une recherche clinique, pour cela, j'ai eu avec un autre stagiaire biostatisticien et un alternant data manager une formation d'une à deux heures consacrer pour initiation au protocole avec la maîtresse de stage. Il est important de comprendre le protocole, car on appuie sur ce dernier pour effectuer des analyses statistiques par la suite. Parce que le protocole est le document de référence de la recherche, les biostatisticiens s'intéressent notamment la partie des objectifs, les critères d'évaluation et la partie concerne les aspects statistiques.

3. La base de données

3.1 Gestion des données

La base de données sur laquelle j'ai travaillé est déjà existante, même si non définitive. Les données des patients ont été saisies sur un e-CRF qui est un cahier d'observation électronique, soit une interface numérique qui, pour chaque patient participant à une recherche, rassemble les informations définies par le protocole correspondant. Ensuite, ce sont les data-managers qui sont responsables des nettoyages, des formatages et de contrôler ces données. Puis, ce sont les biostatisticiens qui sont responsables des analyses statistiques de ces données. Nous utilisons tous le logiciel SAS version 9.4 pour traiter et analyser ces données.

3.2 les données du protocole GABRIEL

Pour le protocole GABRIEL, il y a 7 tables différentes, dont la table principale est la table "gabriel" qui contient les informations du patient. Les autres 6 tables contiennent des informations supplémentaires sur les patients, autrement dit ces tables détaillent les informations qui ne se trouvent pas dans la table "gabriel" des patients. Par exemple, la table "tabeing" détaille les informations sur les événements indésirables non graves des patients, la table "tabrehospitm3" donne les informations sur les patients, s'ils sont réhospitalisés après 3 mois de transplantation.








 gabriel_20240415	16/04/2024 10:04	SAS Data Set	1 792 Ko
 tabeing_20240415	16/04/2024 10:04	SAS Data Set	576 Ko
 tabrehospitm3_20240415	16/04/2024 10:04	SAS Data Set	272 Ko
 tabsuivialat_20240415	16/04/2024 10:04	SAS Data Set	192 Ko
 tabttcon_20240415	16/04/2024 10:04	SAS Data Set	3 200 Ko
 tbleig_procact_20240415	16/04/2024 10:04	SAS Data Set	128 Ko
 tbleig_trtco_20240415	16/04/2024 10:04	SAS Data Set	192 Ko

IMAGE 1 : LES 7 TABLES DE LA BASE DE DONNEES

Les tableaux ci-dessous représentent une partie de la base de données, avec les informations de la table SAS associé, grâce à la procédure contents de SAS. Ici, la base de donnée de la table z.gabriel_20240415 que j'ai travaillé ne peut pas être afficher, car les données sont confidentielles. Mais on voit que la table gabriel compte 610 variables, avec seulement 206 observations soient une observation par patient. La variable SUBJECT_REF de la table signifie la référence du patient, qui est unique dans la table gabriel. Or, la table tabeing qui détaille les informations sur les patients, donc qui compte ici plus d'observations soit 779, car un patient peut avoir plusieurs observations, comme on peut voir dans la table eing ci-dessous. Et il n'y a seulement 25 variables pour cette table.

Le Système SAS

La procédure CONTENTS

Nom de la table	Z.GABRIEL_20240415	Observations	206
Type de membre	DATA	Variables	610
Moteur	V9	Index	0
Créée	16/04/2024 10:04:50	Longueur d'observation	6816
Dernière modification	16/04/2024 10:04:50	Observations supprimées	0
Protection		Compressée	NON
Type de table		Triée	NON
Libellé			
Représentation des données	WINDOWS_64		
Codage	wlatin1 Western (Windows)		

TABEAU 1 : LA SORTIE SAS PAR LA PROCEDURE CONTENTS DE LA TABLE GABRIEL

Le Système

Obs.	SUBJECT_REF	EINGYN	EINGYN_V	NUMCTR	RANDOMIZATION_R1	FORM_TABEING	NATEING
1	001-0001-B-M	Oui	Oui	001	B	1	HYPOPHOSPHATEMIE
2	001-0001-B-M	Oui	Oui	001	B	2	DOULEURS AU NIVEAU DU GREFFON
3	001-0002-V-P	Oui	Oui	001	B	1	HEMATOME PERI GREFFON RENALE ET SIGMOIDITE AIGUE
4	001-0002-V-P	Oui	Oui	001	B	2	HYPOPHOSPHOREMIE < 0.5 MMOL/L
5	001-0003-M-J	Oui	Oui	001	A	1	ANEMIE
6	001-0004-K-C	Oui	Oui	001	B	1	ANEMIE
7	001-0006-J-X	Oui	Oui	001	A	1	CYTOLYSE HEPATIQUE LEGERE
8	001-0007-O-G	Oui	Oui	001	B	1	HYPOPHOSPHOREMIE
9	001-0007-O-G	Oui	Oui	001	B	2	ANEMIE
10	001-0007-O-G	Oui	Oui	001	B	3	CYTOLYSE HEPATIQUE
11	001-0008-P-O	Oui	Oui	001	A	1	HYPHOSPHOREMIE
12	001-0008-P-O	Oui	Oui	001	A	2	LYMPHOCELE
13	001-0009-L-J	Oui	Oui	001	A	1	ANEMIE NORMOCYTAIRE A 7.8
14	001-0010-D-D	Oui	Oui	001	B	2	HYPOPHOSPHOREMIE
15	001-0010-D-D	Oui	Oui	001	B	3	HTA
16	001-0010-D-D	Oui	Oui	001	B	4	ANEMIE NORMOCYTAIE A 7.7
17	001-0010-D-D	Oui	Oui	001	B	5	HYPERKALIEMIE
18	001-0011-P-M	Oui	Oui	001	B	1	DYSPNEE
19	001-0011-P-M	Oui	Oui	001	B	2	ANEMIE NORMOCYTAIRE A 7.8
20	001-0012-H-Y	Oui	Oui	001	A	1	ANEMIE
21	001-0013-M-M	Oui	Oui	001	A	1	ANEMIE

TABEAU 2 : UNE PARTIE DE LA BASE DE DONNEES EING

Le Système SAS

La procédure CONTENTS

Nom de la table	Z.TABEING_20240415	Observations	779
Type de membre	DATA	Variables	25
Moteur	V9	Index	0
Créée	16/04/2024 10:04:50	Longueur d'observation	648
Dernière modification	16/04/2024 10:04:50	Observations supprimées	0
Protection		Compressée	NON
Type de table		Triée	NON
Libellé			
Représentation des données	WINDOWS_64		
Codage	wlatin1 Western (Windows)		

TABEAU 3 : LA SORTIE SAS PAR LA PROCEDURE CONTENTS DE LA TABLE EING

4. Les missions

Durant mon stage, j'ai mobilisé la compétence 1 « Traiter des données à des fins décisionnelles », telle que la programmation statistique en SAS et la compétence 2 « Analyser statistiquement des données » telle que la statistique descriptive que j'ai acquise pendant ces 2 années en BUT Science des données.

La mission principale se concentre autour de la statistique descriptive et de la programmation statistique en utilisant le logiciel SAS à partir des variables qui sont en lien avec le critère d'évaluation principal et des critères d'évaluation secondaires du protocole.

4.1 programmation SAS : statistique descriptive

Dans un premier temps, j'étais amenée à faire une analyse descriptive sur les variables de la caractéristique initiale des patients que la responsable a choisi. En utilisant les procédures SAS adaptée que j'ai déjà utilisé dans mon cours de SAS, telles que le proc freq pour les variables qualitatives, qui permet d'obtenir la fréquence de différentes modalités et le nombre de valeur manquante. Puis, le proc means pour les variables quantitatives, qui permet d'obtenir la moyenne, le minimum, le maximum, l'écart-type, la médian, et p25 et p75 pour l'écart interquartile. L'option SAS p25 et p75, je n'ai jamais vu en cours, c'est la maitresse de stage qui m'a appris cette option. Et, s'il s'agit une variable quantitative, on s'intéresse aussi à sa distribution normale, ainsi on produit un histogramme. Comme nous montre l'image ci-dessous.

```
***** ;
*   La variable: Créatinine (µmol/L) =CREATJ0      ;
***** ;

❏ PROC MEANS DATA=a.gab N nmiss MEAN STD MEDIAN p25 p75 MIN MAX
    maxdec=2;
    var CREATJ0_V;
run;

/* Créer un histogramme de créatinine (µmol/L) */
❏ proc sgplot data=a.gab;
    histogram CREATJ0_V;
run;

***** ;
*   La variable: Cause du décès =CAUDECES          ;
***** ;

❏ proc freq data=a.gab;
    tables CAUDECES;
run;
```

IMAGE 2 : UNE PARTIE DU PROGRAMME SAS POUR LA STATISTIQUE DESCRIPTIVE

D'autre part, il existe une autre manière pour réaliser les statistiques descriptives. Les biostatisticiens utilisent un programme SAS Macro déjà existant. Ce SAS Macro permet de donner plus d'information sur la descriptive des variables en fonction du groupe de randomisation.

Ainsi, pour mieux comprendre ce programme SAS Macro et savoir l'utiliser par la suite, moi et un autre stagiaire avaient une formation de 2 heures avec une biostatisticienne, pour l'utilisation de ce SAS Macro.

```
/*Table Antecedent*/
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=CAUMALRSP1_v , ngroupe=2, testl=non);
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=CAUMALRSP2_v , ngroupe=2, testl=non);
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=CAUMALRSP3_v , ngroupe=2, testl=non);
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=CAUMALRSP4_v , ngroupe=2, testl=non);
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=DIABET_v , ngroupe=2, testl=non);
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=HTA_v , ngroupe=2, testl=non);
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=ATCDCARD_v , ngroupe=2, testl=non);
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=ATCDDIAL_v , ngroupe=2, testl=non);
%descrquali1 (tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=ATCDGREF_v , ngroupe=2, testl=non);

/*Tableau pour ExamenBio*/
%descrquantil(tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=HBJ0_v , testl=student);
%descrquantil(tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=PNNJ0_v , testl=student);
%descrquantil(tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=CREATJ0_v , testl=student);
%descrquantil(tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=EGFRJ0_v , testl=student);
%descrquantil(tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=ALATJ0_v , testl=student);
%descrquantil(tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=ALATJ0NSL_v , testl=student);
%descrquantil(tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=GGTJ0_v , testl=student);
%descrquantil(tablel=a.gabR,groupe=RANDOMIZATION_R1_S , varl=GGTJ0NSL_v , testl=student);
```

IMAGE 3 : UNE PARTIE DU PROGRAMME SAS MACRO POUR STATISTIQUE DESCRIPTIVE

Cette image représente le code SAS Macro pour faire de la statistique descriptive. Les différents paramètres représentent le nom de table, le groupe de randomisation, la variable qu'on veut analyser et le test de student par défaut pour une variable continue.

On voit que le code se diffère en fonction du type de la variable, soit quantitative ou qualitative. Donc, avant de l'utiliser, on doit savoir si les variables sont quantitatives ou qualitatives. S'il s'agit d'une variable quantitative, on doit également vérifier sa distribution normale par un histogramme. L'intérêt de la vérification de la distribution normale s'explique par si on peut faire le test de student ou d'Anova pour les analyses statistiques avancées.

Par exemple, l'image représente la distribution de la variable quantitative de l'âge :

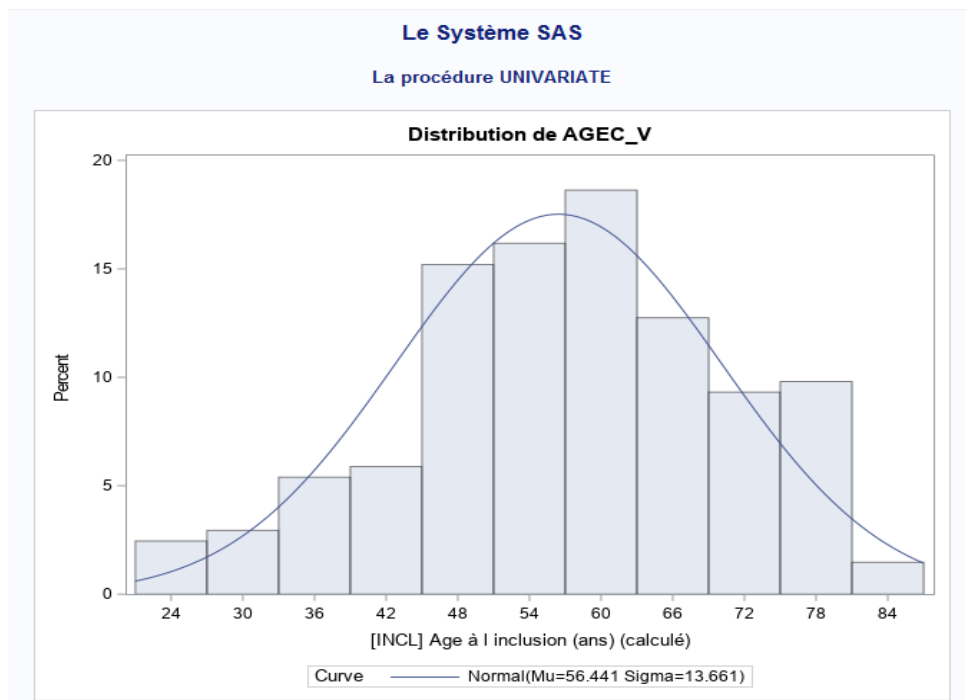


IMAGE 4 : UN EXEMPLE DE L'HISTOGRAMME POUR VERIFIER LA DISTRIBUTION NORMALE

4.2 Programmation SAS : Création de variables

Une fois, j'ai fini de faire la statistique descriptive, l'étape suivante consiste à la création des variables. Celle-ci est basée sur les variables qui ont le lien avec les critères d'évaluation. En commençant par repérer ces variables utiles qui vont se servir à coder les critères d'évaluations. Elles se retrouvent sur un document nommé "dhm" qui signifie en anglais "data handling manual" soit en français : manuel de traitement de données, qui correspond à un dictionnaire des données de la base de données.

L'objectif de la création des variables est de permettre de faire des analyses statistiques à la suite, pour répondre à l'objectif principal et secondaire.

4.2.1 : Le critère d'évaluation principal

Critère d'évaluation principal est : Le ratio de diminution de la créatinine calculé à POD2 (2ème jour post- opératoire) selon la formule suivante :

$$CRR2(\%) = \frac{([Cr1 - Cr2] \times 100)}{Cr1}$$

D'où Cr1 et Cr2 correspondent à la créatinine sérique du matin au jour 1 et au jour 2 post-opératoire

L'image ci-dessous illustre une partie de code SAS qui consiste à la création d'une variable pour le critère d'évaluation principal. Ce dernier est une formule. Donc la variable CRR2 que j'ai créé est une variable quantitative. J'ai commencé par repérer les variables utiles, c'est-à-dire CREATSJ1 et CREATSJ2 depuis le document "dhm". Puis j'ai construit cette variable en commençant par la condition si la variable CREATSJ1 et la variable CREATSJ2 ne sont pas vide, alors la variable CRR2 peut être construite. Et je fais un print, c'est-à-dire un affichage du résultat pour vérifier si la variable construite est correcte.

```
*****;  
* Critère de jugement principal ;  
* La variable: CRR2 ;  
*****;  
  
data a.gabvar;  
set a.gabvar;  
if CREATSJ1_V ne . and CREATSJ2_V ne . then CRR2=((CREATSJ1_V-CREATSJ2_V)*100) /CREATSJ1_V;  
run;  
  
/*view dataset*/  
proc print data=a.gabvar;  
var CREATSJ1_V CREATSJ2_V CRR2;  
run;
```

IMAGE 5 : PROGRAMME SAS POUR LA CREATION DE LA VARIABLE POUR LE CRITERE D'EVALUATION PRINCIPAL

4.2.2 : Les critères d'évaluation secondaire

Ensuite, pour les critères secondaires, j'ai plus de variables à construire.

Les critères d'évaluation secondaire sont nombreux. Parmi l'ensemble des critères d'évaluation secondaire, j'ai travaillé sur les critères :

- Proportion de patients avec une résiduelle élevée de tacrolimus (> 15 ng/mL) à POD2, ou POD7.
- Proportion de patients présentant au moins un épisode de rejet prouvé par biopsie selon les critères de classification de Banff 2017 dans les 3 mois suivant la transplantation.

On voit qu'il s'agit de la création des variables pour la proportion ou encore pour la corrélation, qui permettent de faire les analyses statistiques.

L'image ci-dessous montre une partie de code SAS qui représente les 2 variables binaires que j'ai créé pour les critères d'évaluation secondaires. Par exemple, pour la création de la variable " tacrohigh " pour la concentration résiduelle de Tacrolimus élevée (>15 ng/L) à POD7. J'ai commencé par initialiser la variable et ajouter la condition, si la variable RESTACRO(résiduelle de Tacrolimus) est supérieure à 15, ce qui signifie que la concentration résiduelle de Tacrolimus est élevée et sinon. Ensuite, j'applique le format qui est codé par le data manager pour les modalités, soit 1=Oui et 0=Non de la variable que j'ai créé. Puis je donne un label à cette nouvelle variable.

D'autre part, pour construire la variable rejet, je dois me référer à la table eing que j'ai déjà présenté dans la partie 3.2 les données du protocole GABRIEL. L'objectif est de trouver les patients qui ont au moins un rejet. Pour cela, j'ai commencé avec une condition qui permet de rechercher les modalités qui contiennent le mot rejet. Si le mot rejet est retrouvé alors la variable est construite. Et j'ai un print pour regarder ce que donne le résultat.

```

* Critère de jugement ;
* secondaire ;
*****;
data a.gabvar;
    set a.gabvar;

    /* variable: "tacrohigh": La concentration résiduelle de tacrolimus élevée (> 15 ng/mL) à POD7. */

    if RESTACRO_v=. then tacrohigh =.;
    else if RESTACRO_v>15 then tacrohigh = 1;
    else tacrohigh=0;
    format tacrohigh CODE_3_LIST.;
    label tacrohigh="Tacrolimus résiduel (>15ng/mL) POD7";
run;

data a.transp;
    set a.eing(keep=subject_ref nateing);

    /* Créer une nouvelle variable "rejet": au moins un épisode de rejet...dans les 3 mois suivant la transplantation*/
    if prxmatch("/REJET/i",nateing) then rejet = 1; /*i pour ignorer majuscule et minuscule*/
    else rejet = 0;
    format rejet CODE_3_LIST.;
    label rejet=" 3M- Au moins un épisode de rejet";

proc print data=a.transp;
run;

```

IMAGE 6 : PROGRAMME SAS POUR LA CREATION DE LA VARIABLE POUR LES CRITERES D'EVALUATION SECONDAIRE

Une fois, j'ai ces nouvelles variables, je peux faire des analyses statistiques avec les options SAS, comme proc freq pour avoir la proportion.

5. Les difficultés rencontrées

La première difficulté que j'ai rencontrée est pendant la lecture du protocole.

En effet, le protocole complet de 59 pages est rédigé en anglais, donc parfois, je peux avoir des difficultés pour la compréhension de certains termes médicaux en anglais, le vocabulaire propres à la recherche clinique.

Concernant la programmation en SAS, parfois, je dois utiliser les fonctions SAS qu'on n'avait pas vues en cours de SAS, mais on avait vu en cours de R, comme `proc sgplot` pour les histogrammes qui permettent de faire des analyses statistiques. Donc je dois consacrer un peu de temps pour la recherche d'utilisation de ces fonctions.

D'autre part, parfois, j'ai confondu la clause `where` et `if`. Autrement dit, à la place de `where`, j'ai mis un `if`, ce qui empêche le programme de fonctionner. Et au lieu d'enregistrer mon programme dans l'environnement de travail, je l'ai enregistré dans la data source, ce qui risque d'écraser la donnée de la data source. Sinon, je n'ai pas rencontré des grandes difficultés, puisque les missions sont simples qui correspondent à ce que j'ai déjà fait en cours et je peux demander aux collègues si j'ai des questions.

6. Conclusion

Ce stage en tant que stagiaire biostatisticienne à l'Urcest n'est pas encore terminé. Mais elle m'a permis de découvrir le métier de biostatisticien. J'ai eu l'occasion de mettre en pratique les connaissances que j'ai acquises lors de mes 2 années en BUT sciences des données, comme la statistique descriptive et la programmation en SAS. J'ai compris l'intérêt de la statistique au service du domaine de la santé. J'aime aussi les missions qu'on avait confiées.

D'autre part, je suis bien accompagné par ma maîtresse de stage. Sauf que si elle est occupée, sinon chaque semaine, on fait un point d'une demi-heure pour discuter ce qu'il va bien et ce qui faut corriger de ce que j'ai fait. Mais elle reste disponible pour me guider pour chaque étape de missions. Je suis assignée à des tâches significatives qui contribuent aux objectifs du service, permettant ainsi l'acquisition de compétences précieuses, tel que la programmation en SAS appliquée au domaine de la santé. Donc, ce stage me confirme mon souhait de poursuivre dans le domaine de la data. J'aime les statistiques et analyser les données. Et j'aimerais approfondir mes connaissances en statistique en faisant un Master qui se spécialise dans la science des données.

Bibliographie

Hélène Kontchou Kouomegni et Olivier Decourt, *Maîtriser SAS Base et SAS Macro SAS 9.2 et version antérieures* (Dunod, 2007)

Tables des illustrations

Image 1 : Les 7 tables de la base de données	12
Image 2 : Une partie du programme SAS pour la statistique descriptive	15
Image 3 : Une partie du programme SAS Macro pour statistique descriptive	16
Image 4 : Un exemple de l'histogramme pour vérifier la distribution normale	17
Image 5 : Programme SAS pour la création de la variable pour le critère d'évaluation principal	18
Image 6 : Programme SAS pour la création de la variable pour les critères d'évaluation secondaire .	20
Tableau 1 : La sortie SAS par la procédure contents de la table Gabriel	13
Tableau 2 : Une partie de la base de données Eing	13
Tableau 3 : La sortie SAS par la procédure contents de la table Eing	14
Schéma 1 : Organigramme de l'URC-Est	6
Schéma 2 : Schéma de randomisation	10

Démarche portfolio

Tout d'abord, ma mission a un lien avec la compétence 1 « Traiter des données à des fins décisionnelles », car j'ai eu l'occasion de mettre en pratique mes compétences de la programmation en SAS. C'est-à-dire, j'ai écrit un programme correctement structuré et documenté, respectant les bonnes pratiques, tels que mettre des commentaires sur les programmes que j'ai écrit. Et concernant l'apprentissages critiques, j'ai compris l'organisation des données de l'entreprise. La différence entre les données de l'entreprise et les données que j'ai vues en cours, c'est qu'il existe une base de données qui contient plusieurs tables SAS. Donc, au début, c'est un peu difficile pour moi de savoir quelle table utilisée. Mais avec le temps, je commence à comprendre le fonctionnement de cette base de données. Et j'apprécie l'intérêt de briques logicielles existantes et savoir les utiliser, c'est-à-dire le logiciel SAS que j'ai eu l'occasion de l'utiliser pour la programmation statistique. L'avantage est que ce logiciel correspond au même logiciel que j'ai utilisé dans le cours de SAS. Donc je n'ai pas rencontré des problèmes pour l'utilisation de ce logiciel pendant mon stage.

Ensuite, ma mission a un lien avec la compétence 2 « Analyser statistiquement des données ». Pour la composante essentielle, il existe un contexte de l'étude. La responsable m'a confié une mission en rapport avec une étude clinique. Pour l'apprentissages critiques, j'ai pris conscience de la différence entre modélisation statistique et analyse exploratoire. En effet, ma mission de ce début de stage se concentre notamment sur l'analyse exploratoire, c'est-à-dire, j'ai réalisé les statistiques descriptives. Ainsi, grâce au cours de statistiques descriptives, je sais distinguer les variables quantitatives et les variables qualitatives facilement. Cependant, je dois choisir les bonnes variables, c'est-à-dire sélectionner les variables utiles qui permettent de faire les analyses. Comme la table SAS que j'ai travaillé compter 610 variables ce qui est énorme. Je n'ai jamais travaillé en cours une table qui contient autant de variables.

Donc je dois consacrer plus de temps dans la recherche de la variable utile. Et à travers de ce stage, je comprends l'intérêt des analyses multivariées pour synthétiser et résumer l'information portée par plusieurs variables. Par exemple, pour la mission création de la variable, je dois utiliser plusieurs variables pour construire une nouvelle variable pour le critère d'évaluation. Par contre, je n'ai pas eu l'occasion de faire la modélisation statistique, car elle intervient lorsque la statistique descriptive est finalisée. Or, je suis toujours à l'étape statistiques descriptives.

La compétence 3 « Valoriser une production dans un contexte professionnel » et la compétence 4 « Modéliser les données dans un cadre statistique » ne sont pas mobilisées durant mon stage.