La gestion des données manquantes pour l'analyse statistique de scores de douleur dans les essais cliniques

Thomas Perretti

Thomas PERRETTI
Master 2 Statistique

Année scolaire : 2010-2011

RAPPORT DE STAGE DE FIN D'ETUDES

Structure d'accueil : **QUINTILES, service biostatistique**

Sujet du stage : La gestion des données manquantes pour l'analyse statistique de scores de douleur dans les essais cliniques.

Date: du 14/02/2011 au 12/08/2011

Maitre de stage : Mme Claire le Bolay – Senior Biostatistician

Lieu du stage : Quintiles Benefit France - Biostatistics Department
Parc d'innovation - Rue Jean Dominique Cassini
B.P 50137
67404 ILLKIRCH Cedex (FRANCE)

Remerciements:

Je remercie l'ensemble des membres du département de statistique de Quintiles France, pour leur accueil, leur gentillesse et leur disponibilité, ce qui m'a permis de mener à bien ma tâche.

Je tiens à remercier tout particulièrement Mme Geneviève Jehl, directrice du service de biostatistique France, qui m'a donné l'opportunité de réaliser ce stage. Ainsi que Mme Claire Le Bolay qui m'a suivi tout au long de ces 6 mois et qui m'a beaucoup appris.

Je remercie aussi toutes les personnes avec qui j'ai travaillé pour leur sympathie et leurs précieux conseils.

Table des matières

ABREVIATIONS

Liées aux essais cliniques :

ARC	Attaché de Recherche Clinique
CDISC	Clinical Data Interchange Standards
CRF	Case Report Form
CRO	Contract Research Organization
DM	Data Management
FDA	Food and Drug Administration
IP	Produit d'Investigation
SAP	Pla d'Analyse Statistique

Liées à l'étude:

AAdSS	Average Adjusted Symptom Score
AdSS	Adjusted Symptom Score
ARTSS	Average Rhinoconjunctivitis Total Symptom Score
ARMS	Average Rescue Medication Score
FAS	Full Analysis set
OAS	Oral Allergy Syndrome
PPS	Per Protocol Set
RMS	Rescue Medication Score
RSS	Rhinoconjunctivitis Symptom Score
RTSS	Rhinoconjunctivitis Total Symptom Score
SPT	Skin Prick Test

Statistique :

ANCOVA	Analyse de la Covariance
EM	Expectation Maximization
MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Non At Random
LOCF	Last Observation Caried Forward
MCMC	Markov Chain Monte Carlo
MMRM	Mixed effects Model for Repeated Measures
WGEE	Weighted Generalized Estimating Equations

Remarque : Tous les documents de l'étude étant rédigés en anglais, certains noms de variables, graphiques ou tables n'ont pas été traduits.

INTRODUCTION

Le développement d'un nouveau médicament est très long et nécessite de nombreuses études. Il y a tout d'abord la recherche de molécules qui formeront la base du médicament. Ensuite, les laboratoires pharmaceutiques réalisent des études précliniques sur des animaux. Enfin si cela est concluant, on passe à des études cliniques (ou essais cliniques), c'est-à-dire sur l'homme. On commence par réaliser des études sur un petit nombre de volontaires, afin de tester la toxicologie, de trouver la dose optimale, de tester d'éventuelles interactions avec d'autres traitements et pour se donner une idée de l'efficacité du traitement (études de phases I et II). Puis, si cela est concluant, on passe aux études de phase III, réalisées sur un plus grand nombre de patients afin de tester plus précisément l'efficacité. Toutes ces étapes durent en moyenne une dizaine d'années et coûtent des centaines de millions d'euros. Une fois l'autorisation de mise sur le marché obtenue, il y a encore des études de phase IV qui permettent de suivre le médicament et de montrer d'éventuels effets néfastes que l'on n'aurait pas pu détecter auparavant.

Pour commencer, quelques mots sur le déroulement d'un essai clinique afin de se familiariser avec l'environnement dans lequel j'ai été immergé durant ces six mois de stage : Un essai clinique fait intervenir divers acteurs, dont des statisticiens, qui jouent un rôle important. Il faut commencer par trouver des centres de recrutement ainsi que des médecins investigateurs, qui s'assurent de trouver des patients éligibles afin de les inclure dans l'étude. Ces investigateurs doivent aussi traiter les patients selon les directives du protocole, document de base contenant toutes les informations sur le déroulement de l'étude. Les données des patients sont ensuite recueillies dans un cahier d'observations ou « Case Report Form » (CRF), sous le regard des Attachés de Recherche Clinique (ARC). Ces données sont ensuite saisies et traitées par une équipe de Data Management (DM), puis enfin analysées par des statisticiens.

Une problématique de taille pour les statisticiens est la gestion des données manquantes. En particulier dans des études où les patients doivent rentrer leurs données eux-mêmes par le biais de questionnaires de qualité de vie et d'échelles de douleur. Cela augmente considérablement le nombre de données manquantes, et de ce fait, biaise les résultats obtenus. L'objectif principal de mon stage a été d'approfondir la méthodologie sur le traitement de ces données manquantes, en particulier pour une étude en allergies, basée sur des échelles de douleur.

La présentation de Quintiles et en particulier du département de biostatistique ainsi que les différentes étapes du stage, feront l'objet de ma première partie. Dans une seconde partie, je présenterai l'étude sur laquelle j'ai principalement travaillé, et sur laquelle j'ai appliqué la méthode d'imputation multiple afin de traiter les données manquantes, que j'expose dans la dernière partie.

1 PRESENTATION DE L'ENTREPRISE ET DEROULEMENT DU STAGE

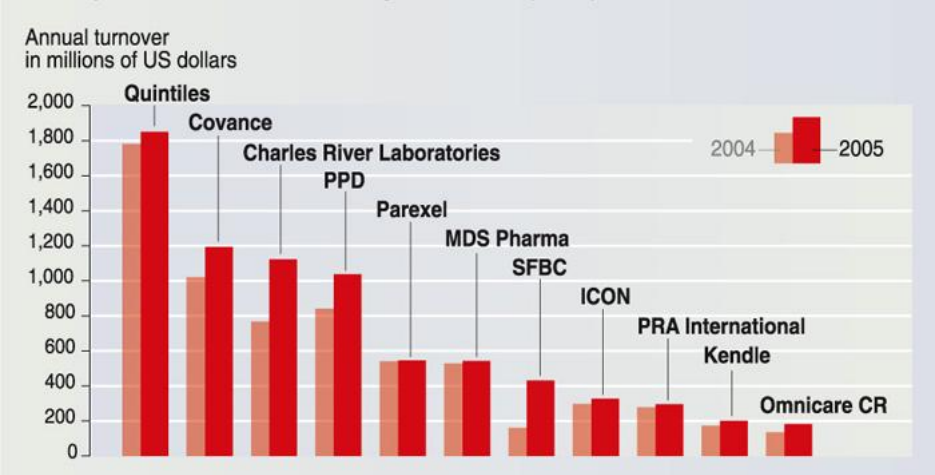
1.1 QUINTILES

1.1.1 Historique

C'est dans les années 1980 que l'on voit apparaître les premières CRO, sociétés de service dans le domaine de la recherche clinique, aux Etats-Unis. Ces sociétés proposent aux laboratoires pharmaceutiques ou aux centres hospitaliers souhaitant réaliser une étude clinique de sous-traiter une partie, voire la totalité, d'un essai clinique.

C'est dans ce contexte, qu'en 1974, le professeur de biostatistiques Dennis Gillings signe son premier contrat de consultant extérieur afin de s'occuper de la gestion des données ainsi que des analyses statistiques d'essais cliniques. Il fonde ensuite Quintiles en 1982, en Caroline du Nord. En 1990 la société s'implante en Europe : Angleterre, Irlande, Allemagne puis en France en 1996 où le groupe acquiert le centre international de recherche clinique Roche à Strasbourg. Basé sur une stratégie d'expansion, Quintiles continue ensuite à s'implanter sur tous les continents et devient ainsi rapidement le leader mondial des sociétés de services dans le domaine pharmaceutique et biotechnologique en dépassant le milliard de revenus nets (1999).

Figure 1 : Revenu annuel des principales CRO



Comme le montre le graphique ci-dessus représentant les principales CRO en 2005, Quintiles était déjà de loin la plus sollicitée, et est encore aujourd'hui en perpétuelle expansion. Actuellement implantée dans plus de 60 pays répartis sur tous les continents, elle emploie plus de 24 000 collaborateurs. Depuis sa création, elle a contribué au développement des 30 molécules les plus vendues dans le monde, et compte parmi ses clients les plus importantes structures pharmaceutiques internationales.

1.1.2 Domaines d'activités

Quintiles accompagne les industries pharmaceutiques et de biotechnologie dans le développement et l'enregistrement de leurs produits. Les services proposés par la compagnie incluent le management des

différents sites lors d'un essai clinique, la gestion et l'analyse des données (biostatistique), la soumission aux autorités de santé, la vente et le marketing des produits pharmaceutiques ainsi que le développement et le conditionnement du matériel nécessaire aux essais cliniques.

Son expérience et ses nombreux clients lui ont permis de maitriser de nombreux domaines d'activités, tel que l'oncologie, le diabète, les maladies cardiovasculaires, le système nerveux, les maladies infectieuses, la médecine interne, etc…

1.1.3 Quintiles France

En France, Quintiles est implantée sur deux sites, un en région parisienne et un à Strasbourg, comprenant environ 200 employés chacun. De plus un réseau d'ARC est implanté dans tout le pays afin de s'assurer du bon déroulement de la récolte des données sur les différents sites d'investigations.

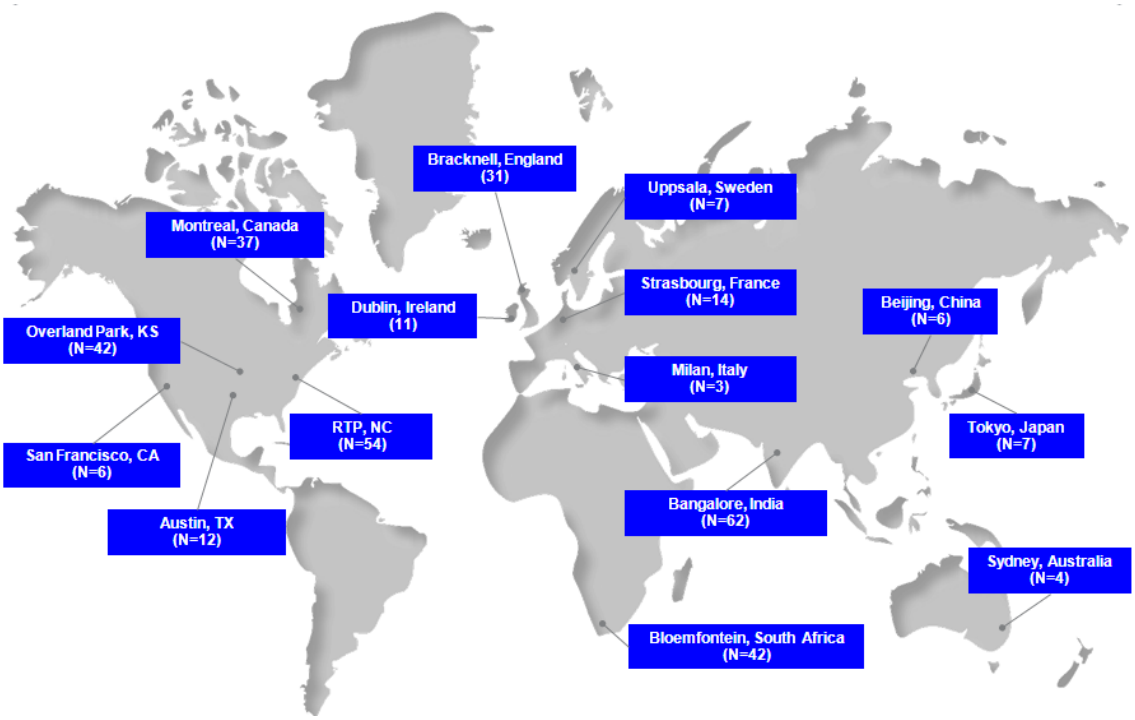
Ce stage s'est donc déroulé sur le site de Strasbourg qui comprend les services suivants :

- Opérations cliniques : chargé de coordonner les essais cliniques en France et à l'étranger.
- Management des données : saisie, validation et codage des données.
- Biostatistique : analyses statistiques des données récoltées.
- Affaires règlementaires : vérifications du bon déroulement de l'étude par rapport aux différentes règlementations des pays impliqués.
- Les départements de supports : L'informatique, les ressources humaines, la comptabilité.

1.2 LE DEPARTEMENT BIOSTATISTIQUE

En 2010 le département était composé de 345 professionnels (biostatisticiens et programmeurs) à travers le monde, comme illustré ci-dessous :

Figure 2 : Répartition du département statistique de Quintiles



1.2.1 Les différentes tâches du biostatisticien

Le département est en contact avec de nombreuses autres équipes impliquées dans le processus des essais cliniques, notamment le Data Management et les rédacteurs médicaux. Le rôle du biostatisticien est large. Celui-ci intervient en effet tout au long du développement d'un médicament en recherche clinique, de la phase I jusqu'à la phase IV. De surcroît, il est important que le statisticien s'assure que les bases de données qui lui sont transmises soient en accord avec les spécifications initiales et qu'elles soient cohérentes, pour lui permettre d'effectuer les analyses statistiques.

Ainsi le biostatisticien peut être amené à réaliser les tâches suivantes :

- Contribution à la rédaction du protocole notamment le design de l'étude. Il s'agit d'évaluer le nombre de sujets nécessaire, les méthodologies statistiques des critères principaux et secondaires.
- Estimation de la taille de l'échantillon de patients à recruter dans l'étude, nécessaire pour avoir suffisamment de patients randomisés, afin d'atteindre les objectifs statistiques fixés.
- Rédaction du plan d'analyse statistique ou « Statistical Analysis Plan » (SAP). Il définit les populations d'analyses, décrit la gestion des données manquantes et des données dérivées, présente les différentes variables étudiées et détermine les analyses et les tests statistiques qui seront réalisés. Il décrit aussi les tableaux, listings et graphiques qui présenteront les résultats.
- Dérivation des jeux de données brutes, afin d'obtenir les données dans un format facilitant la production de tables et de listings. Par exemple, décoder certaines variables ou comparer des dates d'apparition d'effets secondaires afin de savoir s'ils sont apparus avant ou après la première administration du traitement, etc… De plus, pour certaines, études il faut générer ces nouvelles variables dans un format standard CDISC « Clinical Data Interchange Standards » recommandé par la FDA « Food and Drug Administration ».
- Programmation et production des résultats (tableaux, listings, graphiques) pour des analyses intermédiaires ou finales en suivant exactement ce qui a été préalablement rédigé dans le SAP.
- Contrôle Qualité (QC) : le statisticien contrôle les résultats d'une étude dont il n'a pas la charge afin de vérifier et rectifier les éventuelles erreurs commises lors de la programmation.
- Rédaction du rapport statistique en collaboration avec le rédacteur médical pour produire le rapport d'études cliniques qui répond aux questions formulées dans le protocole, mais aussi le rapport d'expert qui synthétise tous les essais réalisés au sein d'un même projet.

Cette liste de tâches n'est pas exhaustive, la profession et les études cliniques évoluant sans cesse.

1.2.2 Le travail effectué durant le stage

Durant le stage j'ai donc été amené à réaliser plusieurs des tâches précédemment citées. Le logiciel utilisé pour la programmation statistique sur le site de Strasbourg est SAS version 9.1.

L'une de mes premières missions a été de rédiger le SAP d'une étude de phase I en oncologie. Pour cela je me suis principalement basé sur le protocole ainsi que sur le SAP d'une étude similaire.

J'ai eu l'occasion de programmer plusieurs listings et tables sur différentes études. Les listings sont des reports de données de la base. Par exemple, lister les antécédents médicaux de chaque patient, ou encore les résultats des analyses médicales obtenus lors des différentes visites. Pour la programmation, ce sont donc en majorité des étapes DATA et de la mise en page (PROC REPORT, ods rtf,…). Les tables présentent des résultats sur les différentes variables que l'on veut étudier. Pour les études de phases I et II, des statistiques descriptives sont programmées en général avec des PROC FREQ et UNIVARIATE. Pour les études de phases III, il faut également tester l'efficacité du traitement, ce qui nécessite la mise en place de différents tests et modèles statistiques.

J'ai aussi été amené à faire plusieurs QC. On s'assure alors que les données du CRF correspondent bien avec la base de données, que les bonnes variables sont reportées ; on vérifie la mise en page et tout ce qui ne serait pas cohérent. Pour les jeux de données dérivées ainsi que pour les tables, une double programmation (par deux programmeurs différents) est réalisée. Puis, à l'aide d'une PROC COMPARE, on compare les résultats des deux programmations afin de s'assurer qu'il n'y ait pas d'erreurs.

Cependant, la principale partie du travail a été réalisée sur une étude en particulier, décrite dans la partie suivante.

1.3 DEROULEMENT DU STAGE

J'ai donc été affilié à une étude spécifique de phase III en allergie. La mission peut être décomposée en deux parties :

- Une partie pratique : le traitement et l'analyse de cette étude en particulier. Cela a nécessité notamment la dérivation de scores et l'utilisation d'échelles de douleur/symptômes, ainsi qu'un travail d'automatisation à l'aide du logiciel SAS.
- Un travail méthodologique sur les données manquantes : gestion des données manquantes, les différentes méthodes d'imputation ainsi que les directives réglementaires sur le sujet.

A mon arrivé l'étude commençait à peine. En attendant que les données soient récoltées, j'ai été formé aux différents processus de Quintiles et pris connaissance de l'étude, notamment à l'aide du protocole et du CRF. Une fois ces bases assimilées et les premières données récoltées, j'ai commencé par la partie pratique : participation à la rédaction du plan d'analyse statistique, programmation des bases de données dérivées, des listings et des tables.

En parallèle, j'ai approfondi la méthodologie sur le traitement des données manquantes et cherché des méthodes qui pourraient être appliquées à cette étude. Mais les données d'efficacité que je devais utiliser pour mes recherches nous sont parvenues peu de temps avant la fin du stage, ce qui ne m'a pas laissé beaucoup de temps pour appliquer toutes les méthodes envisagées. Néanmoins, cela m'a permis de privilégier la pratique et de me perfectionner dans ce domaine.

2 L'ETUDE

L'essai clinique, de phase III, fait partie d'un ensemble d'études ayant pour but de développer un médicament soignant les personnes atteintes d'allergies au pollen (notamment du bouleau pour cet essai en particulier). En effet, actuellement il n'existe aucun traitement de fond contre ces allergies, mais uniquement des médicaments traitant les symptômes.

Ce possible futur médicament, que l'on appellera produit d'investigation (IP), a déjà été testé lors d'études précliniques et cliniques de phases I et II. Ces précédentes études ont permis d'étudier la tolérance, la dose optimale, ainsi que diverses interactions possibles avec d'autres traitements.

2.1 OBJECTIFS

2.1.1 Objectif principal

L'objectif principal est de tester l'efficacité de l'IP à l'aide de la moyenne ajustée des scores de symptômes ou « Average Adjusted Symptom Score » (AAdSS). Cette moyenne est calculée en fonction des 6 symptômes de la rhino-conjonctivite et de la quantité de médicaments administrés pour lutter contre ces symptômes, afin d'estimer plus précisément l'état des patients.

2.1.2 Autres objectifs

L'efficacité du traitement est aussi testée par rapport à d'autres variables que je ne détaillerai pas ici. Cependant ces variables sont principalement basées sur les mêmes variables que l'AAdSS, mais restreintes à différents ensembles de patients ou sur différentes durées. Une étude sur les tests d'allergies effectués sur la peau ou « Skin Prick Test » (SPT) sera aussi effectuée.

La tolérance est également étudiée sur toute la population, ainsi que sur deux sous-populations de patients, en fonction de l'intensité de leurs syndromes d'allergie orale ou « Oral Allergy Syndrome » (OAS). Le choix de ces variables découle des résultats d'études précédentes.

2.2 DEROULEMENT DE L'ETUDE

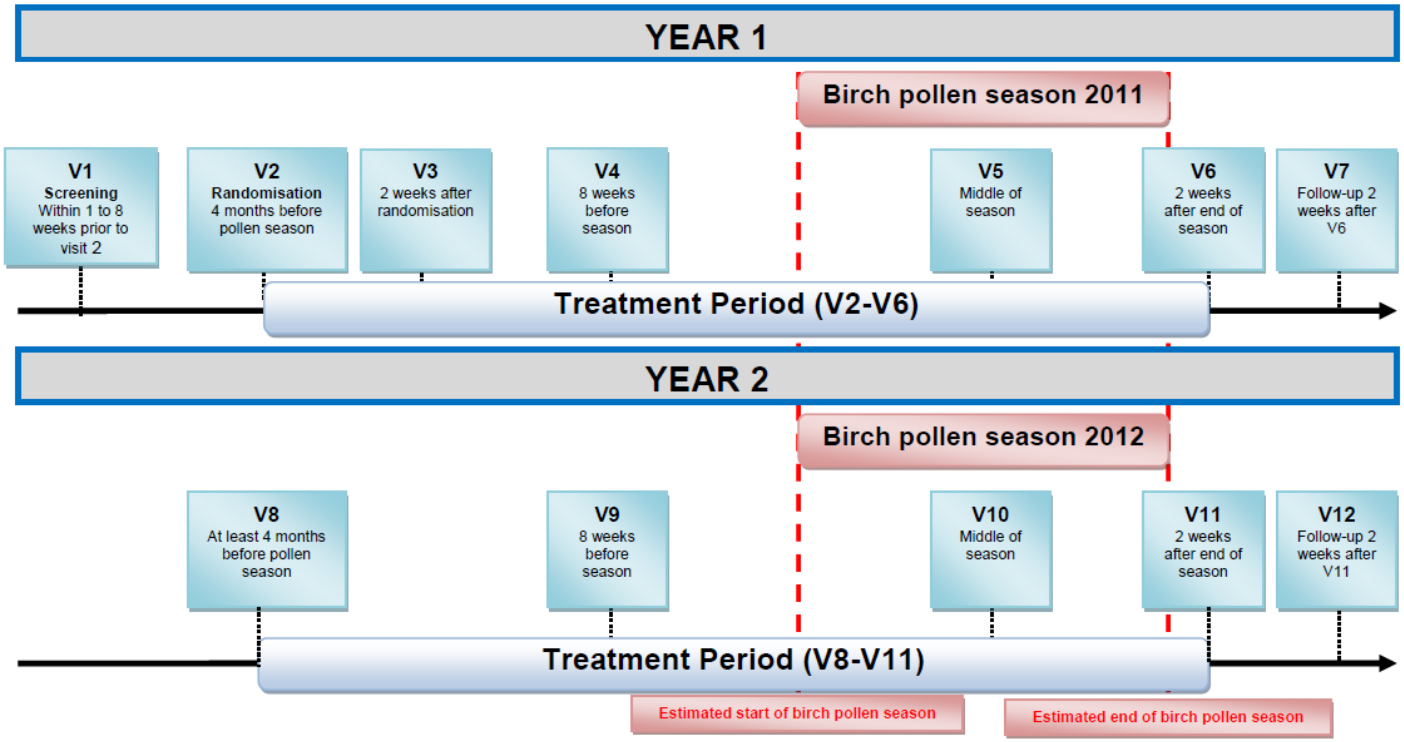
2.2.1 Description générale

Cette étude est dite :

- Randomisée, placebo control (les patient sont répartis aléatoirement dans deux groupes de traitement, Placebo ou IP),
- En double aveugle (personne ne sait quel traitement est administré aux patients avant la fin de l'étude, ni les patients, ni même le médecin),
- Multinationale,
- De phase IIIb.

Le traitement est administré une fois par jour durant approximativement 5 mois par an en fonction de la saison de pollen, pendant deux ans. Les patients devront se présenter à leur centre pour sept visites la première année et cinq l'année suivante, comme illustré ci-dessous :

Figure 3 : Déroulement de l'étude



L'analyse sur laquelle j'ai travaillé est une analyse intermédiaire qui a eu lieu à la fin de la première année. Suite à cette analyse, il sera décidé ou non de continuer l'étude durant la deuxième année. L'analyse principale sera quant à elle réalisée à la fin de la deuxième année.

2.2.2 La récolte des données

On distingue 2 types de données : les données qui sont récoltées lors des différentes visites par l'investigateur et les données d'efficacité qui nécessitent d'être collectées chaque jour à l'aide d'un boitier individuel, durant les périodes de traitement.

Les données récoltées lors des différentes visites sont les suivantes :

- Données administratives,
- Démographiques (sexe, âge,…),
- Historique médicale,
- Examens physiques (signes vitaux, taille, poids,...),
- Tests de laboratoire (analyses sanguines),
- Présence d'asthme, OAS, SPT, marqueurs immunologiques,…
- Effet indésirables,
- Médicaments concomitants.

Les principales données journalières sont :

- La dose de traitement prise,
- L'échelle de douleur des différents symptômes de la rhino-conjonctivite,
- La quantité de médicaments administrés pour lutter contre ces symptômes.

2.2.3 Les patients

<u>Remarque importante :</u> Comme précisé précédemment, cette étude est réalisée en aveugle. Les vrais traitements pris par les patients ne sont pas encore connus ! L'aveugle ne sera levé qu'au dernier moment par une autre équipe de biostatisticien. En attendant, dans la base de données, de faux traitements ont été assignés aux patients. Ce qui implique que les résultats d'analyses ne pourront pas être interprétés correctement.

Les patients sont regroupés en plusieurs ensembles de patients :

- Les **patients examinés** ou « Screened » : Cette population regroupe les patients qui ont été approchés dans le cadre de cette étude, et qui ont signé un consentement écrit.
- Les **patients randomisés** : Cette population regroupe tous les patients qui ont été randomisés dans l'étude, c'est-à-dire assignés à un des deux bras de traitement.

Et pour chacune des 2 années de l'étude :

- « **Safety Set** » incluant tous les patients ayant pris au moins une dose du traitement.
- **« Full Analysis set » (FAS)** incluant tous les patients qui ont été randomisés, qui ont pris au moins une dose du traitement et sur lesquels on a au moins récolté une donnée d'efficacité.
- **« Per Protocol Set » (PPS)** qui est une sous-population du FAS. Cet ensemble inclut les patients dont on a récolté un minimum de 14 données d'efficacité valides, durant au moins 50% de la période de pollen et qui ne présentent aucune violation majeure du protocole (médicaments interdit...).

Les patients randomisés dans l'étude souffrent tous d'allergies au pollen du bouleau, et doivent respecter, de plus, un certain nombre de critères d'inclusion et d'exclusion. Afin d'atteindre les objectifs, il a été planifié de randomiser dans l'étude un minimum de 544 patients répartis équitablement dans les 2 bras de traitement (placebo ou IP).

Le tableau 1 présente le nombre de patient présents dans chacune de ces populations, les dates de débuts de recrutements et de dernières visites. Les tableaux 2 et 3 présentent les paramètres principaux qui seront pris en compte dans le modèle statistique.

Tableau 1: Disposition des patients durant la première année

Treatment Group	First Patient IC Date	First Patient Randomization Date	Last Patient Visit 7 Date	Patients Screened n	Patients Randomized n	Analysis Sets Safety n (%)	Analysis Sets FAS n (%)	Analysis Sets PPS n (%)
Total	08NOV2010	03DEC2010	07JUL2011	777	572	570 (99.7)	542 (94.8)	542 (94.8)
- Placebo	08NOV2010	03DEC2010	28JUN2011		288	287 (50.2)	273 (47.7)	273 (47.7)
- IP	09NOV2010	03DEC2010	07JUL2011		284	283 (49.5)	269 (47.0)	269 (47.0)

FAS: Full analysis set, IC: Informed consent, PPS: Per protocol set, Randomized: Randomized in Year 1. Percentages were calculated relative to Randomized in Year 1.

Tableau 2 : Caractéristiques démographiques (catégorielles) des patients randomisés

Description	Category	Treatment Group IP (N=284) n (%)	Placebo (N=288) n (%)	Total (N=572) n (%)
Gender	Male	138 (48.6)	134 (46.5)	272 (47.6)
	Female	146 (51.4)	154 (53.5)	300 (52.4)
	Total	284 (100.0)	288 (100.0)	572 (100.0)
	Missing	0	0	0
Oral Allergy Syndrom	Yes	147 (51.8)	165 (57.3)	312 (54.5)
	No	137 (48.2)	123 (42.7)	260 (45.5)
	Total	284 (100.0)	288 (100.0)	572 (100.0)
	Missing	0	0	0
Asthma	Yes	66 (23.2)	74 (25.7)	140 (24.5)
	No	218 (76.8)	214 (74.3)	432 (75.5)
	Total	284 (100.0)	288 (100.0)	572 (100.0)
	Missing	0	0	0
Asthma Evaluation (GINA)	Intermittent asthma (GINA 1)	49 (74.2)	61 (82.4)	110 (78.6)
	Mild persistent asthma (GINA 2)	17 (25.8)	13 (17.6)	30 (21.4)
	Total	66 (100.0)	74 (100.0)	140 (100.0)
	Missing	218	214	432
Sensitization Status	Mono-sensitised	47 (16.6)	36 (12.5)	83 (14.5)
	Poly-sensitised	236 (83.4)	252 (87.5)	488 (85.5)
	Total	283 (100.0)	288 (100.0)	571 (100.0)
	Missing	1	0	1

Tableau 3 : Caractéristiques démographiques (continues) des patients randomisés

Description	Statistic	Treatment Group IP (N=284)	Placebo (N=288)	Total (N=572)
Age (years) at Visit 1 (Screening)	n	284	288	572
	Mean	38.3	37.1	37.7
	SD	11.17	11.50	11.34
	95% CI for the mean	[37.0, 39.6]	[35.8, 38.4]	[36.8, 38.6]
	Minimum	18	18	18
	P25	28.0	28.0	28.0
	Median	39.0	37.0	38.0
	P75	46.0	44.5	45.0
	Maximum	65	65	65
Height (cm) at Visit 1 (Screening)	n	284	288	572
	Mean	172.7	172.1	172.4
	SD	9.71	9.42	9.56
	95% CI for the mean	[171.5, 173.8]	[171.0, 173.2]	[171.6, 173.2]
	Minimum	150	150	150
	P25	165.0	166.0	165.0
	Median	173.0	171.0	172.0
	P75	180.0	179.0	180.0
	Maximum	197	197	197
Weight (kg) at Visit 1 (Screening)	n	284	288	572
	Mean	76.24	76.25	76.25
	SD	15.760	15.843	15.788
	95% CI for the mean	[74.40, 78.08]	[74.42, 78.09]	[74.95, 77.54]
	Minimum	47.0	46.0	46.0
	P25	65.00	64.00	64.90
	Median	75.10	75.00	75.00
	P75	85.00	85.50	85.00
	Maximum	150.0	132.3	150.0

CI: Confidence interval, P25: 25th percentile, P75: 75th percentile, SD: Standard deviation, 95% CI from a t-test.

Sur les 777 patients examinés dans l'étude, 572 seront finalement randomisés, et 542 seront inclus dans l'analyse d'efficacité. Le nombre de patients du PPS sera surement revu à la baisse car toutes les données ne sont pas encore récoltées.

On constate qu'il y a autant d'hommes que de femmes, répartis équitablement entre les deux bras de traitement. L'âge des patients varie entre 18 et 65 ans, avec une moyenne de 38 ans environ. La majorité des patients (432) ne souffre pas d'asthme, cette pathologie étant un paramètre important dans l'étude. La plupart des patients (488) sont allergiques à d'autres types de pollens que celui du bouleau.

2.3 L'ANALYSE STATISTIQUE

Dans cette partie seront tout d'abord présentées les différentes variables nécessaires à l'analyse statistique, puis les méthodes établies afin d'analyser au mieux les résultats.

2.3.1 Les Variables

La variable principale qui va permettre d'évaluer l'efficacité du traitement est l'**AAdSS**. Cette variable est calculée à partir des données **journalières** suivantes :

- **RSS** « Rhinoconjunctivitis Symptom Scores » : Durant la période de traitement, le patient doit répertorier chaque jour la sévérité des 6 symptômes de la rhino-conjonctivite (éternuements, rhinorrhée, prurit nasal, congestion nasale, prurit oculaire et larmoiement), appelés les RSS. Chacun de ces RSS est compris entre 0 (absence de symptôme) et 3 (symptômes sévères). La somme de ces 6 scores de symptômes, comprise entre 0 et 18, est le **RTSS** « Rhinoconjunctivitis Total Symptom Scores ». La moyenne de ces RTSS sur une durée donnée est l'**ARTSS** « Average Rhinoconjunctivitis Total Symptom Scores ».

- **RMS** « Rescue Medication Score » : Ce score est assigné à différents types de médicaments pris pour lutter contre les symptômes de l'allergie. Les RMS sont compris entre 0 (aucun médicament pris) et 3 (médicament le plus fort) correspondant à différents types de médicaments. La moyenne de ces RMS sur une durée donnée est l'**ARMS** « Average Rescue Medication Score ». Si un patient prend plusieurs catégories de médicaments le même jour, le RMS sera égal au score le plus élevé.

L'**AAdSS** est la moyenne des **AdSS** « Adjusted Symptom Score », calculée de la façon suivante pour chaque patient (on indice ici chacune des données journalières, par exemple $RTSS_d$ correspond au RTSS du jour numéro *d*) :

- Le premier jour, $AdSS_1=RTSS_1$.
- Si le patient ne prend pas de médicament le jour *d* ni le jour *d-1* (ie. RMS_d et RMS_{d-1} sont nuls), $AdSS_d=RTSS_d$.
- Si le patient prend un médicament le jour *d* alors,
  $AdSS_d = \max (RTSS_d, AdSS_{d-1})$,
  $AdSS_{d+1} = \max (RTSS_{d+1}, AdSS_d)$.

Dans le cas de données manquantes :

- Si $RTSS_d$ est manquant, $AdSS_d$ est manquant.
- Si RMS_d est manquant, alors $AdSS_d = RTSS_d$.
- Si le patient a pris des médicaments le jour *d* (ie. RMS_d n'est pas manquant et vaut 1, 2 ou 3) et $RTSS_{d-1}$ est manquant, alors $AdSS_d = RTSS_d$.
- Toutes les données d'une journée sont considérées comme manquantes si un ou plusieurs RSS (éternuements, rhinorrhée, prurit nasal, congestion nasale, prurit oculaire ou larmoiement) sont manquants, et /ou si le patient n'a pas pris le traitement ce jour là.

Pour chaque patient, l'AAdSS sera calculé comme étant la moyenne des AdSS journaliers non manquants, durant la période d'évaluation. Les AdSS et AAdSS sont donc compris entre 0 et 18 tout comme les RTSS.

D'autres variables d'efficacité et de tolérance sont étudiées en tant qu'objectifs secondaires, mais nous nous concentrerons uniquement sur l'objectif principal.

2.3.2 Le modèle statistique

L'AAdSS sera analysé à l'aide d'une analyse de la covariance (ANCOVA) sur les patients appartenant au FAS, avec comme covariables :

- Le traitement (IP ou placebo),
- Le statut d'OAS (avec ou sans OAS),
- Le groupement des centres,
- Le sexe (homme ou femme),
- La présence ou non d'asthme (asthmatique ou non),
- La sensibilité au pollen distinguant les patients allergiques uniquement au pollen du bouleau, de l'aulne et du noisetier (mono-sensibilité) des patients présentant d'autres allergies (poly-sensibilité).

Et comme facteur l'âge des patients.

Des analyses exploratoires seront aussi effectuées lors de la deuxième année sur les patients appartenant au FAS. Les modèles statistiques ajustés seront les mêmes que précédemment avec les interactions suivantes en plus :

- Modèle A : interaction entre le traitement et le regroupement du centre du patient.
- Modèle B : interaction entre le traitement et l'Asthme.
- Modèle C : interaction entre le traitement et la sensibilité au pollen.
- Modèle D : interaction entre le traitement et OAS.

Les interactions seront considérées comme significatives si la p-valeur est inférieure au risque $\alpha = 0.10$.

Les hypothèses de l'ANCOVA d'homoscédasticité et de normalité des résidus seront vérifiées à l'aide de tests statistiques : tests de Shapiro-Wilk et Kolmogorov-Smirnov pour la normalité et test de Levene pour l'homoscédasticité. Les variables sont considérées indépendantes les une des autres.

2.3.3 Justification du choix du modèle

Dans cette partie, je vais expliquer pourquoi le choix de la variable principale (et donc du modèle statistique qui en découle) s'est porté sur une moyenne (AAdSS) et non par exemple sur un modèle à mesures répétées qui prendrait en compte l'évolution dans le temps des symptômes.

Souvent, dans des études cliniques longitudinales, l'efficacité du traitement se mesure à l'aide d'une variable qui évolue de façon monotone dans le temps. Cette variable est mesurée lors de plusieurs

visites (pas plus d'une dizaine en général), par exemple la taille d'une tumeur, ou le taux de cholestérol. Dans ce cas, il est donc naturel d'étudier cette variable en prenant en compte son évolution dans le temps, par exemple à l'aide de la différence entre la valeur récoltée lors de la dernière visite et la valeur au début de l'étude.

Cependant, dans notre étude, les scores des symptômes (RSS) oscillent grandement au cours du temps car les patients ne sont pas exposés de façon continue au pollen. Les données sur le taux de pollen dans l'air sont collectées et seront prises en compte dans l'analyse. Ces données ne sont pas encore disponibles pour l'année en cours. De plus, les RSS sont recueillis chaque jour durant plus de 2 mois, ce qui augmente encore ces oscillations. Les figures 4 et 5 ci-dessous représentent les AdSS de deux patients (pris au hasard) par jour, durant toute la durée du traitement.

Figure 4 : AdSS par patient (1)

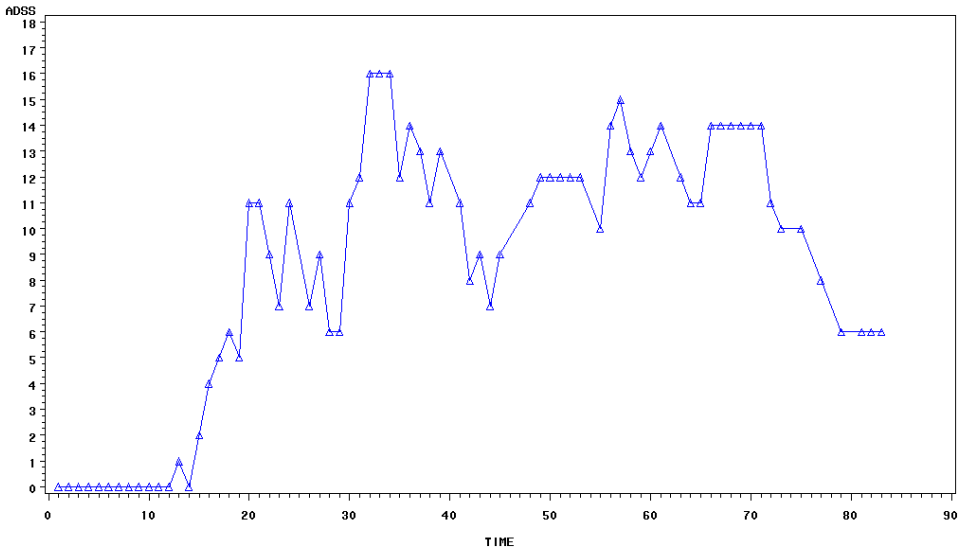
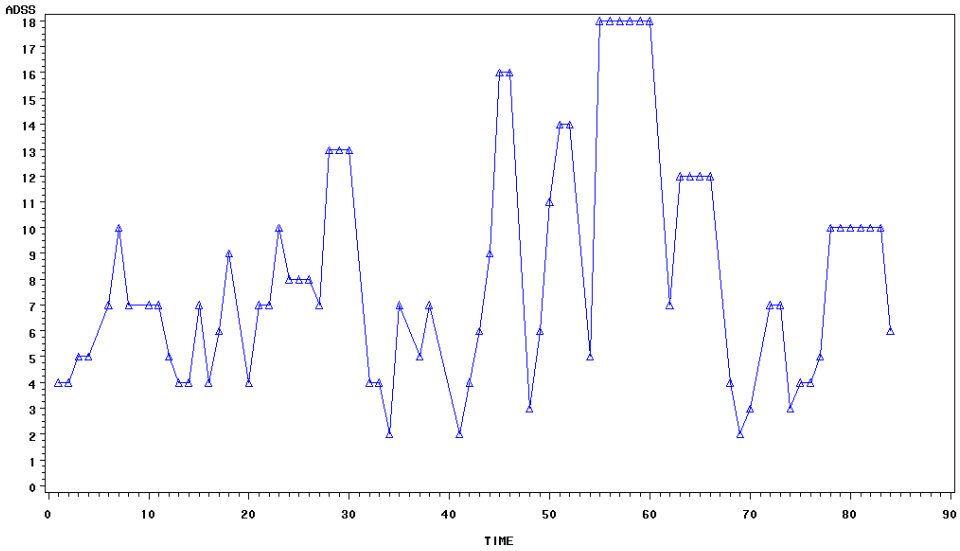


Figure 5 : AdSS par patient (2)



On constate donc que d'une journée à l'autre, l'AdSS peut varier énormément.

Il a donc été décidé de travailler sur une moyenne par patient (AAdSS) sur différentes périodes (l'entière saison de pollen et les 2 semaines avec le taux de pollen le plus fort), ceci afin de ne garder qu'une seule valeur par patient, valeur représentant l'efficacité du traitement sur toute cette période.

2.3.4 Présentation des résultats

Les résultats présentés ici ne sont pas représentatifs de l'efficacité réelle de l'IP car, comme mentionné précédemment, l'étude est en aveugle. De plus certaines variables du modèle ne sont pas encore disponibles, comme le regroupement des centres des patients ainsi que les dates de début et de fin de saison de pollen. Les AAdSS utilisés ici sont donc calculés sur toute la durée du traitement et non pas sur la saison de pollen. Cependant, cela nous donne un aperçu du travail effectué et de la façon dont les résultats seront interprétés.

Le tableau suivant décrit les résultats de l'AAdSS obtenus pour tous les patients faisant partie du FAS :

Tableau 4 : Moyenne des AdSS par groupe de traitement

---------- Treatment (Character)=IP ----------

The MEANS Procedure

Analysis Variable : SCAADSS Average Adjusted Symptom Score

N	Mean	Std Dev	Minimum	Maximum
269	3.9817593	2.6350826	0	12.2545455

---------- Treatment (Character)=Placebo ----------

Analysis Variable : SCAADSS Average Adjusted Symptom Score

N	Mean	Std Dev	Minimum	Maximum
273	3.8261209	2.6282468	0.0238095	13.2087912

Les résultats sont donc similaires dans les deux groupes. Cependant, une fois les vrais traitements pris en compte, on espère détecter une différence significative entre les deux groupes en faveur de l'IP.

Le modèle statistique a été programmé avec la PROC GLM de SAS de la façon suivante :

```
proc glm data=sc;
  class KVTRTC KVOASC KVCOUNTC KVSEXC KVASTHC KVSENSC;
  model SCAAdSS = KVTRTC KVOASC KVCOUNTC KVAGE KVSEXC KVASTHC KVSENSC
                / clparm;
  lsmeans KVTRTC / stderr pdiff cl;
run;quit;
```

Les données concernant les groupements de centres dans lesquels les patients ont été recrutés n'étant pas encore disponibles, on utilisera pour l'instant le pays du patient.

Le tableau 5 ci-dessous suivants présente les résultats, toujours pour les patients faisant partie du FAS, sur toute la saison de pollen :

Tableau 5 : Résultats de l'ANCOVA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	365.146403	22.821650	3.54	<.0001
Error	524	3377.605615	6.445812		
Corrected Total	540	3742.752018			

R-Square	Coeff Var	Root MSE	SCAADSS Mean
0.097561	65.02544	2.538860	3.904411

Source	DF	Type I SS	Mean Square	F Value	Pr > F
KVTRTC	1	3.3778157	3.3778157	0.52	0.4694
KVOASC	1	0.1640670	0.1640670	0.03	0.8733
KVCOUNTC	10	305.8265522	30.5826552	4.74	<.0001
KVAGE	1	5.1708294	5.1708294	0.80	0.3708
KVSEXC	1	16.1923801	16.1923801	2.51	0.1136
KVASTHC	1	32.7728904	32.7728904	5.08	0.0246
KVSENSC	1	1.6418681	1.6418681	0.25	0.6140

Source	DF	Type III SS	Mean Square	F Value	Pr > F
KVTRTC	1	2.1031117	2.1031117	0.33	0.5681
KVOASC	1	3.3296223	3.3296223	0.52	0.4726
KVCOUNTC	10	301.9773983	30.1977398	4.68	<.0001
KVAGE	1	5.0594897	5.0594897	0.78	0.3760
KVSEXC	1	19.6552490	19.6552490	3.05	0.0814
KVASTHC	1	33.8672279	33.8672279	5.25	0.0223
KVSENSC	1	1.6418681	1.6418681	0.25	0.6140

Least Squares Means

KVTRTC	SCAADSS LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1=LSMean2 Pr > |t|
IP	4.37365047	0.25532424	<.0001	0.5681
Placebo	4.24719197	0.25239579	<.0001	

KVTRTC	SCAADSS LSMEAN	95% Confidence Limits	
IP	4.373650	3.872066	4.875235
Placebo	4.247192	3.751360	4.743024

Least Squares Means for Effect KVTRTC

i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.126459	-0.308460	0.561377

On utilisera les résultats basés sur les sommes des carrés de Type I, qui prend en compte l'ordre dans lequel les variables explicatives apparaissent. La p-valeur associée au test de significativité du traitement vaut 0.4694. L'effet du traitement n'est donc pas significatif. Cependant, on espérera qu'une fois l'aveugle levé, le traitement aura une influence significative. Le risque α est fixé à 0.05.

Les moyennes des moindres carrés des AAdSS sont aussi similaires d'un groupe de traitement à l'autre. On pourra s'attendre néanmoins à une moyenne plus élevée dans le groupe de traitement de l'IP. Un intervalle de confiance de la différence entre ces moyennes est aussi fournit. Cet intervalle de confiance est obtenu à l'aide du test de Student.

On remarque notamment que le R^2 est très faible, ce qui est normal vu qu'il reflète l'ajustement du modèle, et que pour l'instant, certains paramètres importants du modèle ne sont pas disponible.

L'hypothèse d'homoscédasticité est vérifiée à l'aide du test de Levene produit avec la PROC GLM ci-dessous :

```
proc glm data=sc;
  class KVTRTC;
  model SCAADSS = KVTRTC / clparm;
  means KVTRTC / hovtest = LEVENE;
run;quit;
```

On obtient le résultat suivant:

Tableau 6 : Résultat du test d'homoscédasticité

Levene's Test for Homogeneity of SCAADSS Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
KVTRTC	1	0.1705	0.1705	0.00	0.9683
Error	540	58098.2	107.6		

La p-valeur vaut ici 0.9683. On ne rejette donc pas l'hypothèse nulle de ce test qui est H_0 : les variances sont égales. Toujours avec un seuil α fixé à 0.05. L'homoscédasticité est donc vérifiée.

L'hypothèse de normalité des résidus est vérifiée à l'aide des tests de Shapiro-Wilk et Kolmogorov-Smirnov, par traitement. Le code SAS utilisé est le suivant (la variable *resid* représentant les résidus):

```
proc univariate data=sc normaltest;
  var resid;
  by KVTRTC;
run; quit;
```

On obtient dans ce cas :

Tableau 7 : Résultats des tests de normalité des résidus par traitement

---------- Treatment (Character)=Placebo ----------

The UNIVARIATE Procedure
Variable: RESID

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.944647	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.088184	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.513488	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.510803	Pr > A-Sq	<0.0050

---------- Treatment (Character)=IP ----------

The UNIVARIATE Procedure
Variable: RESID

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.956657	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.086053	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.404719	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.689468	Pr > A-Sq	<0.0050

Les p-valeurs de ces deux tests pour chacun des traitements sont toutes inférieures au seuil α fixé à 0.05. On rejette donc les hypothèses nulles qui sont pour ces deux tests H_0 : l'échantillon est issu d'une loi normalement distribuée. La normalité des résidus n'est donc pas vérifiée.

On arrive à la même conclusion en regardant les "QQ plots" des résidus pour chaque groupe de traitement. Ils permettent de comparer les valeurs ordonnées d'une variable avec les quantiles d'une distribution théorique spécifiée, dans notre cas une loi normale.

Figure 6 : QQ Plot des résidus – IP

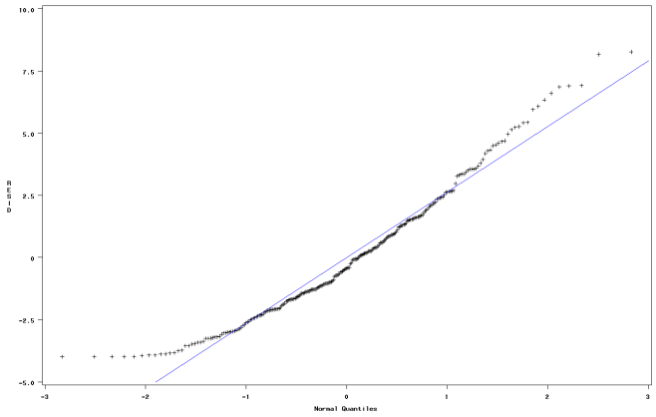
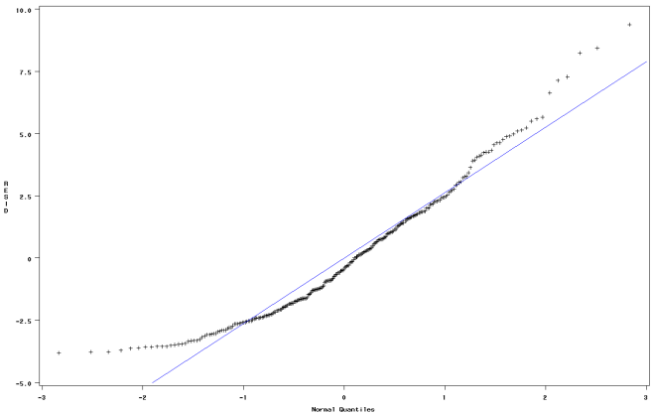


Figure 7 : QQ Plot des résidus - Placebo



On peut voir qu'il y a une tendance qui se dégage. Il reste donc de l'information dans les résidus. Cela implique que dans ce cas les tests de significativités risquent d'être biaisés. Cependant on remarque que le biais est principalement situé aux valeurs extrêmes. Il pourrait être dû au fait que les AdSS récoltés avant la saison de pollen ont été pris en compte dans le calcul de l'AAdSS. En effet ces AdSS sont quasiment tous nuls avant la saison de pollen. Si la normalité des résidus n'est toujours pas vérifiée une fois toutes les données prisent en compte, un autre modèle sera surement ajusté en fonction des résultats.

Plusieurs autres analyses secondaires sont planifiées pour la première année, cependant elles ne seront pas présentées ici. En effet elles sont basées sur le même modèle statistique et ne présente pas grand intérêt pour l'instant puisque l'aveugle n'a pas encore été levé.

3 LES DONNEES MANQUANTES DANS LES ESSAIS CLINIQUES

En statistique, et particulièrement lors d'études cliniques, il est fréquent que le jeu de données soit incomplet. On entend par là que toutes les observations planifiées n'ont pas été récoltées.

Dans cette partie nous allons commencer par donner un aperçu de la méthodologie sur ce sujet, les formulations et les différentes méthodes qui ont été élaborées, et en particulier la méthode dite d'imputation multiple. Ensuite nous appliquerons cette méthode à l'étude en allergie précédemment mentionnée.

3.1 METHODOLOGIES ET NOTATIONS

La terminologie utilisée ici est celle de Rubin (1976) et Little et Rubin (2002). Cette terminologie permet de formaliser le mécanisme des données manquantes, et ainsi de pouvoir appliquer des méthodes adaptées.

3.1.1 Notations

Supposons que pour chaque unité indépendante i = 1, ..., N dans l'étude, il est planifié de collecter un ensemble d'observations Y_{ij} (j = 1, ..., $n_{(i)}$). Dans une étude longitudinale, i indique alors le patient et j la j-ème mesure.

On regroupe les résultats dans un vecteur $Y_i = (Y_{i1}, \ldots, Y_{in(i)})'$ représentant les données complètes.

On définie $R_{ij} = \begin{cases} 1 \text{ si } Y_{ij} \text{ est observé} \\ 0 \text{ sinon} \end{cases}$

Cet indicateur de donnée manquante R_{ij} est organisé en un vecteur R_i de la même façon que Y_i.

On partitionne Y_i en deux sous-vecteurs Y_i^{obs} et Y_i^{mis} tels que $Y_i = (Y_i^{obs}, Y_i^{mis})$.

Y_i^{obs} représentant les données observées, i.e. contenant les Y_{ij} tels que $R_{ij} = 1$ et Y_i^{mis} les données manquantes i.e. contenant les Y_{ij} tel que $R_{ij} = 0$.

On étudie le mécanisme des données manquantes à l'aide de la densité jointe $f(y_i, r_i \mid X_i, W_i, \theta, \psi)$, où

X_i est la matrice de design des données complètes,
W_i la matrice de design des données manquantes,
θ et ψ les vecteurs de paramètres correspondant respectivement à X_i et W_i.

Plusieurs décompositions de cette densité ont été élaborées afin d'avoir différente approches possibles. Les trois principales sont présentées ci-dessous:

La première est appelée **« selection model »** et correspond à la factorisation suivante :

$$\begin{aligned} f(y_i, r_i \mid X_i, W_i, \theta, \psi) &= f(y_i \mid X_i, \theta)\, f(r_i \mid y_i, W_i, \psi) \\ &= f(y_i \mid X_i, \theta)\, f(r_i \mid y_i^{obs}, y_i^{mis}, W_i, \psi) \end{aligned} \qquad \textbf{(1)}$$

où le premier facteur est la densité marginale des données complètes et le deuxième la densité des données manquantes conditionnellement aux y_i.

Le deuxième modèle possible est appelé **« pattern-mixture models »** :

$$f(y_i, r_i \mid X_i, W_i, \theta, \psi) = f(y_i \mid r_i, X_i, \theta)\, f(r_i \mid W_i, \psi)$$

Et le dernier, appelé **« shared-parameter models »,** est défini de la façon suivante :

$$f(y_i, r_i \mid X_i, W_i, \theta, \psi, b_i) = f(y_i \mid r_i, X_i, \theta, b_i)\, f(r_i \mid W_i, \psi, b_i)$$

où b_i est un vecteur (éventuellement aléatoire) dont les composantes sont partagées entre les deux facteurs de la distribution jointe.

Chacun de ces trois modèles ont des interprétations différentes. Je me concentrerai uniquement sur le « selection model », les autres étant cités à titre informatif.

3.1.2 Types de données manquantes

Rubin (1976) a introduit une classification statistique des données manquantes, permettant de différencier trois types de données manquantes :

- **« Missing Completely At Random » (MCAR)** : La probabilité d'avoir une valeur manquante est indépendante des données observées Y_i^{obs} et des données manquantes Y_i^{mis}. On a donc l'égalité suivante : $f(r_i \mid y_i^{obs}, y_i^{mis}, W_i, \psi) = f(r_i \mid W_i, \psi)$
  L'équation **(1)** peut alors être simplifiée et devient :

  $$f(y_i, r_i \mid X_i, W_i, \theta, \psi) = f(y_i \mid X_i, \theta)\, f(r_i \mid W_i, \psi)$$

  Ce qui implique l'indépendance entre les deux facteurs.

- **« Missing At Random » (MAR)** : La probabilité d'avoir une valeur manquante, conditionnellement aux données observées, est indépendante des données manquantes Y_i^{mis}. On a donc l'égalité suivante : $f(r_i \mid y_i^{obs}, y_i^{mis}, W_i, \psi) = f(r_i \mid y_i^{obs}, W_i, \psi)$

- **« Missing Non At Random » (MNAR) :** La probabilité d'avoir une valeur manquante, conditionnellement aux données observées, dépend de la valeur (inconnue) des données manquantes Y_i^{mis}. Dans ce cas on ne peut pas simplifier $f(r_i \mid y_i^{obs}, y_i^{mis}, W_i, \psi)$.
  Le mécanisme des données manquantes ne doit alors surtout pas être ignoré.

Dans la suite nous n'indiceront plus par *i* afin d'alléger les notations.

Remarque : Dans le cas de données longitudinales, la classification est modifiée (Schafer) et prend en compte les variables explicatives du modèle (notées X_k, k = 1, …, p). Dans ce cas le mécanisme est MCAR s'il ne dépend ni des X_k ni de Y ; MAR s'il dépend des X_k et des Y^{obs} ; et MNAR s'il dépend en plus des Y^{mis}.

Les données MNAR sont très difficiles à étudier et il n'existe que peu de méthodes fiables. En général, lors d'un essai clinique, les données manquantes sont de type MAR. En effet, la probabilité que la variable d'intérêt soit manquante dépend souvent d'autres données observées, comme par exemple le traitement, l'âge du patient ou le centre dans lequel il a été assigné. Cependant, les investigateurs ne sont généralement pas en mesure de rejeter l'hypothèse de MNAR.

Il faut aussi distinguer deux types autres de données manquantes : Les données monotones et non-monotones :

- Lors d'un essai clinique, il arrive souvent que des patients quittent l'étude prématurément, entraînant ainsi des données manquantes. Certaines méthodes ont donc été élaborées afin de traiter ces données. Les patients qui quittent l'étude sont listés dans une partie à part du CRF. Les motifs sont alors spécifiés : effet néfaste, non efficacité de l'IP, maladie non liée à l'étude, patient non coopérant, violation du protocole, etc... Ces données manquantes sont dites monotones, c'est-à-dire que le vecteur R_i s'écrit sous la forme (1, …, 1, 0, …, 0). Par exemple, les données manquantes sont réparties de la façon suivante (les * représentant les données manquantes) :

Figure 8 : Disposition monotone

Patient	Visite 1	Visite 2	Visite 3	Visite 4	Visite 5
1	3	7	9	8	7
2	2	2	6	*	*
3	7	1	*	*	*
4	3	3	*	*	*
5	0	*	*	*	*

- Cependant, il est possible qu'un patient ne se présente pas à une visite pour une raison quelconque, mais reste dans l'étude. Il se représentera donc aux visites suivantes, comme le montre l'exemple ci-dessous :

Figure 9 : Disposition non-monotone

Patient	Visite 1	Visite 2	Visite 3	Visite 4	Visite 5
1	3	7	5	8	7
2	2	*	6	8	9
3	1	5	*	7	*
4	*	*	5	7	*
5	2	4	*	5	6

Les données manquantes sont dites alors non-monotones, et elles sont plus difficiles à prendre en compte.

3.2 METHODES D'ANALYSES

Avant toute chose, il est nécessaire d'étudier le nombre de données manquantes et non manquantes, d'observer leurs proportions, les variables et les individus concernés. Les graphiques sont un des outils les plus importants du statisticien pour essayer de comprendre le mécanisme de ces données manquantes.

Il existe de nombreuses méthodes, plus ou moins efficaces. Le tableau 8 présente les principales en fonction du type de données manquantes:

Tableau 8 : Méthodes de traitement des données manquantes

MCAR	MAR	MNAR
Cas complet	Imputation multiple	?
Imputation simple	Algorithme EM	« Sensitivity analysis »
…	« Weighted Generalized Estimating Equations » (WGEE)	
	« Mixed-effects Model for Repeated Measures » (MMRM)	

Certaines de ces méthodes sont beaucoup discutées, et les avis divergent souvent quant aux hypothèses sur le type de données manquantes qu'elles nécessitent. Je me suis concentré sur l'imputation multiple qui est une des méthodes les plus utilisées.

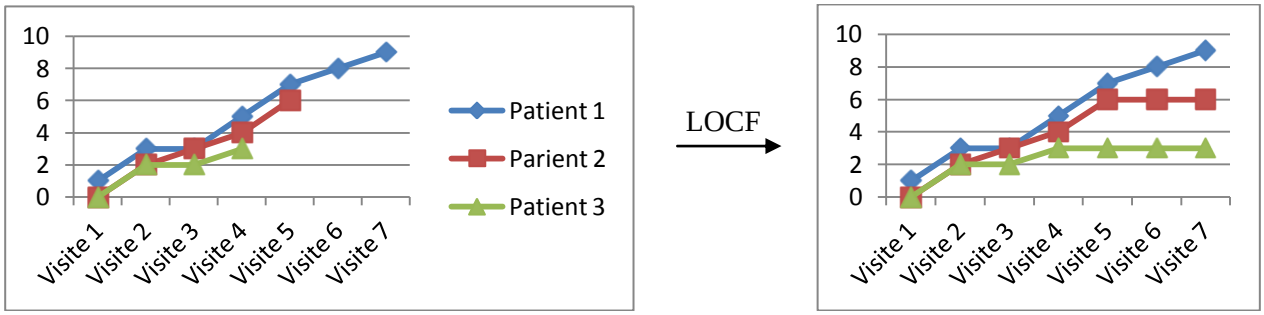
Concernant les données manquantes MNAR, une variété de modèles a été proposée pour les analyser durant la dernière décennie. Mais le problème fondamental de ces méthodes est qu'elles dépendent d'hypothèses non vérifiables. Ces modèles sont appelés « sensitivity analysis ».

3.2.1 Méthode sous l'hypothèse MCAR

L'une des méthodes les plus utilisées et qui est implémenté par défaut dans les logiciels est celle du cas complet. Elle consiste simplement à ne pas prendre en compte les individus qui ont des données manquantes. Par exemple, dans un essai clinique, si une donnée sur un patient, nécessaire à l'analyse statistique, est manquante, on retire entièrement le patient de l'étude. Cette méthode implique souvent une perte de données considérable.

Les méthodes d'imputation simple sont très nombreuses. Elles consistent à substituer à la valeur manquante une valeur choisie de façon définitive. La méthode la plus utilisée dans les essais cliniques est « Last Observation Caried Forward » (LOCF). Cette méthode est souvent utilisée sur des données monotones lors d'études longitudinales. Elle consiste à imputer toutes les données manquantes d'un même individu par la dernière observation qui a été récoltée, comme illustré ci-dessous :

Figure 10 : Exemple LOCF



Cependant, cette hypothèse sous-entendant que la dernière valeur récoltée sur un patient ayant quitté l'étude resterait constante, ne peut être testée et est potentiellement irréelle. Même l'hypothèse forte de MCAR ne garantit pas qu'une analyse LOCF soit valide. De plus, cette méthode résulte systématiquement d'une sous-estimation de l'écart type.

Une autre méthode d'imputation simple fortement plébiscité par les autorités de santé est le « worst case ». Elle consiste à imputer des valeurs qui défavorisent le résultat attendu.

L'inconvénient est que ces méthodes d'imputation simple ne tiennent pas compte de l'incertitude supplémentaire liée au fait que la donnée manquante était manquante avant imputation. Les données imputées sont donc considérées de la même façon que les autres dans l'analyse. D'où l'intérêt de l'imputation multiple.

3.2.2 Imputation multiple

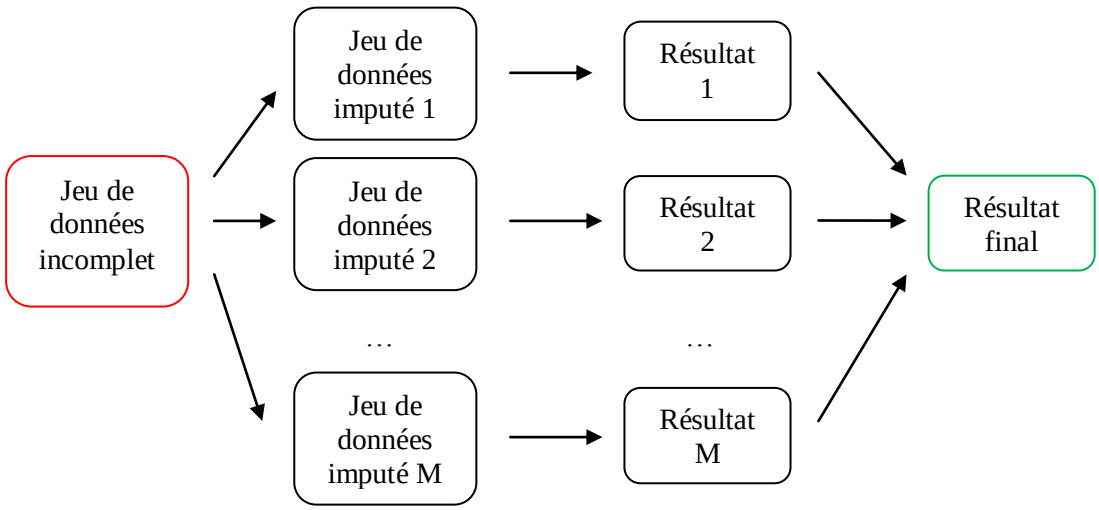
Introduite par Rubin (1978), l'imputation multiple est devenue une approche importante pour traiter les données manquantes. Cette méthode est en général à utiliser sous l'hypothèse MAR. Les logiciels qui l'ont implémentée (SAS en particulier) font cette hypothèse. Cependant dans certaines situations il est possible de l'utiliser sous l'hypothèse MNAR.

L'imputation multiple se déroule en trois étapes :

- Les données manquantes sont remplacées par un ensemble de M valeurs possibles (M $\in$ $\mathbf{N}^*$). Les imputations produisent alors M jeux de données « complets ».
- Chacun de ces jeux de données est alors analysé indépendamment, à l'aide de méthodes standards, que l'on aurait appliquées si le jeu de données avait été complet.
- Les résultats des M analyses sont combinés en une seule inférence.

La figure 9 ci-dessous illustre ces trois étapes :

Figure 11 : Schéma d'une imputation multiple



Le logiciel SAS permet d'effectuer des imputations multiples à l'aide de procédures spécifiques pour chacune des étapes :

- La PROC MI qui permet de générer M jeux de données imputées.
- Les procédures standards (GLM, LOGISTIC, etc) qui permettent d'analyser les résultats de ces M jeux de données indépendamment.
- La PROC MIANALYZE permettant de fusionner les résultats obtenus.

3.2.2.1 Générer M jeux de données

Les M jeux de données sont générés à partir de la distribution à posteriori des valeurs de Y^{mis}. Le modèle utilisé doit prendre en compte toutes les variables qui pourraient avoir un effet sur les valeurs manquantes. Ces méthodes d'imputations dépendent naturellement de la nature des variables à imputer. Le tableau suivant présente les principales méthodes d'imputation implémenté dans la PROC MI de SAS :

Tableau 9 : Méthodes d'imputations possibles

Disposition des données manquantes	Type de la variable à imputer	Méthodes possibles (option dans la PROC MI)
Monotone	Continue	Paramétrique: • **Méthode de régression** (MONOTONE REG) • « Predictive mean matching » (MONOTONE REGPMM) Non-paramétrique: • Score de propension (MONOTONE PROPENSITY)
Monotone	Catégorielle	Régression logistique (MONOTONE LOGISTIC)
Non-monotone	Continue	« Markov Chain Monte Carlo » (MCMC) • Imputation totale • **Imputation partielle** *
Non-monotone	Catégorielle	Pas de méthode disponible

***Imputation des données non-monotone avec MCMC afin d'obtenir une disposition des données manquantes monotones. Puis imputer le reste à l'aide de méthodes monotones.**

Nous nous intéresseront tout particulièrement à la méthode de régression et à la méthode d'imputation partielle basée sur l'algorithme.

La méthode de régression permet donc d'imputer des variables continues contenant des valeurs manquantes monotones. L'idée proposée par Rubin (1987) est la suivante :

Pour chaque variable continue Y_j contenant des données manquantes, le modèle de régression

$$Y_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_j$$

est ajusté, utilisant uniquement les observations avec des valeurs non manquantes pour Y_j et ses variables explicatives $X_1, \ldots, X_k$.

A l'aide de ce modèle, on estime les paramètres $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ et la matrice de covariance $\hat{\sigma}_j^2 V_j$ de façon standard.

A chaque itération m = 1, …, M, on calcule les paramètres $\beta^{(m)} = (\beta_0^{(m)}, \beta_1^{(m)}, \ldots, \beta_k^{(m)})$ et $\sigma_j^{2(m)}$ qui sont obtenus à partir de la distribution à posteriori des paramètres. Ce calcul ne sera pas détaillé ici, mais dépend des paramètres $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ et $\hat{\sigma}_j^2 V_j$ et prend aussi en compte le nombre d'observations non manquantes de Y_j.

Les données manquantes Y_{ij}^{mis} sont alors remplacées par

$$\beta_0^{(m)} + \beta_1^{(m)} x_{1i} + \beta_2^{(m)} x_{2i} + \ldots + \beta_k^{(m)} x_{ki} + z_i \sigma_j^{(m)}$$

Où les x_{1i}, x_{2i}, … x_{ki} sont les valeurs observées des variables explicatives, et z_i un bruit normal aléatoire.

Cette méthode suppose que les variables inclues dans le modèle d'imputation sont distribuées de façon normale. Cependant, si l'hypothèse de normalité n'est pas vérifiée, on peut se tourner vers la méthode « Predictive mean matching ». En effet, cette méthode pourtant assez similaire, assure que les imputations restent plausibles, même si l'hypothèse de normalité n'est pas vérifiée.

L'algorithme MCMC, qui est basé sur les chaines de Markov ne sera pas expliqué en détails. Cependant l'idée générale est présentée ci-dessous :

Une chaine de Markov est une séquence de variables aléatoires dans laquelle la distribution de chaque élément dépend uniquement de la valeur de la précédente, de telle sorte que

$$Pr\,(X_{t+1} \mid X_t, X_{t-1}, \ldots, X_1) = Pr\,(X_{t+1} \mid X_t) \xrightarrow[t \to \infty]{} \pi(X)$$

Où les X_i, i Є **N**, sont des variables aléatoires.
L'idée est de construire une chaine assez longue afin que la distribution des éléments se stabilise à la distribution d'intérêt, noté ici π.

La PROC MI simule une chaine de Markov ($Y^{mis\,(1)}$, $\theta^{(1)}$) ($Y^{mis\,(2)}$, $\theta^{(2)}$) … qui converge vers une distribution $P\,(Y^{mis}, \theta \mid Y^{obs})$ où θ représente les paramètres du modèle d'imputation.

D'autres options utiles de la PROC MI seront décrites dans la présentation des résultats.

3.2.2.2 Combiner les résultats

Supposons que l'on souhaite estimer k paramètres $\beta_1, \ldots, \beta_k$ représentés par le vecteur $\beta = (\beta_1, \ldots, \beta_k)'$.

Après avoir imputé les M jeux de données on obtient M vecteurs β^m que l'on estime avec les méthodes usuelles (m = 1, …, M). On obtient alors $\hat{\beta}^m$ et V^m respectivement l'estimateur de β et sa matrice de covariance pour le m-ième jeu de données imputées.

L'estimateur final de β est simplement la moyenne de ces estimateurs :

$$\hat{\beta}^* = \frac{1}{m} \sum_{m=1}^{M} \hat{\beta}^m$$

Afin de mesurer la précision des $\hat{\beta}^*$ on utilise les matrices de covariances intra-imputation notées V^{intra} et inter-imputation notées V^{inter} définies de façon intuitive (Rubin) :

$$V^{intra} = \frac{1}{m} \sum_{m=1}^{M} V^m$$

$$V^{inter} = \frac{1}{m-1} \sum_{m=1}^{M} (\hat{\beta}^m - \hat{\beta}^*)(\hat{\beta}^m - \hat{\beta}^*)'$$

Ces variances reflètent l'incertitude des imputations.

La matrice de covariance de $\hat{\beta}^*$ est donnée par :

$$V = V^{intra} + \left(\frac{M+1}{M}\right) V^{inter}$$

Dans la plupart des cas un petit nombre M d'imputations suffit à obtenir d'excellents résultats.
Rubin (1987) a montré que l'efficacité d'une estimation basée sur M imputations est approximativement

$$RE := \left(1 + \frac{\gamma}{M}\right)^{-1}$$

où γ est la fraction d'informations manquantes pour la quantité estimée. γ quantifie à quel point l'estimation aurait été plus précise si aucune donnée n'avait été manquante. On notera cette quantité RE pour « Relative Efficiency ».

L'imputation multiple tient donc compte de l'incertitude supplémentaire liée aux manquantes. De plus, elle préserve les aspects de la distribution des données (moyenne, tendance, variation inter et intra sujets). Un point fort de cette méthode est sa flexibilité. En effet on peut utiliser différents modèles pour imputer les données et pour analyser ces données imputées.

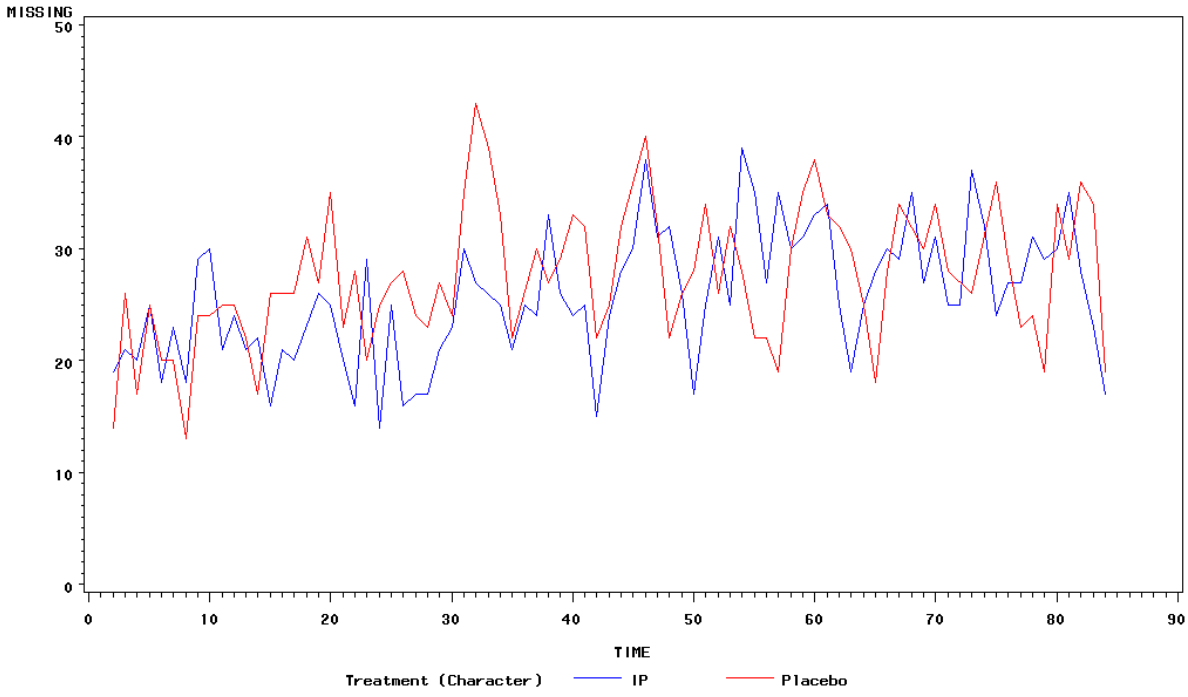
3.3 APPLICATION SUR UNE ETUDE EN ALLERGIE

Nous allons maintenant étudier les données manquantes de l'étude présentée précédemment.

3.3.1 Description des données manquantes

Dans cette étude on évalue l'efficacité à partir des RSS et des RMS récoltés chaque jour durant plus de deux mois. Ces données sont rentrées dans la base à l'aide d'un boitier électronique, par le patient lui-même, sans qu'il ait besoin de se déplacer à son centre. Cependant cela implique la présence de nombreuses données manquantes non-monotones. En effet on peut imaginer que sur une durée aussi longue, il arrive que le patient oublie ou ne puisse tout simplement pas remplir ses données. Le graphique ci-dessous nous montre le nombre d'AdSS manquants en fonction des deux groupes de traitements, par jour :

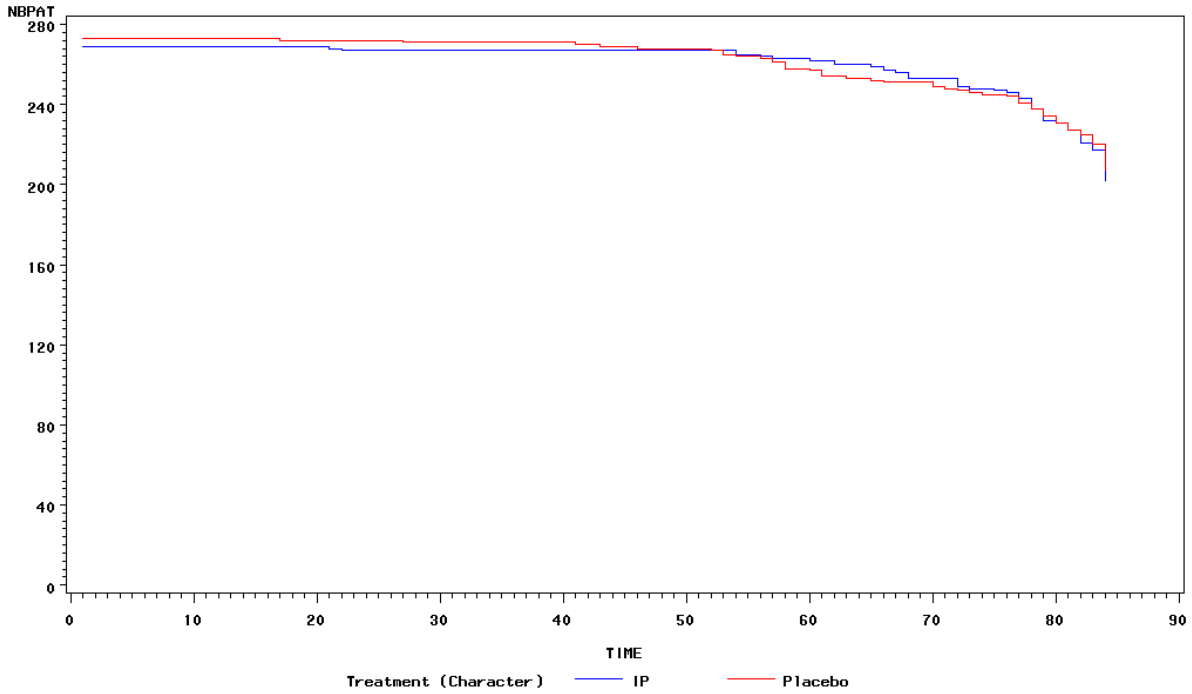
Figure 12 : Représentation des données manquantes



En moyenne, on remarque qu'environ 50 patients (25 par groupe) ne rentrent pas du tout ou pas correctement leurs données chaque jour. On remarque aussi une légère tendance à la hausse dû aux patients qui ont quitté l'étude.

Le graphique suivant montre le nombre de patients encore présents dans l'étude :

Figure 13 : Patients encore présents dans l'étude



On constate que les patients commencent à quitter l'étude à partir du cinquantième jour environ, qui correspond au début de la pire saison de pollen.

Ce qui nous intéresse est de savoir à quoi peut être lié le fait que les données soient manquantes, dans le but d'essayer de déterminer le type de données manquantes (MCAR, MAR ou MNAR) auxquelles nous devrons faire face.

Dans le cas des données non-monotones, on peut supposer que cela est principalement dû à des raisons personnelles, non liée à l'étude. En effet, le premier graphique nous montre que le nombre de données manquantes est assez constant. Cependant, certains paramètres liés à l'étude peuvent avoir une influence, comme par exemple l'âge des patients ou le traitement. Les figures 10 et 11 ne permettent pas de distinguer de réelles différences entre les deux groupes, du fait que les vrais traitements ne sont pas encore dans la base de données. Dans le doute nous rejetterons donc l'hypothèse MCAR et nous nous tournerons plutôt vers des données manquantes de type MAR. Nous écarterons aussi l'hypothèse MNAR. En effet, le fait qu'un patient ne rentre pas ses données ne dépendrait apparemment pas de la valeur de ces données manquantes (inconnue). Cependant il sera impossible de vérifier cette supposition.

En ce qui concerne les données manquantes monotones, on supposera de la même façon que le traitement peut influer sur le fait que les patients quittent l'étude. En effet, il est probable que des patients étant dans le groupe placebo quittent l'étude à cause du manque d'efficacité du traitement. Cela impliquerait donc aussi l'hypothèse MAR.

3.3.2 Imputation des données

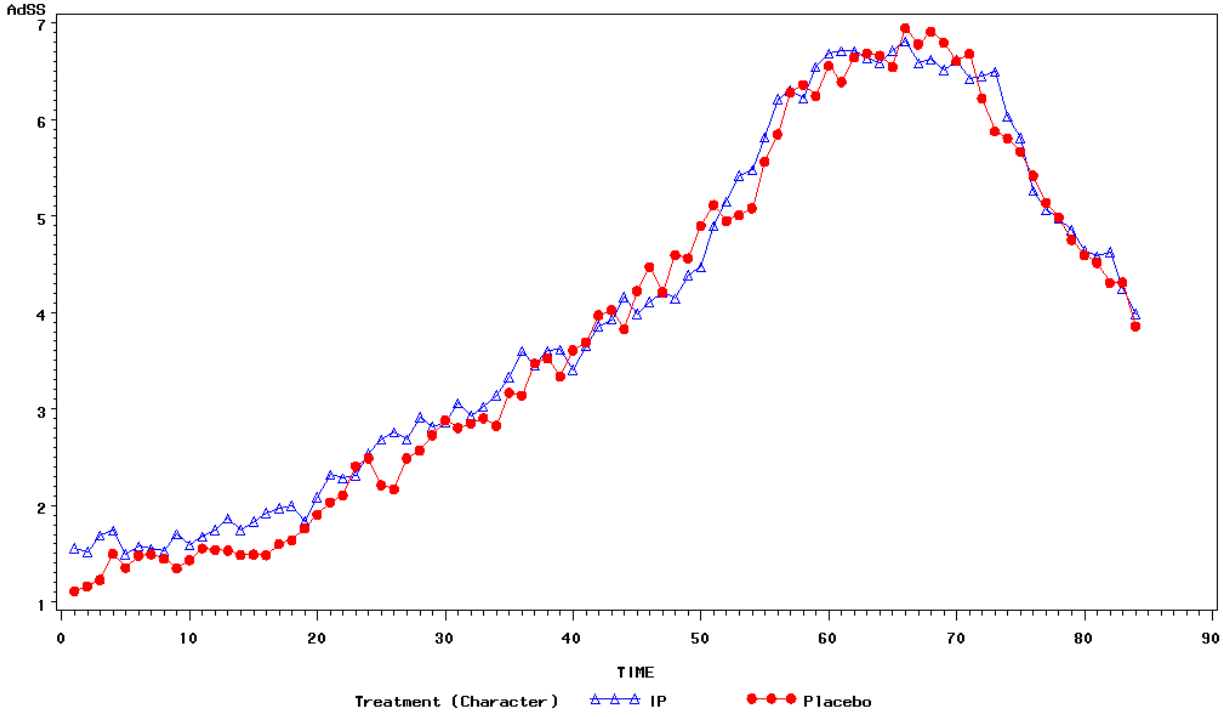
Afin d'étudier ces données manquantes, nous allons reconsidérer le modèle statistique. L'idée de départ est d'étudier les AdSS de façon longitudinale, et non d'étudier la moyenne sur toute la période de

pollen (AAdSS). En effet, l'AAdSS n'est jamais manquant, étant donné que c'est la moyenne des AdSS non manquants. Pour cela, nous nous concentrerons sur la pire saison de pollen qui se situe entre le 56ème jour et le 70ème, et on évaluera l'efficacité du traitement par la différence d'AdSS entre le 70ème et la valeur dite à « baseline » qui correspond au jour 1, où le patient n'était pas encore sous traitement.

Comme expliqué précédemment, les résultats de l'AdSS oscillent fortement, ce qui peut être un inconvénient avec ce type d'ajustement de modèle. En effet, comme on compare l'AdSS à deux points fixes dans le temps (en fin de saison et au début), il se peut que l'un de ces points fixes dans le temps soit par exemple une journée où le patient ait été en contact plus qu'habituellement avec du pollen, ce qui engendre un AdSS élevé ; ou au contraire, une journée sans pollen et donc un score d'AdSS faible. Cependant, on peut supposer que, globalement, sur l'ensemble des patients, une tendance pourra se dégager.

La figure 12 représente la moyenne des AdSS de tous les patients, par jour, de la visite 4 à la visite 6. On peut voir qu'il y a bien un pic des symptômes durant cette pire période de pollen.

Figure 14 : Moyenne des AdSS par jour



Remarque : tous les graphiques sont présentés en fonction du traitement. Même si cela n'a pas d'intérêt pour l'instant, cela permet de donner une idée sur la façon de raisonner une fois l'aveugle levé. Par exemple, dans le dernier graphique on pourra espérer détecter une différence entre les deux groupes en faveur de l'IP.

Partant de l'hypothèse MAR, nous allons appliquer la méthode d'imputation multiple de la façon suivante :

Nous utiliserons tout d'abord la méthode d'imputation partielle basée sur l'algorithme MCMC, sur les AdSS manquants qui sont répartis de façon non-monotone uniquement. Puis, sur les AdSS manquant qu'il restera, nous appliquerons la méthode de régression afin d'obtenir M jeux de données complets.

Ensuite, nous analyserons chacun de ces jeux de données indépendamment à l'aide d'une ANCOVA. Enfin, les résultats seront combinés pour n'obtenir qu'un seul résultat final.

3.3.3 Présentation des résultats

Le jeu de données utilisé (appelé *dym_ps_t*) contient une ligne par patient appartenant au FAS. Les différentes colonnes représentent les variables suivantes : Le traitement, l'OAS, le pays, l'âge, le sexe, l'asthme, la sensibilité au pollen ainsi que la valeur de l'AdSS à « baseline ». De plus les valeurs de l'AdSS pour chaque jour, du 56ème au 70ème, sont aussi représentées sous forme de colonnes. Ce sont ces dernières que l'on souhaite imputer lorsqu'elles sont manquantes. Toutes les autres variables étant connues.

En premier lieu, on peut afficher la façon dont les données manquantes sont réparties (monotone ou non-monotone) à l'aide de la PROC MI de la façon suivante :

```
proc mi data=dym_ps_t nimpute=0;
     var T56 - T70;
run;
```

Une partie des résultats obtenus sont présentés ci-dessous, les points représentant les données manquantes :

Figure 15 : Disposition des données manquantes

Group	T56	T57	T58	T59	T60	T61	T62	T63	T64	T65	T66	T67	T68	T69	T70	Freq
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	217
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	7
3	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	5
4	X	X	X	X	X	X	X	X	X	X	X	X	X	.	.	1
5	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	6
...																

On remarque que, sur les 527 patients, seulement 217 ont rempli toutes les données nécessaires au calcul des AdSS durant la pire saison de pollen.

La première étape consiste à imputer les données non-monotones afin de générer M (que l'on fixera à 5 pour commencer) jeux de données. On utilise pour cela la PORC MI de la façon suivante :

```
proc mi data=dym_ps_t out=dym_ps_mcmc simple nimpute=5 round=0.1;
    var KVTRTN KVCOUNTN KVAGE KVSEXN KVOASN KVASTHN KVSENSN T56 - T70;
    mcmc impute=monotone
         displayinit
         chain=multiple
         initial=EM
         outiter(mean cov)=IterParms
         outest=ImputParms;
run;
```

Les AdSS sont représentés par les colonnes T56 à T70. Une version numérique des autres variables est nécessaire. Les variables contenant des valeurs manquantes sont imputées chacune leur tour, en fonction de l'ordre dans lequel elles sont entrées lors de la programmation. Les seules variables

contenant des valeurs manquantes étant T56 à T70, les T56 seront donc imputés en premier, puis T57, etc…

L'option *impute=monotone* précise que l'on veut imputer uniquement les données non-monotones.

Displayinit affiche les valeurs des paramètres initiaux obtenus grâce à l'algorithme EM, nécessaire à l'initialisation de l'algorithme MCMC (correspondant au premier maillon de la chaine).

L'option *chain=multiple* permet d'utiliser des chaines de Markov multiples. Cela permet de s'assurer que les paramètres des différents modèles d'imputation, noté $\theta^{(m)}$, m = 1, …, M, sont des échantillons indépendant. Cette hypothèse est nécessaire afin d'obtenir des imputations « propres ».

Les deux dernières options (*outiter(mean cov)=IterParms* et *outest=ImputParms*) permettent de stocker les paramètres calculés à chaque itération, pour chacune des imputations, et les paramètres du modèle final de chaque imputation. Cela nous génère donc cinq jeux de données stockés dans une seule table (appelé ici *dym_ps_mcmc*).

Une fois les données non-monotones imputées, on impute les données monotones. Pour cela on applique la méthode basée sur la régression, toujours avec la PROC MI comme ci dessous:

```
proc mi data=dym_ps_mcmc out=dym_ps_reg nimpute=1;
  by _Imputation_;
  var KVTRTC KVCOUNTC KVAGE KVSEXC KVOASC KVASTHC KVSENSC T56 - T70;
  class KVTRTC KVCOUNTC KVSEXC KVOASC KVASTHC KVSENSC;
  monotone regression ;
run;
```

Tout comme l'étape précédente, les T56 seront donc imputés en premier, puis T57, etc… Cela génère un jeu de données pour chacun des cinq jeux déjà existants. On se retrouve alors avec 5 jeux de données complets stockés dans la table appelée *dym_ps_reg*.

L'étape suivante consiste à analyser les résultats de ces 5 jeux de données indépendamment. Pour cela on ajuste un modèle d'ANCOVA sur la valeur de l'AdSS le dernier jour, avec comme covariable le traitement et comme facteur la valeur à « baseline », tous deux à effets fixes. Cela permet de modéliser la différence des moyennes des moindres carrés de l'AdSS, entre le dernier jour et « baseline », à l'aide de la PROC MIXED, de la façon suivante :

```
proc mixed data=dym_ps_reg ;
  class KVTRTC;
  model T70 = KVTRTC T1/ solution;
    lsmeans KVTRTC / diff=control('IP') cl;
    ods output LSMeans=lsm Diffs=lsdiffs solutionF=Parms;
    by _Imputation_;
run;
```

La commande *by _Imputation_* permet d'analyser chacun des 5 jeux de données indépendamment.

On stocke les données obtenues dans 3 tables afin de pouvoir les combiner lors de la dernière étape. La première table (appelé *lsm*) contient les moyennes des moindres carrés de chaque traitement. Les résultats obtenus sont les suivants :

Tableau 10 : Moyennes des moindres carrés de l'AdSS

IMPUTATION	EFFECT	KVTRTC	ESTIMATE	STDERR	DF	TVALUE	PROBT	ALPHA	LOWER	UPPER
1	KVTRTC	IP	6.4152	0.2963	539	21.65	<.0001	0.05	5.8331	6.9972
1	KVTRTC	Placebo	6.5052	0.2941	539	22.12	<.0001	0.05	5.9275	7.0829
2	KVTRTC	IP	6.6608	0.2914	539	22.85	<.0001	0.05	6.0883	7.2333
2	KVTRTC	Placebo	6.6969	0.2893	539	23.15	<.0001	0.05	6.1286	7.2652
3	KVTRTC	IP	6.5699	0.2910	539	22.58	<.0001	0.05	5.9983	7.1415
3	KVTRTC	Placebo	6.4655	0.2888	539	22.38	<.0001	0.05	5.8982	7.0329
4	KVTRTC	IP	6.4674	0.2962	539	21.83	<.0001	0.05	5.8856	7.0493
4	KVTRTC	Placebo	6.4335	0.2940	539	21.88	<.0001	0.05	5.8560	7.0111
5	KVTRTC	IP	6.5641	0.3006	539	21.84	<.0001	0.05	5.9736	7.1546
5	KVTRTC	Placebo	6.4855	0.2984	539	21.74	<.0001	0.05	5.8994	7.0717

On remarque que les résultats sont assez similaires d'une imputation à l'autre. On espérera détecter des différences entre les 2 traitements une fois l'aveugle levé, avec des moyennes plus élevées pour les groupes de patients ayant pris un placebo.

La deuxième table (*lsdiffs*) permet de stocker les différences des moyennes des moindres carrés pour chaque imputation ainsi que les p-valeurs associées (PROBT) au test suivant :

H_0 : Il n'y a pas de différence entre les deux traitements

H_1 : Il y a une différence entre les deux traitements

Le niveau du test α est fixé à 5%.

On obtient les résultats suivants :

Tableau 11 : Différences des moyennes des moindres carrés de l'AdSS

IMPUTATION	EFFECT	KVTRTC	_KVTRTC	ESTIMATE	STDERR	DF	TVALUE	PROBT	ALPHA	LOWER	UPPER
1	KVTRTC	Placebo	IP	0.09006	0.4185	539	0.22	0.8297	0.05	-0.7319	0.9121
2	KVTRTC	Placebo	IP	0.03615	0.4116	539	0.09	0.9300	0.05	-0.7724	0.8447
3	KVTRTC	Placebo	IP	-0.1044	0.4110	539	-0.25	0.7996	0.05	-0.9116	0.7029
4	KVTRTC	Placebo	IP	-0.03390	0.4183	539	-0.08	0.9354	0.05	-0.8556	0.7878
5	KVTRTC	Placebo	IP	-0.07862	0.4245	539	-0.19	0.8532	0.05	-0.9126	0.7553

Dans tous les cas les p-valeurs sont supérieures à 0,05. On ne peut donc pas rejeter l'hypothèse nulle, ce qui signifie qu'il n'y a pas de différence significative entre les deux traitements.

La dernière table (*Parms*) fournit les estimations des paramètres du modèle :

Tableau 12 : Estimation des paramètres du modèle

IMPUTATION	EFFECT	KVTRTC	ESTIMATE	STDERR	DF	TVALUE	PROBT
1	Intercept		6.0093	0.3102	539	19.37	<.0001
1	KVTRTC	IP	-0.09006	0.4185	539	-0.22	0.8297
1	KVTRTC	Placebo	0				
1	T1		0.3712	0.09034	539	4.11	<.0001
2	Intercept		6.0264	0.3051	539	19.75	<.0001
2	KVTRTC	IP	-0.03615	0.4116	539	-0.09	0.9300
2	KVTRTC	Placebo	0				
2	T1		0.5020	0.08886	539	5.65	<.0001
3	Intercept		5.9549	0.3046	539	19.55	<.0001
3	KVTRTC	IP	0.1044	0.4110	539	0.25	0.7996
3	KVTRTC	Placebo	0				
3	T1		0.3823	0.08872	539	4.31	<.0001
4	Intercept		5.9633	0.3101	539	19.23	<.0001
4	KVTRTC	IP	0.03390	0.4183	539	0.08	0.9354
4	KVTRTC	Placebo	0				
4	T1		0.3520	0.09031	539	3.90	0.0001
5	Intercept		5.9649	0.3147	539	18.95	<.0001
5	KVTRTC	IP	0.07862	0.4245	539	0.19	0.8532
5	KVTRTC	Placebo	0				
5	T1		0.3898	0.09166	539	4.25	<.0001

La dernière étape consiste à fusionner les résultats de chaque imputation pour chacune de ces trois tables. Pour cela on utilise la PROC MIANALYZE de la façon suivante :

```
proc mianalyze parms(classvar=full)=lsm;
  class KVTRTC;
  modeleffects KVTRTC;
    ods output ParameterEstimates=MIAN_lsm;
run;

proc mianalyze parms(classvar=full)=lsdiffs;
  class KVTRTC;
  modeleffects KVTRTC;
    ods output ParameterEstimates=MIAN_lsdiffs;
run;

proc mianalyze parms(classvar=full)=parms;
  class KVTRTC;
  modeleffects Intercept KVTRTC T1;
  ods output ParameterEstimates=MIAN_Parms;
run;
```

Les résultats obtenus sont calculé à partir des estimations des paramètres du modèle et de l'écart type des trois tables précédentes (colonnes ESTIMATE et STDERR). On obtient alors les résultats finaux suivants :

Tableau 13 : Résultats finaux

Final Least Squares Means

PARM	KVTRTC	ESTIMATE	STDERR	LCLMEAN	UCLMEAN	DF	MIN	MAX	THETA0	TVALUE	PROBT
KVTRTC	IP	6.535480	0.313275	5.919112	7.151848	315.96	6.415154	6.660757	0	20.86	<.0001
KVTRTC	Placebo	6.517348	0.314244	5.898228	7.136467	233.21	6.433544	6.696910	0	20.74	<.0001

Final Differences of Least Squares Means

PARM	KVTRTC	ESTIMATE	STDERR	LCLMEAN	UCLMEAN	DF	MIN	MAX	THETA0	TVALUE	PROBT
KVTRTC	Placebo	-0.018132	0.426058	-0.85366	0.817392	2167.3	-0.104358	0.090061	0	-0.04	0.9661

Final Solution for Fixed Effects

PARM	KVTRTC	ESTIMATE	STDERR	LCLMEAN	UCLMEAN	DF	MIN	MAX	THETA0	TVALUE	PROBT
Intercept		5.983751	0.310935	5.37430	6.593201	25005	5.954871	6.026358	0	19.24	<.0001
KVTRTC	IP	0.018132	0.426058	-0.81739	0.853657	2167.3	-0.090061	0.104358	0	0.04	0.9661
KVTRTC	Placebo	0					0	0	0		
T1		0.399460	0.110824	0.17435	0.624570	34.462	0.352018	0.501988	0	3.60	0.0010

La p-valeur vaut finalement 0.9661. Une fois l'aveugle levé on espérera que celle-ci soit inférieure à 0.05 afin de pouvoir rejeter l'hypothèse nulle et ainsi de conclure à une différence significative entre les traitements.

En utilisant le même modèle statistique sur les patients ayant des AdSS non manquants le 70ème jour (437 patients), on obtient les résultats suivant :

Tableau 14 : Résultats sans traitement des données manquantes

Least Squares Means

EFFECT	KVTRTC	ESTIMATE	STDERR	DF	TVALUE	PROBT	ALPHA	LOWER	UPPER
KVTRTC	IP	6.5143	0.3258	434	20.00	<.0001	0.05	5.8740	7.1546
KVTRTC	Placebo	6.7062	0.3310	434	20.26	<.0001	0.05	6.0555	7.3568

Differences of Least Squares Means

EFFECT	KVTRTC	_KVTRTC	ESTIMATE	STDERR	DF	TVALUE	PROBT	ALPHA	LOWER	UPPER
KVTRTC	Placebo	IP	0.1919	0.4654	434	0.41	0.6803	0.05	-0.7228	1.1065

Solution for Fixed Effects

EFFECT	KVTRTC	ESTIMATE	STDERR	DF	TVALUE	PROBT
Intercept		6.0692	0.3482	434	17.43	<.0001
KVTRTC	IP	-0.1919	0.4654	434	-0.41	0.6803
KVTRTC	Placebo	0				
T1		0.5727	0.1174	434	4.88	<.0001

On remarque que la moyenne des moindres carrés est quasiment identique. La p-valeur quant à elle est plus faible dans le modèle sans imputation mais reste loin du seuil de significativité.

Afin de mesurer l'efficacité, la PROC MIANALYZE permet de calculer le RE. Les sorties SAS ci-dessous nous montrent les résultats de ce RE pour M = 5, 10 et 50 imputations, « Fraction Missing Information » étant le γ mentionné précédement.

Tableau 15 : Efficacité en fonction du nombre d'imputations M

The MIANALYZE Procedure

Model Information

PARMS Data Set	WORK.LSM
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	KVTRTC	Variance: Between	Variance: Within	Variance: Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
KVTRTC	IP	0.009202	0.087099	0.098141	315.96	0.126781	0.118081	0.976929
KVTRTC	Placebo	0.010777	0.085816	0.098749	233.21	0.150702	0.138323	0.973080

The MIANALYZE Procedure

Model Information

PARMS Data Set	WORK.LSM
Number of Imputations	10

Multiple Imputation Variance Information

Parameter	KVTRTC	Variance: Between	Variance: Within	Variance: Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
KVTRTC	IP	0.006236	0.088975	0.095835	1756.6	0.077098	0.072634	0.992789
KVTRTC	Placebo	0.018649	0.087665	0.108179	250.29	0.234001	0.196026	0.980774

The MIANALYZE Procedure

Model Information

PARMS Data Set	WORK.LSM
Number of Imputations	50

Multiple Imputation Variance Information

Parameter	KVTRTC	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
		Between	Within	Total				
KVTRTC	IP	0.008693	0.089629	0.098497	6045.9	0.098932	0.090327	0.998197
KVTRTC	Placebo	0.011481	0.088310	0.100020	3574.6	0.132605	0.117573	0.997654

On remarque que l'efficacité n'augmente que faiblement avec le nombre d'imputations, et qu'elle est déjà satisfaisante avec M=5. On pourra éventuellement augmenter M à 10 mais au-delà le gain d'efficacité est minime.

3.3.4 Conclusion sur l'essai

L'avantage de l'imputation multiple est sa flexibilité. En effet, une seule méthode a été présentée ici alors que l'on aurait pu analyser les modèles obtenus après imputations de différentes manières, par exemple en prenant en compte d'autres covariables, en rajoutant des effets aléatoires ou en travaillant sur un modèle à mesures répétées (MMRM). Combinées à cela, différentes méthodes d'imputations auraient aussi pu être implémentées, comme des méthodes non paramétriques basées sur les scores de propension.

Cependant, le modèle statistique le mieux adapté à cet essai clinique reste celui basé sur les AAdSS, pour les raisons précédemment mentionnées (principalement les fortes oscillations des résultats dûs à la spécificité de cette étude). De plus, le contrôle des autorités de santé reste très strict en ce qui concerne le traitement des données manquantes. En effet, rares sont les laboratoires pharmaceutiques qui tentent d'implémenter des méthodes autres que des imputations simples, les méthodes les plus utilisées étant la méthode LOCF et la méthode dite « worst case » qui ne risque pas de favoriser l'IP.

CONCLUSION

Ce stage a été pour moi très enrichissant, autant professionnellement qu'humainement. Ayant mis l'accent sur la partie pratique en participant à l'analyse statistique d'une étude en particulier, j'ai pu parfaire mes connaissances dans le domaine pharmaceutique ainsi que dans celui des statistiques appliquées aux essais cliniques, secteur dans lequel je souhaitais me spécialiser. Cela m'a d'ailleurs conforté dans l'idée de faire carrière dans ce domaine. De plus, ce stage m'a permis de me familiariser avec le logiciel SAS, compétence actuellement précieuse dans le domaine professionnel des statistiques. En effet, de nombreuses heures de programmation ont été nécessaires, principalement pour dériver les données (par exemple pour le calcul des AdSS), produire des tables et des graphiques, ajuster les modèles statistiques… Soulignons également deux aspects essentiels de la pratique sur lesquels j'ai eu l'occasion de me perfectionner de façon quotidienne : le travail en équipe ainsi que la compréhension et l'expression de l'anglais.

Par ailleurs, le travail de recherche sur les données manquantes m'a permis de découvrir un domaine des statistiques que je ne connaissais quasiment pas, même si, par manque de temps, je n'ai pas pu approfondir davantage le sujet. En effet, la théorie des données manquantes est un domaine très vaste, souvent controversé et pouvant être basé sur une théorie forte, en particulier dans le cas de données MNAR. C'est en faisant ce genre de recherches que l'on se rend compte de l'importance de minimiser le nombre de données manquantes lors de la récolte des données. Tout mettre en œuvre pour limiter le nombre de données manquantes reste la meilleure option.

Cette immersion au sein du département de biostatistique de Quintiles m'a permis d'appréhender au mieux la transition entre la vie étudiante et professionnelle.

Liste des figures

Liste des tables

REFERENCES

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd edn), Wiley.

Molenberghs, G., Kenward, M.G. (2007). *Missing Data in Clinical Studies*, Statistics in Practice, Wiley.

PSI missing data expert group (2010). Missing data: Discussion points, *Pharmaceutical Statistics*, **9**, 288-297.

Rubin, D.B. (1976). Inference and missing data, *Biometrika*, **63**, 581-592.

SAS 9.1 User's Guide, PROC MI and PROC MIANALYZE.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schafer, J.L. (1999). Multiple imputation : a primer, *Statistical Methods in Medical Research*, **8**, 3-15.

White, I.R., Carlin, J.B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values, *Statistics in Medecine*, **29**, 2920-2931.