Licence Informatique parcours Mathématiques-Informatique
Université des Sciences et des Techniques de Nantes
2008-2009

Contrôle de cohérence de DIVAT Cohorte observationnelle de données médicales des patients transplantés rénaux

Chloé Le Tétour

INSERM Unité 643
CHU de Nantes

Maître de stage : Yohann FOUCHER
Responsable universitaire : Chantal ENGUEHARD

Table des matières

Remerciements

Je tiens à remercier et à témoigner toute ma reconnaissance à toutes les personnes du service de transplantation, pour l'expérience enrichissante et pleine d'intérêt qu'elles m'ont fait vivre durant ces trois mois de stage.

Je remercie plus particulièrement mon tuteur, Yohann FOUCHER, pour m'avoir intégré rapidement et m'avoir accordé toute sa confiance ; pour le temps qu'il m'a consacré tout au long de cette période, sachant répondre à toutes mes interrogations. Je lui suis très reconnaissante pour toutes les connaissances qu'il m'a apporté.

Merci à Pascal DAGUIN et Magali GIRAL de m'avoir accordés leur confiance pour ce projet.

Merci aussi à Magali GIRAL et Morganne GOSSELIN pour la place qu'elles m'ont fait dans leur bureau, ainsi que l'accueil qu'elles m'ont accordé dès mon arrivée dans le service.

Je remercie également, Mesdames Cécile GIRAL et Assia HAMI pour leur soutien et leur bonne humeur. Elles ont su me donner d'excellents conseils tout au long de ces trois mois.

Introduction

Le stage de fin de licence d'informatique à l'université de Nantes s'inscrit dans une démarche de prise en main du projet professionnel. Il permet de connaître le milieu professionnel. Ce stage permet à l'étudiant de découvrir ce que pourra être une mission d'informaticien au sein d'une entreprise.

Ce stage, de 3 mois, s'est déroulé au CHU de Nantes, plus précisément dans l'Unité 643 de l'Institut Nationnal des Sciences Et de la Recherche Médicale (INSERM). La mission était de réaliser un contrôle de cohérence sur une base de données contenant les données médicales des patients qui ont bénéficié d'une transplantation rénale : DIVAT.
Cette base de données mémorise les dossiers patients et est également utilisée comme source de données pour les recherches cliniques sur les transplantations rénales. Ce contrôle de cohérence vise à améliorer la fiabilité des études cliniques.

Afin de bien comprendre le contexte dans lequel se place cette mission nous présenterons dans un premier temps sur l'environnement dans lequel ce stage a eu lieu.
Nous expliquerons l'utilité de ce projet pour la base de données DIVAT.
Nous reviendrons ensuite sur le travail réalisé.
Puis nous développerons la mission en elle-même afin d'avoir une vision la plus globale possible de ce qui a été réalisé lors de ces trois mois de stage.

Chapitre 1

Présentation de l'environnement

1.1 Présentation l'unité 643 de INSERM

L'unité 643 de l'Institut Nationnal de la Santé Et de la Recherche Médicale (INSERM) fait partie de l'Institut de Tranplantation Et de Recherche en Transplantation (ITERT). Cet institut regroupe plusieurs services : l'INSERM U643, le service de néphrologie, le service de transplantation et le service de consultation.
L'ITERT coordonne ces services. Le service de néphrologie accueille les patients insuffisants rénaux, celui de transplantation est destiné aux patients transplantés et le service de consultation de jour est le lieu ou les patients, transplantés ou non, viennent en consultation.
L'INSERM U643, créé en 2004, est spécialisé dans *l'immunointervention dans les allo et xénotransplantations*. Cette unité de l'INSERM est composée de 104 personnes réparties en neuf groupes de travail. L'équipe qui m'a accueillie est une équipe transversale (dirigée par le Dr. Pr. Magali GIRAL) qui coordonne l'activité de recherche clinique.

1.2 Présentation de DIVAT

1.2.1 Historique

DIVAT est une base de données contenant les données médicales des patients ayant bénéficié d'une transplantation rénale.
Elle a été créée il y une quinzaine d'années à la demande de l'ITERT afin de faciliter l'utilisation des données des patients transplantés rénaux.
Le patient est ajouté à la base lorsqu'il bénéficie d'une transplantation, il est ensuite suivi tout au long de la vie de son greffon. Ce suivi constitue une base de données de grande ampleur qui permet de réaliser des études cliniques. Cette base de données est ce que l'on appel une cohorte observationnelle.
DIVAT a été initié à Nantes et a ensuite été rejoint par cinq autres centres hospitaliers francais, Paris (Necker), Nancy, Toulouse, Montpellier et Lyon. La base contient aujourd'hui près de 16000 dossiers de greffe, dont 5000 à Nantes.

Il y a aujourd'hui plusieurs extensions de DIVAT à d'autres services, la pédiatrie, l'urolo-

gie et le pancréas ont créé leur propre base de données. Outre ces extensions complètement détachées de la transplantation rénale, la volonté d'avoir un dossier patient le plus complet possible a donné lieu à deux extensions dédiées à la transplantation rénale. Pour avoir, en plus des données médicales du patients, ses données biologiques il a été mis en place DIVAT-biocol qui renseigne toutes les informations (emplacement, quantité, ...) sur les échantillons biologiques que possède le patient à l'hôpital. Pour avoir un réel suivi des patients suivi dans un centre extérieur au centre de transplantation, DIVAT-ville assure la complétion du dossier.

1.2.2 DIVAT (Données Informatiques VAlidées en Transplantation)

DIVAT est une base de données de 250 items renseignant le dossier médical d'un patient ayant profité d'une transplantation rénale. L'ensemble du réseau DIVAT contient prés de 16000 dossiers.

Ce dossier informatisé comprend :

- Les données administratives.
- Les données médicales élémentaires.
- Les antécédents.
- Les données immunologiques.
- Les données sur les échantillons biologiques.
- Le suivi médical depuis la transplantation.
- Les renseignements sur les possibles rejets, infections et complications.
- Les renseignements concernant le donneur.

Cette base de données est complétée par des Assistants de Recherche Clinique (ARC) lors de chaque nouvelle transplantation. Nantes est le plus gros centre de greffe rénale en France, près de 190 greffes par an.

La base est consultée quotidiennement par les médecins, le personnel hospitalier et aussi par les chercheurs.
Elle est utilisée administrativement pour gérer les rendez-vous des patients, connaître les dernières directives du service, avoir les formulaires destinés aux patients.
Les médecins l'utilisent comme un dossier patient lors des consultations car DIVAT contient l'ensemble du passé du patient. Ils peuvent grâce à ce dossier informatisé accéder à toutes les données du patient très facilement en naviguant dans les différents onglets. Ils accèdent aux prescriptions qu'a pu avoir le patient depuis sa transplantation, ses résultats biologiques (résultat de biopsie par exemple)... La base renseigne si le patient fait, ou a fait parti d'un protocole, ce qui est pris en compte lors de nouvelles décisions médicales. Effectivement les patients participant à un protocole doivent être soignés avec des traitements spéciaux.
Les chercheurs se servent de la base comme d'une base de données médicales au service des études épidémiologiques. Ils ne se servent évidement pas des données personnelles du patient mais seulement des données nécessaires à leur étude. Usuellement les centres hospitaliers créent des bases de données médicales uniques pour chaque étude. Dans les centres hospitaliers du réseau DIVAT, toutes les études cliniques se font à partir de DIVAT.

Chapitre 2

Présentation de la mission de stage

La mission est de faire un contrôle de cohérence des données contenues dans DIVAT. Ce contrôle est effectué de manière automatique à intervalles réguliers.

2.1 L'utilité de ce contrôle

Cette mission s'inscrit dans l'objectif de validation des données de DIVAT.
Aujourd'hui le contrôle et la validation des données de DIVAT se fait par un cross-audit une fois par an. Lors de cet audit il y a un contrôle de 30 dossiers par centre, soit 180 dossiers sur 16000. Ces dossiers sont contrôlés avec le dossier papier du patient, sur tous les items, par les ARC des autres centres. DIVAT s'est fixé un pourcentage de moins de 1% d'erreur.

L'objectif de DIVAT est d'effectuer un contrôle sur d'avantage de dossiers mais sur un nombre d'items restreint à ceux qui sont essentiels. Si, après ce contrôle, plus de 1% des dossiers vérifiés contiennent des erreurs il y aura alors vérification de tous les dossiers sur ces items.
En attente de la réalisation de ce contrôle, DIVAT voudrait un contrôle plus régulier de ses données. C'est pourquoi il met en place ce contrôle de cohérence automatique.

Ce contrôle est nécessaire car, malgré les audits annuels, actuellement lors des études cliniques il y a près de 50% des sujets concernés que l'on doit exclure uniquement parce qu'ils ont des items non renseignés. Lorsque l'information est inconnue les modèles statistiques peuvent le gérer mais lorsque l'information n'a pas été complétée par oubli cela ne peut se modéliser, il faut alors exclure le patient de l'étude. Il y a un réel souci de différentiation de ces deux types de données manquantes afin d'avoir des études de plus grande ampleur, c'est le sujet de la première partie du contrôle.

Ce contrôle est conçu pour la fiabilité du dossier patient c'est pourquoi nous contrôlons aussi la cohérence des données du patient entre elles, ce qui assurera aussi la fiabilité des études épidémiologiques.

2.2 Le contrôle

Ce contrôle est effectué à intervalles réguliers il émet un rapport, au format pdf, par centre et est consultable par internet sur DIVAT.
A partir de DIVAT on extrait automatiquement l'ensemble des données essentielles inscrites dans la base, soit près de 30 items sur 16000 dossiers. On exécute ensuite un programme, écrit en langage R, qui effectue quatre types de contrôles sur les données par centre. Puis grâce à l'extension `Sweave` et `xtable` des langages R et Latex, on retourne dans DIVAT un rapport au format pdf contenant des tableaux où figure toutes les erreurs repérées. Ce rapport est alors placé dans un module de DIVAT qui lui est propre et envoyé par mail aux ARC et aux responsables de chaque centre. Les ARC consultent ce rapport à chaque émission et corrigent les dossiers des patients.

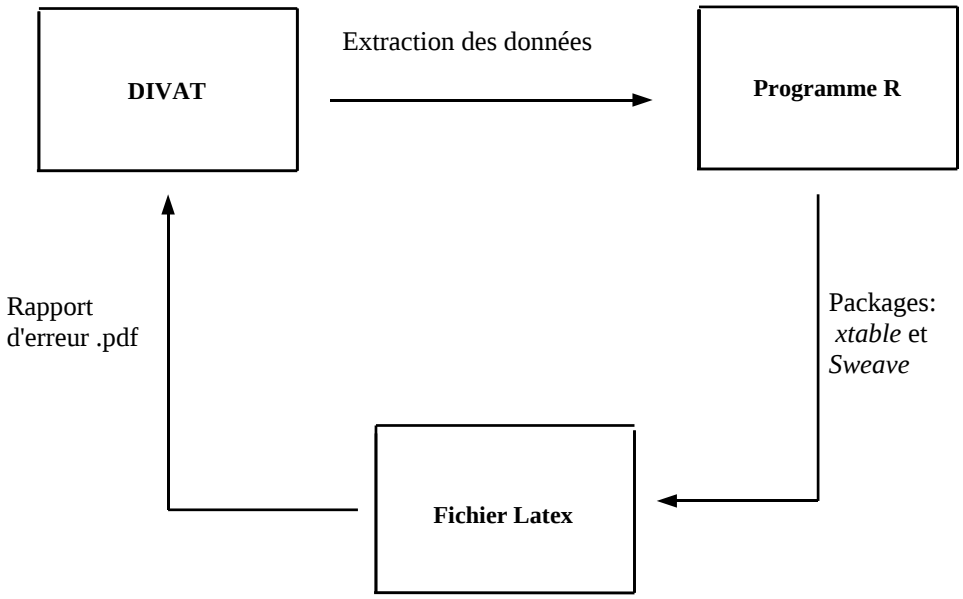

FIGURE 2.1 – Diagramme du cheminement du contrôle de cohérence

Ce contrôle est effectué sur trois types d'erreurs

1. Les données manquantes.
2. Les données incorrectes.
3. Les données suspectes.

La quatrième partie est dédiée à la confirmation des dates de perte de vue des patients.

2.2.1 Les valeurs manquantes

Les données manquantes sont les données non complétées par les ARC.

Ce problème de valeurs manquantes a donné lieu il y a quelques années à la mise en place d'un code couleur. Avant que l'ARC ne complète le dossier d'un nouveau greffé tous les items sont précédés d'une boule rouge. Si l'ARC veut renseigner le champ il doit cliquer sur cette boule, la transformer en verte, et compléter la valeur du champ. Si l'ARC ne possède aucune information pour cet item il doit alors transformer la boule en violette.
Ce code couleur permet de différentier les valeurs complétées, les valeurs non-complétées et les valeurs manquantes au dossier.
Lors de l'extraction des données pour une étude épidémiologique on retrouve grâce à ce code couleur trois types de valeurs :

1. Les champs qui ont été complétés (boule verte) possèdent bien évidemment la valeur entrée.

2. Les champs qui ont été renseignés comme donnée introuvable (boule violette) sont remplacés par -99 pour les numériques, Øpour les chaînes de caractères et 01/01/1100 pour les dates.

3. Les champs oubliés (boule rouge) sont eux des valeurs manquantes NA.

C'est ainsi que nous ferons la différence dans le contrôle et que nous retournerons les valeurs incomplétées.

Ici nous ne testons que les items essentiels de la base, cela représente une trentaine d'items sur les 250. Nous avons déterminé ces items en prenant en compte les besoins des études cliniques. C'est effectivement lors des études qu'il y a un réel besoin de différentiation des données réellement manquantes de celles manquantes par oubli, car pour le dossier du patient cela n'a pas grande importance de savoir pourquoi la donnée est manquante. En revanche ce que cette partie peut apporter au dossier du patient c'est la complétion de celui-ci, en effet si, par oubli, l'assistant avait omis une donnée elle sera retournée en erreur dans cette partie et, après recherche, la donnée sera complétée.

2.2.2 Les valeurs incorrectes

Les données incorrectes sont celles qui, selon des critères médicaux, sont impossibles et celles qui sont incohérentes avec les autres données du patient (l'âge et date de naissance ne correspondant pas, ...).
Les données sont incorrectes médicalement si elles dépassent des bornes établies en accord avec les médecins. Les rapprochements possibles entres deux données ont aussi été validés par les médecins.
Ces tests de cohérence seront bien évidemment évolutifs avec les résultats des recherches épidémiologiques, de nouveaux paramètres de cause à effet pourrons être la source de nouveaux tests.

2.2.3 Les valeurs suspectes

Les données suspectes sont celles qui ne ressemblent pas aux données des autres patients.
Cette partie utilise un modèle statistique pour modéliser les variations des données. Nous ne contrôlons ici que cinq items : l'âge du donneur et du receveur, la taille et le poids du receveur et le temps d'ischémie froide. Ces items sont les plus importants dans les études cliniques, ils sont indispensables il est donc nécessaire qu'ils soient plus fortement contrôlés afin de garantir au maximum leur validité.

2.2.4 Les dates de perte de vue des patients

La date de perte de vue des patients est la dernière date saisie dans la base pour les patients dont on n'a plus de nouvelles depuis 2 ans.
Cette date est affectée automatiquement par DIVAT au bout de 2 ans. Mais classifier quelqu'un comme perdu de vue implique qu'il ne sera plus dans les études cliniques, il faut donc que cette classification soit validée par le personnel et non par le logiciel.
Cette quatrième partie permet d'indiquer aux ARC les nouveaux patients qui ont été définis comme perdu de vue par le logiciel, ils doivent aller rechercher la cause de cette perte de vue et valider ou non le patient comme tel.

2.3 L'évolution du contrôle

Ce contrôle est une possibilité d'améliorer le travail de chacun, pour les médecins (pour qu'ils aient un dossier très fiable), pour les chercheurs (pour qu'ils mènent des études de grande importances et très fiables), à terme la gestion des dossiers en sera simplifié. Dans un premier temps ce contrôle peut être une source de travail supplémentaire pour les ARC. Pour ne pas surcharger les assistants de recherches le rapport sera dans les premiers temps effectué moins fréquemment qu'à terme.
Les ARC seront les premières personnes à être contactées après quelques émissions. Ces échanges pourront ainsi mettre en lumière l'efficacité de détection d'erreur. Si il y a un taux d'erreur très élevé à un test il pourra être mis en place directement à la saisie des données, ce qui ne demanderait pas aux assistants de ressortir les dossiers. A contrario, si on remarque qu'un test mal apprécié, retourne à tort des erreurs, nous pourrons alors supprimer ce contrôle. Les contrôles pourrons être améliorés.
Si certaines études épidémiologiques mettent en avant de nouvelles relations de cause à effet, nous pourrons les intégrer à notre contrôle.
Il pourra y avoir une amélioration des modèles statistiques utilisés, effectivement un spécialiste de ce type de modélisation pourrait sûrement améliorer l'approche statistique que nous avons réalisé.

Ce contrôle est un contrôle qui sera facilement modulable suivant les retours auxquels il donnera lieu. C'est dans un objectif d'amélioration du travail de chacun qu'il a été réalisé, il s'adaptera aux demandes diverses qu'il pourrait susciter.

Chapitre 3

Travail réalisé lors du stage

La réalisation de la mission a commencé par une prise de connaissance de l'outil DIVAT et des raisons pour lesquelles ce contrôle de cohérence est nécessaire. Il a ensuite fallu que je me familiarise avec les langages R et Latex.
Après cette phase d'adaptation, le déroulement de la réalisation de la mission a globalement suivi l'ordre final du rapport émis par ce contrôle : création du contrôle des valeurs manquantes, des valeurs incorrectes, des valeurs suspectes puis des dates de perte de vue.

3.1 La découverte de la mission et de ses enjeux

Ce projet sera évolutif dans les prochaines années, les modifications pourront être pris en charge dans le cadre de la maintenance de DIVAT par IDBC, en revanche sa mise en place nécessitait un travail plus poussé d'où la prise en charge d'un stagiaire par le réseau DIVAT.

3.1.1 Le service de transplantation et DIVAT

Comme la mission porte sur une base de données médicales de patients transplantés rénaux je me suis familiarisé avec ce milieu médical que je ne connaissais pas.

La mission étant de nature informatique aurait pu se dérouler dans les locaux de l'entreprise IDBC. Mon maître de stage, Yohann Foucher, travaillant dans les locaux de l'hôpital, elle s'est déroulée dans les locaux de l'ITERT au CHU de Nantes. Être dans les locaux où se trouve les utilisateurs de la base m'a permi une meilleure compréhension de l'utilité et l'utilisation de DIVAT.
Les premiers jours dans le service ont été très instructifs pour comprendre le milieu médical. Les notions médicales qui étaient encore floues après l'explication du processus de transplantation se sont peu à peu éclaircies grâce aux collègues. La découverte de la recherche dans le monde médical s'est faite à travers différentes conférences [1] de présentation des évolutions des recherches en cours. Bien que complexes, car relevant du domaine médical, ces conférences ont constitué une introduction éclairante au monde de la recherche

1. *Immune evasion mechanisms of human cytomegalovirus in dentric cells*, Fabienne Haspot le 09/04/2009. *Anti-HLA antibodies isotypes in kidney transplantation*, Sandrine Leroux le 16/04/2009.

médicale.

Pour découvrir en profondeur DIVAT, se mettre à la place d'un utilisateur fut très instructif. J'ai ajouté un nouveau patient greffé dans la base. La complétion, en compagnie d'un médecin arrivé le même jour que moi dans le service, des 250 champs à partir du dossier papier du patient n'a pas été aisée mais, aidée pour les notions médicales, cette tâche m'a permis de découvrir les différentes fonctionnalités de cet outil. DIVAT est riche en informations sur le patient, le parcours de ses onglets n'est pas suffisant pour bien en comprendre toutes ses ramifications.
Ajouter un nouveau dossier à DIVAT éclaire sur le travail qu'effectuent les assistants de recherche clinique tous les jours. Or la mission du stage étant de contrôler les dossiers qu'ils complètent, voir ce que ce travail représente est essentiel pour évaluer la forme et la taille optimale que doivent prendre les rapports qui leur seront destinés.

3.1.2 Les enjeux de la mission

Le contrôle de cohérence a tout d'abord été demandé pour améliorer la fiabilité des études cliniques et l'impact qu'elles ont.

Comme nous l'avons déjà expliqué précédemment, la première partie sur la détection des champs non complétés vise à ne pas perdre inutilement des individus lors des études cliniques. Les équipes auront ainsi des études de grande ampleur pouvant avoir un impact majeur sur la communauté internationale car la base de données, avec ses 16000 individus, est l'une des plus importantes au monde.

La seconde partie est axée sur la validité des données elles-mêmes. cette qualité est essentielle pour les médecins afin qu'ils aient des données certifiées exactes lors de leurs prises de décision médicale et également pour la précision des études épidémiologiques.

La troisième partie s'attarde sur les facteurs nécessairement utilisés dans les recherches. Il s'agit d'un second contrôle, ici statistique, qui rend encore plus fiable l'exactitude de ces paramètres.

La dernière section est destinée à faciliter le travail des assistants qui renseignent DIVAT. lors de cette étape ils seront informés des nouveaux patients déclarés comme perdus de vue, ils doivent alors vérifier s'ils le sont réellement. Ce contrôle profite aux ARC car il n'est plus nécessaire de rechercher les patients dans la base, cette quatrième partie l'indique et ils n'ont plus qu'à confirmer.

Être auprès des trois parties concernées, les médecins, les assistants de recherche clinique et les chercheurs, permet d'apprécier l'apport positif de ce contrôle : il permet à tous de travailler avec un meilleur outil, ce qui se ressent sur la qualité de suivi du patient et sur la qualité des recherches menées à partir des informations contenu dans DIVAT.
Cependant,le premier objectif reste l'amélioration des soins apportés aux patients.

3.2 Les langages de programmation utilisés

Le travail a été réalisé grâce au langage R, habituellement dédié aux statistiques, et au traitement de texte Latex.

Un autre langage aurait pû être utilisé si le contrôle n'avait pas comporté la partie statistique. En effet, le contrôle statistique implique l'utilisation d'un langage spécifique, le langage R étant gratuit et utilisé lors des recherches épidémiologiques a semblé approprié. Ce langage permet à la fois de programmer, comme tout autre langage de programmation, et d'effectuer des analyses statistiques sur de grands échantillons.
Pour apprendre ces langages je suis passée par l'analyse de l'ébauche du contrôle de cohérence qui avait déjà été réalisé, ce qui m'a permis de comprendre plus rapidement le langage.

Le traitement de texte Latex est utilisé pour la conversion en pdf du rapport. Ce traitement de texte offre une possibilité de contrôle de la mise en forme tout en y inserrant progressivement les tableaux contenant les erreurs. Ce traitement de texte est un langage, une fois le fichier.tex compilé il émet un fichier pdf. Le rapport peut donc être émis automatiquement, sans intervention humaine pour émission.

Afin d'inclure les tableaux contenant les individus sélectionnés avec le langage R, dans le rapport Latex, il a fallu relier ces deux langages. Pour cela nous avons adopté deux approches différentes.

3.2.1 La première approche

La première approche a été initiée lors de l'ébauche du programme de Yohann Foucher. Cette approche nécessite trois fichiers, dont la compilation de deux. Un fichier latex initial contenant la page de présentation de ce contrôle, située au début du rapport. Un fichier R, contenant l'ensemble du code qui effectue la sélection des erreurs, qu'il faut compiler. Un autre fichier Latex qui, après compilation, émet le rapport final au format pdf .

Dans le premier fichier il y a l'initialisation du fichier latex final. Ce fichier est inscrit au fichier latex final par le biais du fichier R. Cette copie se fait grâce la fonction `write.table` prédéfinie dans R.
On exécute la séquence suivante qui copie le contenu du fichier latex initial :

```
rapport_ini<-read.table("Rapport_initial.tex", sep=";")
```

On l'inscrit ensuite dans le fichier latex final comme ceci :

```
write.table(rapport_ini, "Rapport_final.tex", sep = ";")
```

Ceci initialise le rapport final avec toutes les spécifications voulues (le style du rapport, la langue utilisé etc) qui sont inscrites dans le fichier initial.

Après l'initialisation, suit l'ajout un à un des tableaux dans ce rapport.
Le fichier R analyse la base de données, effectue les contrôles de cohérence et crée des tables qui contiennent les informations contrôlées. Ces tables sont créées en sélectionnant dans la base les valeurs incorrectes, par exemple les patients qui ont un champ non complété pour leur âge :

```
Base.temp<-Base[is.na(Base$ageR),]
```

On créer ensuite des variables, qui sont la syntaxe même des tableaux dans latex. On simule la création d'un tableau latex dans une variable R et on copie ensuite cette variable dans le fichier latex final de la même manière que pour le fichier initial :

```
res <- data.frame ( V1 = c( "\\begin{table}[ht]", "\\begin{center}"
   , "\\begin{tabular}{c}", "\\hline", "Numéro du patient \\\\"
   , "\\hline", paste(Base.temp$num_patient, "\\\\")
   , "\\hline", "\\end{tabular}"
   , paste ("\\caption{"Champs incomplété pour l'âge du receveur"}",sep = " ")
   , "\\end{center}" )  )
write.table(res, "Rapport_final.tex", append = TRUE, sep = ";")
```

Et cela pour chaque tableau ce qui complète le fichier latex copie après copie.

Après toute l'analyse terminée on obtient un fichier latex final, après compilation de ce fichier on a le rapport au format Pdf.

Cette approche est quelque peu instable. On se sert de la copie d'une variable dans un fichier, offerte par le langage R, pour simuler des codes latex créant des tableaux. A la moindre erreur de syntaxe on ne génére pas de tableau et le processus est très lourd.

3.2.2 La deuxième approche

Lors des recherches que j'ai entreprises pour comprendre les deux langages, j'ai découvert de nouveaux packages qui ont changé l'approche du projet. Il existe deux packages, un pour R et un autre pour Latex, qui réalisent une jonction entre ces deux langages. Le package `xtable` , destiné au langage R, traduit un objet R en un objet xtable. Le package `Sweave` , pour Latex, permet la prise en charge et l'affichage des tableaux générés par `xtable` .
Ces packages permettent, et nécessitent, la fusion du fichier R et du fichier Latex. Un fichier d'extension `Rnw` contient à la fois le code latex et le code R. Ce fichier est un fichier contenant par défaut du code latex, pour écrire du code R on le place entre deux balises. Grâce à la découverte de ces deux packages le projet se résume à un fichier et deux compilations.
Le même test que celui présenté au-dessus, se réalise maintenant de la manière suivante :

```
<<results=tex,echo=FALSE>>=
Base.temp <- Base[is.na(Base$ageR),]    #sélection des erreurs
```

```
rest <- data.frame(Base.temp$num.patient) #création de la table à afficher
colnames(rest) <- c("Numero (dossier) Patient")
#transcription en objet xtable de la table
xtable(rest, align="c|c", caption="Age est incompléte", digits=2)
@
```

Cela sélectionne et insère un tableau, indiquant les patients ayant leur valeur sur l'âge incomplétée, dans le rapport.

Comprendre comment bien utiliser ces extensions a été complexe. La documentation est très réduite et lorsqu'elle existe elle est en anglais ce qui ne rend pas la tache aisée.
Cet outil est très efficace c'est pourquoi nous avons décidé de recommencer le travail déjà réalisé avec ces nouvelles extensions, le résultat final en étant amélioré.

3.3 Généralités sur le contrôle

Les contrôles se font sur une table contenant les items choisis de tous les centres. Le fichier contenant ces informations est donc dans un premier temps importé. DIVAT assure la validité de ses données à partir de 1990 pour les centres hospitaliers de Nantes, Paris et Nancy et depuis 2003 pour les trois autres centres. Nous réduisons les données en excluant les dossiers antérieurs.

Les données sont toutes renommées, s'il y a un jour un changement dans le nom des items dans DIVAT le contrôle n'en sera pas affecté. Après se rennommage il est nécessaire de redonner aux variables leur nature originelle. L'export des données dans un fichier texte a transformé toutes les variables en chaînes de caractères. Nous rendons leur type aux variables, tout en renseignant les valeurs manquantes (exportées sous la forme « »).

Les rapports sont émis par centre. Pour connaître le centre dans lequel nous devons effectuer les tests, nous importons un fichier qui contient le nom du centre.
Avec ces fichiers nous sélectionnons la partie des données qui nous intéresse.

Le même processus est effectué pour la base concernant les rejets. Un patient peut faire plusieurs rejets, traiter les données sur les rejets avec les autres aurait pu entrainer une duplication des informations.

Après ces initialisations effectuées nous pouvons nous attarder sur le contrôle à proprement parlé.
Lors de chaque contrôle une fonction a été définie afin de créer un tableau latex. Les fonctions `insert_valmanq, insert_valpb, insert_2valpb et insert_valsusp` forment des tableaux latex, d'une ou plusieurs colonnes.
Les fonctions génèrent des tableaux qui comportent toujours une colonne indiquant le numéro de dossier du patient (non pas son nom dans un souci de confidentialité et d'efficacité lors de la recherche des patients, un patient pouvant avoir plusieurs greffes). Dans la partie concernant les valeurs manquantes seule cette colonne est présente.

Dans la deuxième partie nous ajoutons une colonne qui indique la valeur incorrecte, ou deux colonnes si c'est un problème d'incohérence entre deux données.
La troisième partie exporte des tableaux de taille variable, une colonne par variable utilisée dans la modélisation statistique.

Pour ne pas demander un travail trop conséquent aux ARC, si un contrôle trouve beaucoup d'erreurs nous limitons la table générée à 40 patients. Dans cette hypothèse là, pour informer l'ARC, le tableau est terminé par la mention « à suivre... ».
Lors de notre première approche, nous informions l'ARC par une ligne d'alerte « Attention il y a plus de 40 individus concerné par ce contrôle ». Ce message est plus explicite pour l'assistant que l'ajout de la ligne « à suivre... »dans les tableaux. Mais nous ne pouvons pas procéder de manière identique dans notre seconde approche. Les blocs de langage R ne permettent pas de créer des variables globales pouvant être reprises dans les parties Latex. Nous ne pouvons pas initialiser une variable booléenne, lorsque nous testons s'il y a plus de 40 patients, et afficher une ligne supplémentaire en dessous du tableau lorsque c'est nécessaire.

Il est possible que les tests ne retournent aucun tableau, cela signifie que la base de données est complète et sans erreur, ce qui est à terme espéré. Nous avons envisagé dans la première approche de le signaler aux ARC avec une ligne « Il n'y a pour cette partie aucune valeur incohérente dans la base ». Mais pour les mêmes raisons que précédemment nous n'avons pu dans notre seconde approche afficher ce message.

Nous avons inclu dans le rapport une page d'explication sur comment faire bon usage de ce rapport. Ceci pour expliquer, entre autres, aux assistants de recherche clinique les deux événements précédents.

3.4 Le contrôle des valeurs manquantes

Ce contrôle s'effectue sur un nombre d'items restreint, on ne va pas contrôler les 250 champs de la base mais uniquement ceux qui sont indispensables pour les médecins et les chercheurs. Nous contrôlons environ trente items, qui vont des données de bases du patient (son âge, son poids, sa taille, le type de greffe, etc), à ses données immunologiques les plus précises (ses anti-corps et anti-gène, etc), mais aussi les informations du donneur. Voici la liste de l'ensemble des items testés :

1. Age du receveur.
2. Age du donneur.
3. Date de naissance du receveur.
4. Date de naissance du donneur.
5. Sexe du receveur.
6. Sexe du donneur.
7. Taille du receveur.
8. Poids du receveur.
9. Date d'inscription sur les listes de pré-greffe.
10. Numéro de greffe.
11. Type de greffe.
12. Date de greffe.
13. Relation receveur-donneur.
14. Temps d'ischémie froide (temps entre le prélèvement et l'implantation du greffon)
15. Incompatibilité HLA.
16. Localisation de l'implantation du greffon.
17. Anti-corps CMV du receveur.
18. Anti-corps EBV du receveur.
19. Anti-corps HCV du receveur.
20. Anti-corps HIV du receveur.
21. Anti-gène BK du receveur.
22. HIV du donneur.
23. EBV du donneur.
24. Anti-gène HBS du donneur.
25. Anti-corps HBS du donneur.
26. HCV du donneur.
27. Anti-corps HBS du donneur.
28. CMV du donneur.
29. Créatinine du donneur.
30. Groupe sanguin du receveur.

31. Groupe sanguin du donneur.
32. Traitement d'induction.

Ces items ont été déterminés en accord avec le médecin Magali Giral et le statisticien Yohann Foucher. Certains sont nécessaires pour les études cliniques et d'autres sont indispensables médicalement.
Comme nous l'avons exposé auparavant les données qui sont toujours analysées dans les études épidémiologiques ne doivent pas contenir de valeur manquante pour ne pas réduire inutilement le nombre d'individus.
Certaines données sont essentielles pour la connaissance médicale du patient, il ne doit pas y avoir de doute sur la raison du manque d'information de certains champs. Les valeurs manquantes par oubli seront ainsi complétées.

Grâce au code couleur qui a été mis en place (que nous avons déjà détaillé, cf 2.2.1) à l'extraction des données, nous avons des données qui ont des valeurs, des données qui ont les valeurs par défaut et des données qui n'ont pas de valeur. Les données qui n'ont pas de valeur ne devraient pas exister. En effet, les données sans valeur sont celles qui ont une boule rouge et il ne devrait pas rester de boule rouge après complétion du dossier. Soit la valeur est connue et on la renseigne (boule verte), soit elle est introuvable et on le précise (boule violette) ce qui entraîne à l'extraction les valeurs par défaut.
Ce contrôle informe donc aux ARC les données non complétées.
Grâce aux transformations effectuées en début de programme (qui retranscrivent les données vides, représenté par « », en données manquantes NA) nous n'avons qu'à sélectionner les valeurs manquantes, ce qui se fait de la manière suivante :

```
Base.temp <- Base[is.na(Base$ageR),]
```

La variable `Base.temp` contient maintenant l'ensemble des valeurs manquantes de l'item « Age du receveur ». On appelle la fonction qui ajoute le contenu d'une table dans un objet xtable, ce qui permet l'écriture sur le rapport d'un tableau contenant l'ensemble des individus concernés :

```
insert_valmanq("Patients dont la donnée sur l'âge du receveur est incomplète.")
```

Ce qui produit le tableau suivant :

	Numéro (dossier) Patient
1	3176
2	1828

Tab1 : Patients dont la donnée sur l'âge du receveur est incomplète.

Le programme contient autant de tests, similaire à celui présenté ici, que de variables.
Cette partie va exporter dans le rapport final un tableau par item vérifié. Si un item n'a aucune valeur manquante il n'y aura bien entendu aucun tableau de présent. Le fichier final pourra contenir jusqu'à 32 tableaux en première partie.

3.5 Le contrôle des valeurs incorrectes

Cette partie est la partie centrale du contrôle, c'est ici que nous effectuons les tests de cohérence des variables. Le fichier exporté au début du programme ne contient pas les données sur les rejets. Si nous avions inséré les informations sur les rejets nous aurions pu avoir plusieurs lignes par patient, car un patient peut faire plusieurs rejets. Nous importons ici un nouveau fichier qui contient les informations sur les rejets, ce qui permettra de tester l'exactitude des données sur les rejets.
Nous testons dans un premier temps, lorsque cela est possible, que les valeurs sont bien dans les normes médicales. Suite ma demande de quantifier certaines variables, Magali Giral a donnée des bornes assez larges pour que les variables ne les dépassent en aucun cas, si elles sont correctes.
Nous procédons comme pour les valeurs manquantes à la sélection des variables ne rentrant pas dans ces normes médicales.

```
  Base.temp <- Base[(Base$ageR<10| Base$ageR>90)
    & !is.na(Base$ageR) & Base$ageR!= -99
   ,]
  insert_valpb(Base.temp$ageR, "Patients dont l'âge semble incorrect."
, "Age receveur")
```

Nous aurons ici déterminé et retourné les patients qui n'ont pas un âge correct. La base ne contient pas d'enfants (ces patients sont dans la base DIVAT-pédiatrie) et les centres ne greffent pas de patients de plus de 90ans, ces âges ne peuvent donc pas être réellement ceux des patients.

	Numéro (dossier) Patient	Age du receveur
1	2930	9
2	2057	3
3	3379	7
4	3552	6

Tab32 : Patients dont la donnée sur l'âge du receveur est incorrecte.

Nous effectuons ensuite des tests de cohérence entre les données d'un même patient. Nous testons les cohérences sur les données personnelles, par exemple voir si l'âge à la greffe correspond bien à la date de naissance, si le poids est cohérent avec la taille, etc... . Nous nous intéressons ensuite aux dates présentes dans le dossier, elles doivent correspondre au déroulement chronologique de la greffe (date de naissance, puis date d'inscription sur les liste de pré-greffe, puis date de prélèvement du greffon, puis date de greffe, etc...).
Les informations biologiques sont contrôlées pour déceler des items cliniquement incorrects.
Nous vérifions la concordance des données du receveur et du donneur.

On ne peut pas déterminer quel est l'item qui pose problème, on indique dans le rapport les deux variables qui ne sont pas cohérentes, les ARC vérifient les deux et déterminent celle qui est incorrecte.
Le code est similaire au précédent seul le tableau retourné change car on indique deux variables problématiques à vérifier et non plus une :

```
Base.temp <- Base[Base$taille<130 & Base$ageR>10
    & !is.na(Base$taille) & !is.na(Base$ageR) & Base$taille!= -99
   ,]
```

Cette partie a demandé une recherche afin de déterminer tous les croisements possibles entre les données, ceci à bien entendu été validé par le médecin Magali Giral.

L'ensemble de ces croisements représentent 49 tests. Lors de leur réalisation quelques soucis de programmation ont pût être observés.
Dans le logiciel DIVAT certain items sont remplis grâce à des menus déroulants. Extérieurement des champs sont des chaînes des caractères, par exemple la relation receveur-donneur (cadavre, mère-père, frère-soeur, etc) mais en réalité ces champs sont codés de manière numérique (2 pour cadavre, 3 pour mère-père, etc). Lors de ces tests il a été nécessaire d'obtenir les descriptions des variables champ par champ.
Les dates dans le logiciel R peuvent être traitées de différentes manières. Pour pouvoir comparer ses dates entre elles, elles doivent être au format « Julian Date ». Ce format transforme les dates en nombre de jours séparant la date et le 01/01/1960 (si la date est antérieure à 1960 elle est représentée par un nombre négatif). Lors des transformations nous avons détecté beaucoup d'anomalies ce qui nous a interpellé. Ces anomalies étaient dues au format d'exportation des dates depuis DIVAT. L'exportation compressée de la base réduit l'année des dates à deux chiffres ce qui entraîne une confusion entre les années du 20ème siècle et du 21ème lors de la transformation au format Julian Date.

Ces tests cernent l'ensemble des valeurs essentielles de DIVAT. Avoir effectué ces contrôles permet de détecter les erreurs qui peuvent être présentes dans la base et ainsi les corriger. C'est ce travail qui valide la fiabilité de la base.

3.6 Le contrôle des valeurs suspectes

Dans ce contrôle nous ne garantissons pas l'exactitude des erreurs retournées, d'où l'appellation « valeurs suspectes ». Nous recentrons ici les tests sur cinq variables majeures de la base : l'âge du receveur, l'âge du donneur, le poids, la taille et le temps d'ischémie froide (intervalle entre le prélèvement et l'implantation du greffon).
Nous procédons à une approche statistique des données, nous créons un modèle statistique qui permet d'approcher au mieux les variations des variables, citées ci-dessus, et nous retournons celles qui s'éloignent le plus du modèle.

Ce travail demande une connaissance des statistiques plus poussée que celle que les cours de Probabilité et Statistique, délivrés durant licence, offrent. Lors de ces trois mois il m'a fallut comprendre des notions nouvelles. Assister à des cours destinés aux étudiants de l'école doctorale de Biologie-Santé, m'a beaucoup apporté. Ces cours, dispensés par Yohann Foucher et Jean-Benoit Hardouin, portaient sur les modèles linéaires, entre autres les régressions linéaires simple et multi-variées. L'application au contrôle de cohérence m'a permis d'assimiler ce cours et de comprendre en profondeur les modèles statistiques jusque là inconnus. Cette troisième partie m'a permis de découvrir les statistiques, leur utilité et leurs applications.

Nous tentons de modéliser les variations de cinq variables, pour cela nous les approchons grâce à 21 covariables. Les covariables sont des variables qui ont un lien avec les cinq variables à approcher.

3.6.1 La transformation des données

Nous avons des covariables quantitatives (comme le nombre de greffes du patient, le nombre de dialyses post-greffe), mais il y a aussi des variables qualitatives. Ces variables qualitatives doivent alors être transformées afin qu'elles puissent être insérées dans le modèle linéaire.
Les variables qualitatives sont des variables qui peuvent être nominales (comme le sexe) ou ordinales qui désignent le rang ou la préférence. Dans notre étude nous n'avons que des variables qualitatives nominales. Pour pouvoir interpréter ces variables nous les transformons en variables quantitatives.
Quantifier ou coder une variable qualitative c'est associer à chacune de ses modalités un nombre réel et ainsi transformer la variable qualitative en une variable quantitative. On peut transformer ces variables grâce à un disjonctif complet, cette transformation est celle qui se prête le mieux aux calculs statistiques. La forme disjonctive complète est obtenue en définissant une variable indicatrice pour chacune des modalités ; par exemple `H` est la variable indicatrice de la modalité « Homme ».
H = 1 si l'individu est un homme
H = 0 sinon
Plus généralement, lorsqu'une variable qualitative a n valeurs possibles, on peut la transformer en n-1 variables quantitatives.
Exemple : la variable indicatrice du centre, dans lequel le patient est suivi, a six valeurs possible : Nantes, Necker, Nancy, Montpellier, Toulouse et Lyon. Cette variable centre peut être modélisé avec cinq variables, centreL, centreT, centreM, centreNe, centreNy.

Centre	centreL	centreT	centreM	centreNe	centreNy
Nantes	0	0	0	0	0
Nancy	0	0	0	0	1
Necker	0	0	0	1	0
Montpellier	0	0	1	0	0
Toulouse	0	1	0	0	0
Lyon	1	0	0	0	0

Ces cinq variables centre indiqueront dans quel centre se situent les individus. Si la variable centreM a pour valeur 1 on est alors dans le centre hospitalier de Montpellier sinon on est dans un autre centre, de même pour les quatre autres variables de centre. Si toutes les variables centre sont a zéro on est alors dans le centre, choisi comme étant le centre de référence, de Nantes.
Nous avons effectué de tels disjonctifs pour l'ensemble des variables qui le nécessitaient. Parmi ces variables nous avons tous les antécédents (cardiaques, vasculaires, digestifs, métaboliques, urologiques, néoplasiques), mais aussi le type de greffe (une variable qui vaut 1 lorsque c'est une greffe de rein, zéro sinon), les maladies initiales, etc.

Une fois les variables transformées et exploitables nous pouvons commencer l'approche statistique de nos variables majeures.

3.6.2 La régression linéaire uni-variée

Nous commencons par approcher nos cinq variables de façon uni-variée.
On détermine pour chaque variable majeure sa corrélation avec chacune des covariables. Prenons l'exemple de l'âge du receveur et l'âge du donneur, ce sont des variables qui varient généralement ensemble et dans le même sens. Le lien entre ces deux variables n'est cependant pas un lien absolu : certains patients jeunes recoivent parfois l'organe d'un donneur plus âgé qu'un autre patient plus vieux.
Nous devons chercher une droite qui approche au plus près le nuage de point suivant :

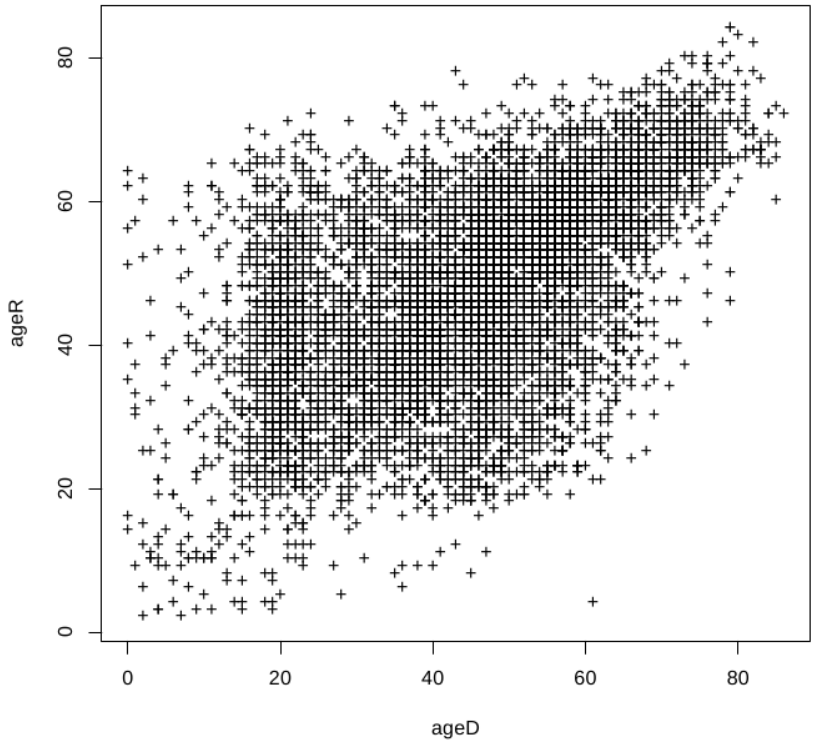

FIGURE 3.1 – Variation de l'âge du receveur en fonction de l'âge du donneur.

Pour mesurer l'intensité de la relation entre ces deux variables on effectue une régression linéaire.
On note x et y les deux variables, soit x_i (respectivement y_i) la valeur prise par l'individu i pour la variable x (respectivement y).
Déterminons s'il existe une relation linéaire vérifiée par les deux variables, c'est à dire s'il existe deux réels a et b tels que :

$$y_i = a + bx_i + e_i \qquad \text{pour i = 1,...,n}$$

où a est l'ordonnée à l'origine (valeur de y_i moyenne quand $x_i = 0$, significative que si la valeur zéro est cohérente pour x)

b est la pente (changement moyen de y_i quand x_i augmente d'une unité).
e_i est le résidu (différence entre la valeur prédicte et la valeur observée).

Notre objectif est de trouver la meilleure droite possible pour le nuage de points observé. Or, la relation entre x et y sera d'autant plus proche d'une relation linéaire exacte que les résidus e seront petits. Algébriquement on détermine a et b selon le critère des moindres carrés, c'est à dire de façon à ce que $\sum_{i=1}^{i=n} e_i^2$ ait une valeur minimale.

L'approximation de la corrélation entre l'âge et le poids se fait avec la régression linéaire suivante : $AgeReceveur = a + b * AgeDonneur + e$
Graphiquement cette régression se modélise par la droite suivante :

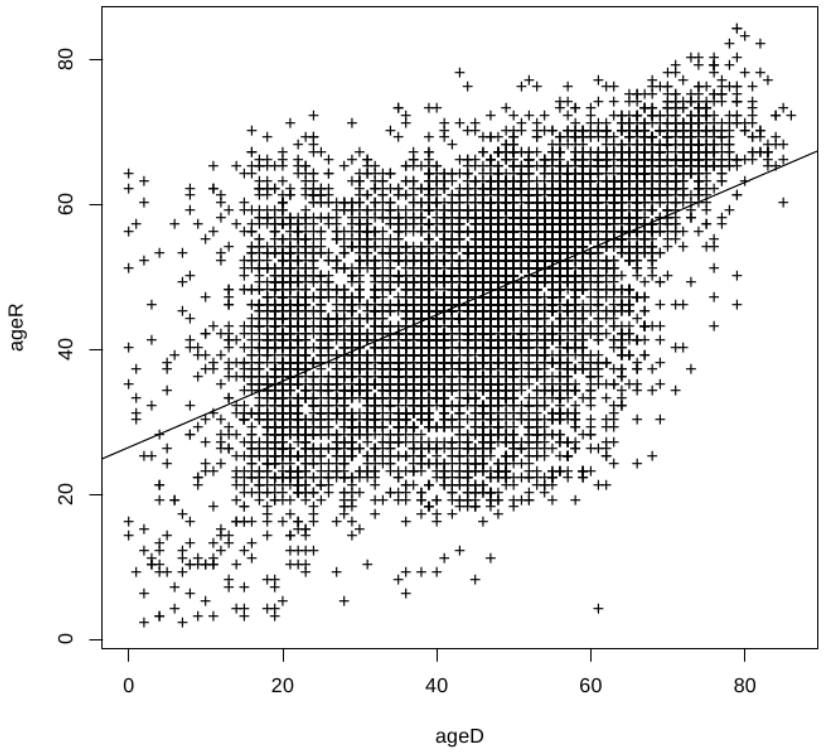

FIGURE 3.2 – Variation de l'âge du receveur en fonction de l'âge du donneur, modélisé par la doirte de régression linéaire.

Pour définir si la droite approche bien notre nuage on utilise la méthode du test de Student. Elle a deux hypothèses : H_0 : b = 0 et H_1 : b $\neq$ 0, avec b le paramètre représentant la pente de la droite de régression.
La méthode calcule la statistique de test : $t = \frac{b}{s(b)}$ ($s(b)$ est l'écart type de b) qui suit une loi de Student à N-2 degrés de liberté.
On définit une région non-critique : $[-t_{\alpha,N-2}; +t_{\alpha,N-2}]$ (avec t lu dans la table de Student).
Si t appartient à la région critique, on rejette H_0, sinon on ne peut pas rejeter H_0.

Dans la pratique pour connaître les données nécessaires à la décision du rejet ou non de l'hypothèse H_O on utilise la fonction *summary* du langage R. Elle nous fournit les

données suivantes :

	Valeur estimée	Ecart type	Test de student	p-value
Intercept	26.509953	0.397374	66.71	<2e-16
AgeDonneur	0.457583	0.008438	54.23	<2e-16

R^2 : 0.2853 ; R^2 ajusté : 0.2852

La p-value est la probabilité de conclure à tort que la covariable est significative.
R^2 représente la proportion de variation de y expliquée par x (R est le coefficient de corrélation).
Dans notre approche uni-variée nous gardons les covariables qui ont une probabilité de conclure à tort (la p-value) inférieure à 0.25, ce qui est vérifié dans notre exemple.
Nous devons maintenant vérifier l'hypothèse de linéarité : l'analyse du graphique précédent nous montre que notre approche linéaire n'est peut être pas la bien venue, de plus voyons si on peut augmenter la valeur du coefficient de corrélation. Essayons d'approcher notre échantillon autrement que linéairement.
Tentons une approche polynomiale : $AgeReceveur = a + b * AgeDonneur + c * AgeDonneur^2 + AgeDonneur^3 + e$, on a alors le graphique suivant (la courbe est l'approche polynomiale et la droite est l'approche linéaire) :

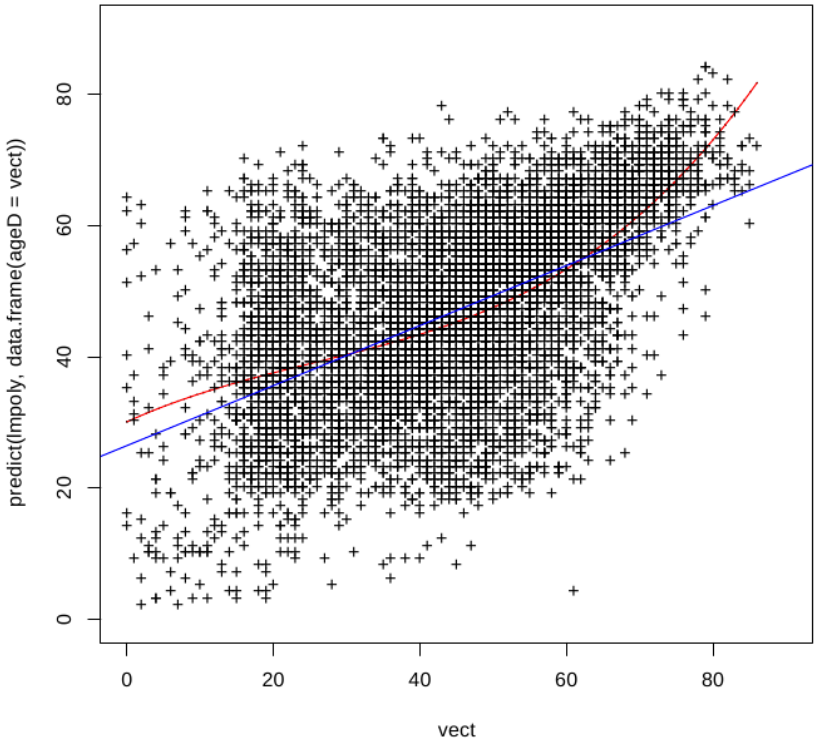

FIGURE 3.3 – Variation de l'âge du receveur en fonction de l'âge du donneur et sa courbe approchant polynomilaement ses variations.

On a, avec cette approche polynomiale, les données suivantes :

	Valeur estimée	Ecart type	Test de student	p-value
Intercept	3.014e+01	1.351e+00	22.29	< 2e-16
AgeDonneur	5.087e-01	1.094e-01	4.650	3.38e-06
AgeDonneur^2	-9.129e-03	2.692e-03	-3.392	0.000699
AgeDonneur^3	1.186e-04	2.035e-05	5.828	5.83e-09

R^2 : 0.3088 ; R^2 ajusté : 0.3085

L'approche polynomiale est plus appropriée pour ces deux variables, elle approche mieux le nuage de points et le R^2 est légèrement plus élevé. On explique près de 31% des variations de l'âge du receveur par l'âge de son donneur contre 28% avec une approche linéaire.

On effectue cette analyse pour l'ensemble des covariables. On exclut celles qui ne sont pas significatives. On remarque que l'approche uni-variée, dans l'ensemble, explique peu les variations des variables majeurs, c'est pourquoi on passe à une approche multi-variée.

3.6.3 La régression linéaire multi-variée

La régression linéaire multi-variée est semblable à l'approche uni-variée. Pour approcher une variable par n covariables explicatives, on a le modèle suivant :

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + ... + b_n x_{ni} + e_i$$

Le paramètre b_1 représente la variation de y quand x_1 augmente d'une unité, b_2 représente la variation de y quand x_2 augmente d'une unité, ainsi de suite.
Pour continuer avec notre exemple la régression linéaire multi-variée est :

$$\begin{gathered} AgeReceveur = \\ a + b_1 ante.cancer + b_2 poids + b_3 poids^2 + b_4 numero.greffe + b_5 maladie.initiale + \\ b_6 nb.dialyse.post.greffe + b_7 ageDonneur + b_8 ageDonneur^2 + b_9 ageDonneur^3 + \\ b_{10} relationDonneur.Receveur + b_{11} antiCorps.cmv + b_{12} tps.ischémie + b_{13} SexeReceveur + \\ b_{14} type.greffe + b_{15} incompatibilitéABDR + b_{16} SexeDonneur + b_{17} antiCorps.hcv + \\ b_{18} antiGène.bk + b_{19} antécédentCardio + b_{20} antécédentHTA + b_{21} antécédentMéta + \\ b_{22} antécédentUro + b_{23} antécédentNeo + b_{24}(centreNe + centreT + centreNy + centreM) + e \end{gathered}$$

L'ensemble des covariables présentes dans le modèle ont été déclarées significatives avec l'analyse uni-variée précédente.
Il est très difficile d'analyser le modèle multi-varié graphiquement, c'est un modèle qui a autant de dimensions que de covariables, dans notre exemple le graphique aurait 24 dimensions. On utilise pour déterminer si l'approche est correcte les paramètres formels.

	Valeur estimée	Ecart type	Test de Student	p-value
(Intercept)	19.3455	8.3284	2.32	0.0203
centre Necker	4.1977	3.7682	1.11	0.2655
centre Toulouse	1.8526	4.6096	0.40	0.6878
centre Montpellier	7.9021	11.3901	0.69	0.4879
sexeReceveur	-2.7421	0.6546	-4.19	2.96e-05
poids	0.2802	0.0245	11.45	$<$ 2e-16
numero.greffe	-1.9659	0.7034	-2.79	0.0053
maladie.initiale	-1.6568	0.6400	-2.59	0.0097
nb.dialyse.post-greffe	-0.3659	0.1329	-2.75	0.0060
tps.ischémie	0.0018	0.0005	3.43	0.0006
type.greffe	1.7141	1.4486	1.18	0.2369
incompatibilité.ABDR	0.2134	0.2173	0.98	0.3263
sexeDonneur	0.1182	0.6188	0.19	0.8485
ageDonneur	0.3161	0.0196	16.17	$<$ 2e-16
relationDonneur.Receveur	5.5148	1.5221	3.62	0.0003
antiCorps.cmv	2.3405	0.5847	4.00	6.56e-05
antiCorps.hcv	-0.2258	1.0781	-0.21	0.8341
antiGène.bk	5.2344	2.0181	2.59	0.0096
antécédent.HTA	-17.4175	8.1972	-2.12	0.0338
antécédent.cardio	6.5684	1.9584	3.35	0.0008
antécédent.meta	0.5594	1.2197	0.46	0.6465
antécédent.uro	1.3837	0.8617	1.61	0.1085
antécédent.neo	5.7914	1.2906	4.49	7.75e-06

R^2 : 0.4048 ; R^2 ajusté : 0.4034

On écarte une à une les covariables qui n'ont pas une p-value inférieure à 0.05, en commençant par la plus élevée (seule les variables des centres ne sont pas testées et restent d'office dans le modèle). On réduit notre modèle aux variables très significatives. Dans notre exemple, cette procédure nous réduit le modèle à 12 covariables significatives :

$$AgeReceveur = a + b_1 ante.cancer + b_2 poids + b_3 poids^2 + b_4 numero.greffe + b_5 maladie.initiale + b_6 nb.dialyse.post.greffe + b_7 ageDonneur + b_8 ageDonneur^2 + b_{10} relationDonneur.Receveur + b_{11} antiCorps.cmv + b_{12} tps.ischémie + b_{13}(centreNe + centreT + centreNy + centreM) + e$$

	Valeur estimée	Ecart type	Test de Student	p-value
(Intercept)	1.6630	2.6574	0.63	0.313743
antecédent.neo	5.5484	0.5517	10.06	$<$ 2e-16
poids	0.8460	0.0652	12.97	$<$ 2e-16
poids^2	-0.0048	0.0005	-10.42	$<$ 2e-16
numero.greffe	-1.3690	0.2768	-4.95	7.75e-070
maladie.initiale	-2.2839	0.2767	-8.25	$<$ 2e-16
nb.dialyse.post-greffe	-0.1149	0.0551	-2.08	0.035035
ageDonneur	-0.1623	0.1133	-1.43	0.035035
ageDonneur^2	0.0055	0.0027	1.98	5.50e-10
relationDonneur.Receveur	5.4537	0.5565	9.80	$<$ 2e-16
antiCorps.cmv	3.1515	0.2697	11.69	$<$ 2e-16
tps.ischemie	0.0009	0.0002	3.62	0.0003
centreNe	-3.0737	0.3576	-8.60	$<$ 2e-16
centreT	0.3912	0.5008	0.78	0.4348
centreNy	-0.2552	0.3548	-0.72	0.4721
centreM	-0.7760	0.4830	-1.61	0.1081

R^2 : 0.4048 ; R^2 ajusté : 0.4034

Les covariables peuvent parfois interagir entre elles. Ces interactions ajustent plus finement le modèle linéaire. Mathématiquement une interaction est représentée par le produit de deux covariables. On teste les interactions possibles entre les covariables. Si elles sont significative s on les ajoute au modèle statistique.
Dans notre exemple, il existe deux interactions : l'âge du donneur avec les centres et le temps d'ischémie froide avec les centres. Notre modèle devient :

$$AgeReceveur = a + b_1 ante.cancer + b_2 poids + b_3 poids^2 + b_4 numero.greffe + b_5 maladie.initiale + b_6 nb.dialyse.post.greffe + b_7(centreNe + centreNy) * ageDonneur + b_8 ageDonneur^2 + b_{10} relationDonneur.Receveur + b_{11} antiCorps.cmv + b_{12} tps.ischémie * centreNy + centreM + centreT + centreNe + e$$

	Valeur estimée	Ecart type	Test de Student	p-value
(Intercept)	1.5911	2.4082	0.66	0.5088
ante.neo	5.4575	0.5484	9.95	$<$ 2e-16
poids	0.8219	0.0649	12.67	$<$ 2e-16
poids^2	-0.0047	0.0005	-10.11	$<$ 2e-160
numero.greffe	-1.3707	0.2752	-4.98	7.02e-07
maladie.initiale	-2.3339	0.2754	-8.48	$<$ 2e-16
nb.dialyse.post-greffe	-0.1270	0.0548	-2.32	0.0205
centreNe	5.1970	1.0102	5.14	3.44e-06
centreNy	4.3965	1.1487	3.83	0.0001
ageDonneur	-0.2327	0.0384	-6.06	9.93e-10
ageDonneur^2	0.0079	0.0004	18.28	$<$ 2e-16
relationDonneur.Receveur	5.3438	0.5532	9.66	$<$ 2e-16
antiCorps.cmv	3.0647	0.2683	11.42	$<$ 2e-16
tps.ischemie	0.0012	0.0003	4.57	5.49e-06
centreT	0.3331	0.5003	0.67	0.5056
centreM	-1.0497	0.4822	-2.18	0.0295
centreNe :ageDonneur	-0.1760	0.0202	-8.72	$<$ 2e-16
centreNy :ageDonneur	-0.0721	0.0208	-3.47	0.000516
centreNy :tps.ischemie	-0.0012	0.0005	-2.43	0.0152

$R^2 : 0.4122\,;\ R^2 : 0.4105$

Les trois dernières lignes modélisent les interactions.
Notre variable *Age du receveur* est expliquée à 41% par notre modèle linéaire multi-varié.

Les résidus

Pour déterminer quels sont les individus suspects nous examinons les résidus de notre modèle statistique. Rappelons qu'un résidu est la distance qui sépare la valeur observée de la valeur prédite par le modèle statistique.

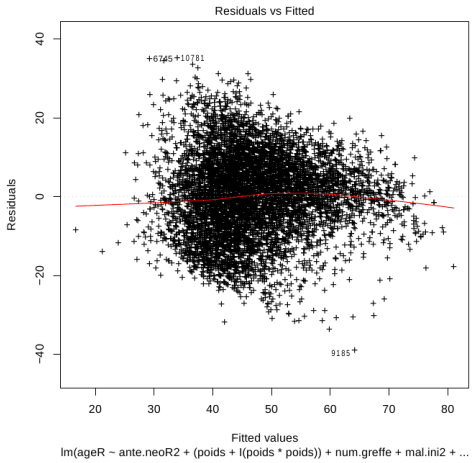

FIGURE 3.4 – Résidus du modèle statistique qui apporche la variable de l'âge du patient.

Nous calculons les résidus grâce à la formule de Jackknife. La méthode de Jackknife calcule le résidu d'un individu en l'excluant du modèle, pour ne pas biaiser le modèle si l'individu est trop influent et fau.
On sélectionne ensuite les résidus qui ont une probabilité inférieure au seuil fixé qu'il existe un individu plus éloigné que lui du modèle. Pour ne pas retourner des individus qui risquent d'être corrects, nous avons fixé le seuil à 1% sur l'ensemble des données de DIVAT (soit 16000 patients).

Individus suspects

Contrairement aux parties précédentes, les individus retournés ne sont que suspect. Si le modèle statistique n'approche pas assez bien les variations des items, les valeurs retournées peuvent être correctes.
Les ARC doivent contrôler l'ensemble des variables ayant servi à modéliser l'item contrôlé. Après avoir contrôlé l'ensemble de ses variables, si toutefois elles sont toutes correctes, les ARC valideront cet individu comme étant correct pour l'item analysé.

Pour faire la validation nous avons ajouté une colonne supplémentaire aux tableaux. Dans la dernière colonne se trouve un lien sur lequel les ARC cliqueront pour signaler que la valeur de la variable est correcte.

Dans le programme avant de retourner une table d'individus suspects, nous contrôlons d'abord si les individus trouvés n'ont pas déjà été validés comme correct par les assistants de recherche lors des dernières émissions du contrôle.

Le lien est un lien vers une application gérée par l'entreprise IDBC. A la fin de ce lien nous indiquons le numéro du dossier patient, l'item contrôlé et la valeur de l'item. L'application récupère ces informations et l'ajoute à la table des erreurs suspectes validées correctes (cette table est récupérée avant chaque contrôle).

	Num dossier	AgeR	Ante cancer	Poids	...	tps ischémie	Validation
1	1699	28	1	78	...	2160	*validation*
2	2765	73	1	62	...	2220	*validation*
3	3680	61	1	39	...	1395	*validation*
4	2060	64	1	61	...	1440	*validation*
5	1478	37	1	72.5	...	2467	*validation*
6	...	...	...	...	...	...	...

Tab52 : Patients qui ont un age suspect.

Résumé

1. Modèle uni-varié : $Var = a + b * coVar + e$
   -Test de la significativité de la covariable (p-value < 0.25) : exclusion du modèle des covariables non significatives.
   -Test de la linéarité des covariables significatives.
   -Ajustement polynomiale (ou exponentielle, logarithmique ...) des covariables non linéaires.

2. Modèle multi-varié : $Var = a + b_1 coVar_1 + b_2 coVar_2 + ... + b_n coVar_n + e$
   -Test de la significativité des covariables (p-value <0.05) : exclusion une à une des covariables non significatives.
   -Test des interactions entre les covariables pour l'amélioration du modèle.

3. Insertion dans le programme du modèle statistique élaboré :
   `lm1<-lm(ageR ~ ante.neo + poids + poids`^{2} `+ numero.greffe + maladie.initiale + nb.dialyse.post-greffe + ageDonneur*(centreNe + centreNy) + ageDonneur`^{2} `+ relationDonneur.Receveur + antiCorps.cmv + tps.ischémie*centreNy + centreT + centreNy + centreM, data=Base.temp)`

4. Calcule des résidus avec la formule de Jackknife :
   `p<-length(lm1$coefficients)+1`
   `n<-length(lm1$residuals)`
   `res.student<-rstudent(lm1)*`$\sqrt{\frac{n-p-1}{n-p-rstudent(lm1)^2}}$

5. Sélection des individus qui ont un résidu extrême :
   `Base.suspecte <-Base.temp[2*(1-pt(abs(res.student), n-p-1))<seuil,]`

6. Sélection des individus du centre analysé par le contrôle de cohérence en cours :
   `Base.suspecte <- Base.suspecte[Base.suspecte$centre == centre,]`

7. Création de la table des individus suspects qui n'ont pas déjà été validés correcte :
   `Base.suspecte <- Base.suspecte[Base.suspecte$num.patient != BaseValidee$num & Base.suspecte$ageR != BaseValidee$valeur & BaseValidee$item == "ageR" & !is.na(Base.suspect$num.patient + Base.suspecte$ageR) ,]`

8. Ajout de la dernière colonne contenant le lien de validation avec toutes les informations sur l'item et sa valeur :
   `Base.suspecte$lien <- paste(« $ href {http ://application.idbc.fr/stat /operation2.aspx ?choix=2009&erreur= », Base.suspecte$num.patient, « ageR », Base.suspecte$ageR, « }{ {validation} }$ »)`

9. Inscription du tableau dans le fichier latex :
   `x <- xtable(rest, caption = "Patients qui ont un age suspect.",)`
   `print(x, sanitize.text.function=function(x)x)`
   Pour afficher le lien nous utilisons l'option `sanitize.text.function=function(x)x` qui transmet le contenu des colonnes du tableau comme tel, sans les transformer en chaîne de caractères, pour que latex comprenne la dernière colonne comme étant des liens et non des caractères.

3.7 Les dates de perte de vue des patients

Dans cette partie nous ne contrôlons pas des erreurs. Nous informons les ARC des nouveaux patients déclarés perdu de vue.
Les patients sont automatiquement déclaré comme perdu de vue par le logiciel DIVAT si depuis deux ans aucune information n'a été insérée dans son dossier. Le logiciel renseigne dans la date de perdu de vue la dernière date présente dans le dossier du patient. Cet item est suivi d'un item qui valide cette date.
Nous contrôlons que les patients qui ont une date de perdu de vue, ont « True »comme valeur pour l'item de validation de cette date. Si des patients n'ont pas la variable de validation à « True », cela signifie que le patient a été déclaré perdu de vue récemment, il faut entamer des procédures pour confirmer cette perte de vue.
Nous sélectionnons les patients qui correspondent à cette description :

```
Base.temp <- Base[!is.na(Base$Dperdu.vu) & is.na(Base$valide.Dperdu.vu),]
```

Nous créons un tableau, de manière identique à ce qui a été fait avant, contenant tous les individus concernés.

Chapitre 4

Bilan du travail réalisé

4.1 Les objectifs de la mission

L'objectif initial de création d'un contrôle de cohérence automatisé est atteint. Il a été réalisé avec des outils différents de ceux utilisés lors du travail initié par Yohann Foucher mais l'approche adoptée le rend plus lisible et plus efficace.
En complément des trois contrôles, initialement prévu, a été ajouté une quatrième partie informant des nouveaux patients perdu de vue. Cette partie apporte un gain de temps non négligeable pour les ARC.

4.2 L'évolution du travail réalisé

Le contrôle sera très évolutif. Il prendra en compte les remarques des premiers utilisateurs, les ARC. Certains contrôles, qui auront pu être mal jugés, seront supprimés, d'autres contrôles, qui détectent beaucoup d'erreurs, pourront être mis en place directement à la saisie des dossiers. La découverte, grâce à des études cliniques, de nouvelles cause à effet pourrons donner lieu à l'ajout de nouveaux contrôles.
Les modèles statistiques, utilisés dans la troisième partie, pourront être mieux ajustés par un spécialiste de ce type de modélisation. Les individus suspects indiqués ici seront ainsi plus incorrects que suspects.

Dans la troisième partie les ARC peuvent valider les patients comme étant correct grâce à un lien. Lors de la présentation de mon travail devant l'équipe de bio-statistique du CHU, on nous a fait remarquer que nous pourrions appliquer les liens aux autres parties du rapport. Nous avons pensé que effectivement il serait bon de créer un lien dans les parties précédentes qui renverrait sur la page du patient. L'ajout de ce lien faciliterait le travail des ARC, ils n'auraient plus à trouver le dossier du patient par eux-mêmes, le lien les dirigerait directement sur le dossier, et pourquoi pas sur la page de l'item problématique. L'application est aujourd'hui possible, mais un futur remaniement de DIVAT en facilitera la réalisation, nous avons donc décidé de remettre à plus tard cette option.

Conclusion

Le contrôle de cohérence de la base de données médicales des patients ayant bénéficié d'une transplantation rénale, DIVAT, a été réalisé pour optimiser la fiabilité des données saisies.

La principale difficulté est qu'il s'agit d'une cohorte observationnelle, plus proche d'un dossier patient que d'un simple registre ou d'un essai thérapeutique (objectif précis et analyse statistique déterminée à l'avance). Pour pouvoir concilier suivi du patient et analyse statistique des données il a été mis en place un contrôle automatisé de cohérence des informations contenues dans DIVAT.

Les tests présents dans ce contrôle ont tous un même but, néanmoins, ils portent chacun sur des objectifs différents.
Le contrôle des valeurs manquantes vise à compléter la base de données, le contrôle des valeurs incorrectes vise à certifier les données contenues dans DIVAT et le contrôle des valeurs suspectes détectent des données potentiellement erronées. La dernière partie du contrôle sur les dates de perte de vue, a été créé pour faciliter le travail des assistants de recherche clinique.
Après quelques émissions du rapport des erreurs présentes dans DIVAT, la consultation du personnel hospitalier pourra amener à faire évoluer le contrôle de cohérence.

Glossaire

1. ARC : Assistant de Recherche Clinique.
2. BK : BK Polyomavirus, virus dont le génome est constitué d'ADN bicaténaire circulaire.
3. Coefficient de corrélation : Pourcentage représentant la proportion des variations d'une variable expliquées par une covariable dans un modèle statistique.
4. CMV : CytoMegalo Virus
5. Créatinine :indicateur de l'insuffisance rénale (plus elle est élevée plus le rein a des problèmes fonctionnel).
6. DIVAT : Données Informtisées VAlidées en Transplantation.
7. EBV : EpsteinBarr Virus.
8. HBS : Hépatite B.
9. HCV : Hépatite C.
10. Incompatibilité HLA : Marqueurs d'histocompatibilité (compatibilité organique entre le tissu d'un greffé et un greffon).
11. IDBC : Informatique de Données Biomédicales à la Carte.
12. INSERM :Institut Nationnal des Sciences Et de la Recherche Médicale.
13. ITERT : Institut de Tranplantation Et de Recherche en Transplantation.
14. Julian Date : format des date dans le langage R, nombre de jour entre le 1er janvier 1960 et la date.
15. Liste de pré-greffe : Liste d'attente nationnale recensant l'ensemble des personnes nécessitant une greffe.
16. NA : codage des valeurs manquantes pour le langage R.
17. Néphrologie : Spécialité de la médecine qui se consacre à l'étude de la physiologie et de la pathologie des reins.
18. p_value : Probabilité de conclure à tort qu'une covariable d'un modèle statistique est significative pour la variable qu'elle modélise.
19. Résidu : Terme qui n'est pas expliqué par les autres variables dans une régression statistique.
20. Temps d'ischémie froide : temps entre le prélèvement du greffon et son implantation.
21. Traitement d'induction : Premier traitement après la greffe qui permet d'éviter le rejet.
22. Urologie : Branche de la médecine qui se consacre à l'étude des maladies des voies urinaires et de l'appareil génital de l'homme.

Annexes

Annexe I : Interface de DIVAT

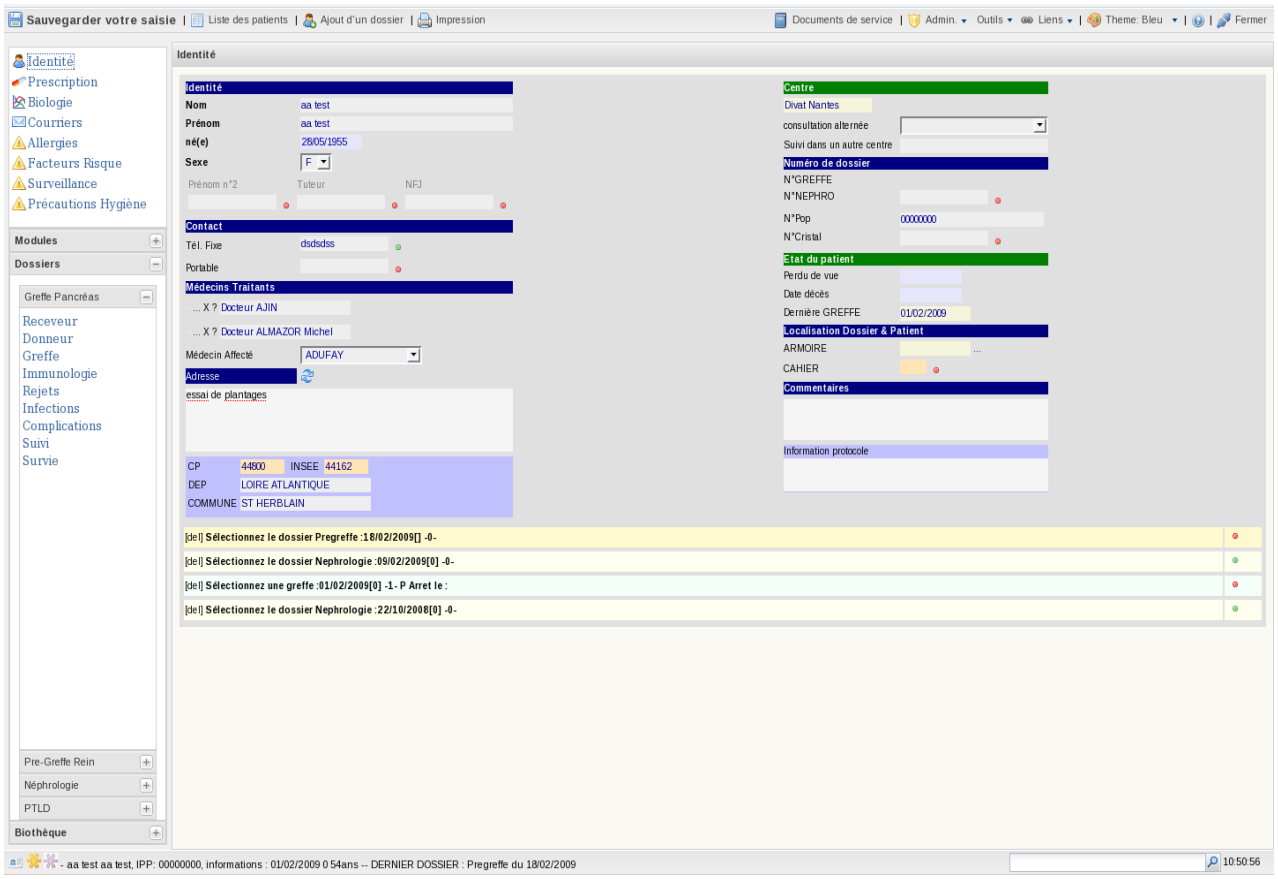

FIGURE 4.1 – Interface web de DIVAT

Annexe II : Interface d'extraction de DIVAT

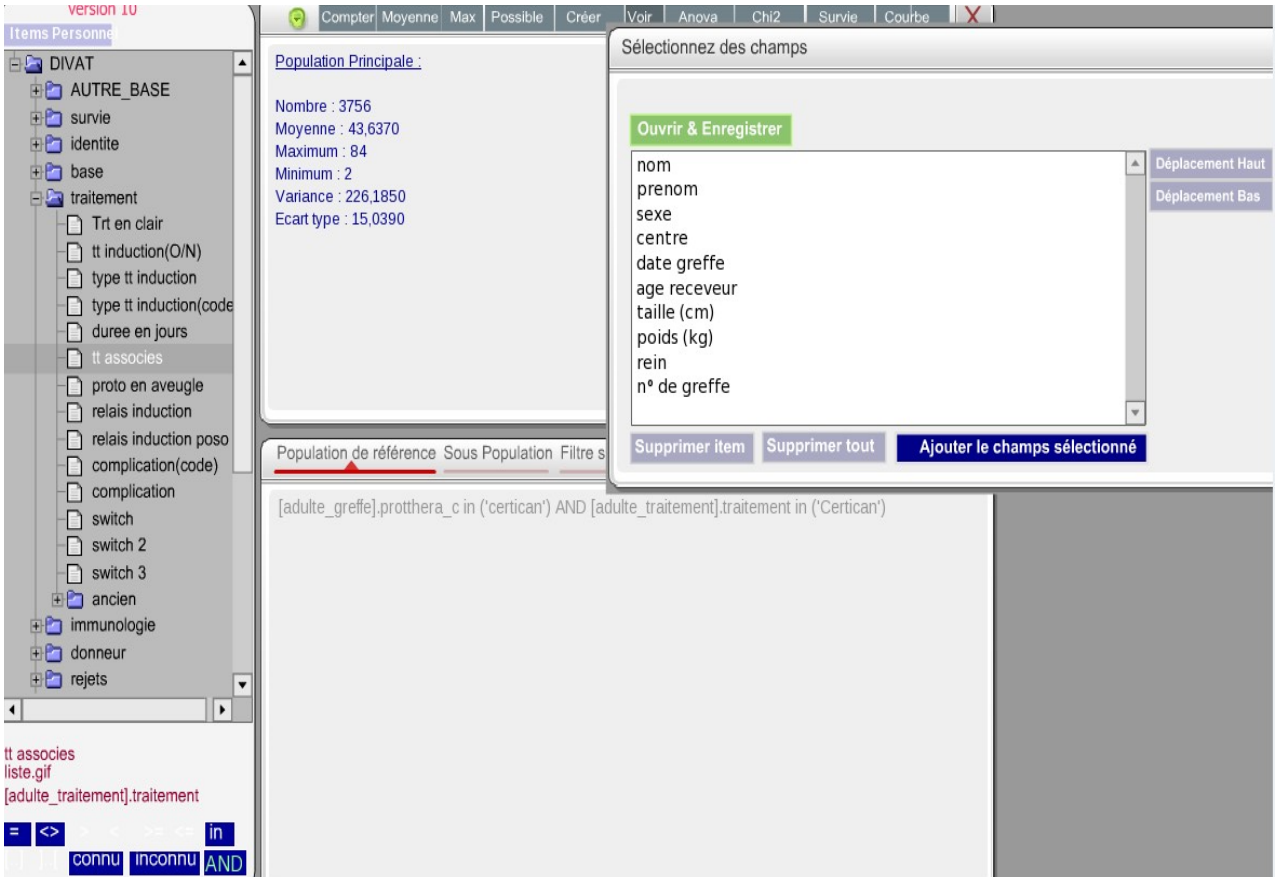

FIGURE 4.2 – Interface d'extraction des données de DIVAT