

|  |
| --- |
| INSERM |
| Rapport d’Alternance |
| Identification d’éventuelles associations entre les biomarqueurs sanguins et les marqueurs de la Maladie d’Alzheimer |

|  |
| --- |
| Geoffrey BOILAY  Alternance du 15/10/22 au 15/09/23  Département Statistique et Informatique Décisionnelle  IUT de Paris - Rives de Seine - Université de Paris  Année Universitaire 2022 – 2023 |



Remerciements

Je tiens tout d’abord à remercier la branche STID de l’IUT Paris Cité pour m’avoir permis de réaliser cette alternance en entreprise. L’écoute et l’aide dont j’ai bénéficié m’ont permis de trouver rapidement une alternance, dans le but d’affiner mon projet professionnel et d’obtenir ma licence professionnelle métiers du décisionnel et de la statistique, parcours data mining.

J’aimerais aussi remercier l’équipe EpiAgeing que j’ai accompagné durant ces trois mois et qui, par leur sympathie et leur bonne humeur, m’ont permis d’avoir une bonne expérience en entreprise de l’INSERM et tout particulièrement Madame Aline DUGRAVOT qui m’a épaulé et conseillé tout au long de mon alternance.

Cette alternance m’a permis d’affiner certaines pistes pour bâtir mon projet d’orientation professionnelle et signer l’aboutissement de cette année passée en licence professionnelle.

Remerciements spéciaux à mes relecteurs et correcteurs qui ont contribué, grâce à leurs conseils et recommandations, à l’élaboration et au bon déroulé de mon rapport d’alternance.

Table des matières

[1. INTRODUCTION 5](#_Toc143764796)

[2. CONTEXTE 6](#_Toc143764797)

[3. PRESENTATION DE L’ALTERNANCE 7](#_Toc143764798)

[3.1 Contexte 7](#_Toc143764799)

[3.2 Ma mission 7](#_Toc143764800)

[4. RÉALISATION 8](#_Toc143764801)

[4.1 Compte rendu : Analyses univariées et bivariées 8](#_Toc143764802)

[4.1.1 Analyses univariées 8](#_Toc143764803)

[4.1.2 Études d'association 13](#_Toc143764804)

[4.1.3 Conception du compte rendu 14](#_Toc143764805)

[4.1.4 Conclusion 14](#_Toc143764806)

[4.2 Compte rendu : Analyses centrées sur le cholestérol 14](#_Toc143764807)

[4.2.1 Fonctionnement du compte rendu 14](#_Toc143764808)

[4.2.2 Conception du compte rendu 16](#_Toc143764809)

[4.2.1 Conception du compte rendu 17](#_Toc143764810)

[4.3 Compte rendu : Analyses exploratoires sur le cholestérol 18](#_Toc143764811)

[4.3.1 Fonctionnement du compte rendu 18](#_Toc143764812)

[4.3.2 Conception du compte rendu 21](#_Toc143764813)

[8. TABLE DES ILLUSTRATIONS 23](#_Toc143764814)

[9.ANNEXES 23](#_Toc143764815)

# 1. INTRODUCTION

Afin de valider ma licence professionnelle, j’ai réalisé une alternance tout au long de ma formation dans une équipe de recherche INSERM, l’Institut National de la Santé et de la Recherche Médicale. Suite à mon stage réalisé précédemment chez eux, ma tutrice m’a offert l’opportunité d’effectuer mon alternance dans le même environnement.

L’alternance entre dans le cadre d’un projet de recherche d’une équipe INSERM dont la problématique est la suivante : Existe-t-il un lien entre les différents biomarqueurs sanguins (cholestérol, glucose, triglycérides…) et les protéines Abeta et Tau, marqueurs de la maladie d’Alzheimer ? Pour répondre à cette problématique, les investigateurs du projet ont préparés la mise en place d’une cohorte [[1]](#footnote-1)de patients dont un certain nombre de données sont recueillies lors du suivi médical. La particularité de cette cohorte est que chaque patient a effectué au minimum une ponction lombaire. Ma contribution dans ce projet est de ressortir dans un compte rendu les résultats des analyses de la base de données issue de cette cohorte afin de résumer les informations pour le chercheur à la tête du projet. Tout d’abord, j’ai pris connaissance du contexte du projet, je me suis approprié la base de données, puis nettoyer cette dernière afin de pouvoir l’analyser le plus justement possible. J’ai rédigé un premier rapport global résumant les analyses univariées et bivariées du jeu de données. Un deuxième rapport concentré sur l’analyse de l’effet du cholestérol sur les protéines marqueurs de la maladie d’Alzheimer. Un troisième et dernier rapport résumant les résultats des analyses exploratoires.

# 2. CONTEXTE

Le secteur d’activité de l’INSERM est comme son nom l’indique la recherche dans la santé et le domaine médical. L’INSERM est le seul organisme public français qui se dédie entièrement à la santé humaine. Il a été créé en 1964 par Raymond MARCELLIN à la suite de l’Institut National d’Hygiène (INH). C’est un établissement public à caractère scientifique et technologique qui dépend du ministère de la Santé et du ministère de la Recherche. Son Directeur Général est Gilles BLOCH.

Les 5 principales missions que s’est fixé l’INSERM sont :

* Tenir le gouvernementinformé de l’état sanitaire du pays et en orienter le control.
* Entreprendre toutes études sur la santé de l’homme et la situation sanitaire du pays.
* Centraliser et mettre à jour toutes informations sur les activités de recherche médicale tant en France qu’à l’étranger.
* Effectuer, susciter, encourager les travaux de recherche médicale ou participer à de tels travaux, apporter son concours aux enseignements préparatoires à la recherche médicale.
* Assurer la publication de tous travaux et études se rapportant à ses activités.

L’équipe EpiAgeing dans laquelle j’ai réalisé mon alternance fait partie du Centre de Recherche en Epidémiologie et Statistiques de Sorbonne (CRESS). L’équipe EpiAgeing est spécialisée dans l’épidémiologie du vieillissement et des maladies neurodégénératives. Elle est dirigée par Archana SINGH-MANOUX et est formée de 19 membres composés de chercheurs, d’ingénieurs et de doctorants (voir organigramme, figure X). Leur approche est de couvrir l’ensemble du processus de vieillissement, en partant du déclin fonctionnel et cognitif jusqu’aux maladies. Les données qu’ils utilisent proviennent de cohortes longitudinales ou de bases de données de l’Assurance Maladie.

Les principaux thèmes de recherche sont la recherche des conséquences et des déterminants du déclin cognitif et physique, et l’épidémiologie des maladies neurodégénératives (démence, Parkinson) et cardiovasculaires.

Mon alternance est inscrite dans le cadre d’un projet porté par Julien DUMURGIER, chercheur, sous l’encadrement d’Aline DUGRAVOT.

# 3. PRESENTATION DE L’ALTERNANCE

## 3.1 Présentation de la maladie d’Alzheimer

La maladie d'Alzheimer est une affection neurodégénérative progressive qui affecte le cerveau, principalement chez les personnes âgées. Elle se manifeste par des symptômes tels que la perte de mémoire, la désorientation temporelle et spatiale, ainsi que des troubles du langage et de la pensée. Cette maladie tire son nom du médecin allemand Alois Alzheimer, qui l'a décrite pour la première fois en 1906.

Les caractéristiques principales de la maladie d'Alzheimer sont la formation de plaques amyloïdes et d'enchevêtrements neurofibrillaires dans le cerveau. Les plaques amyloïdes sont des agrégats anormaux de protéines bêta-amyloïdes (Abeta42) qui s'accumulent entre les cellules cérébrales, tandis que les enchevêtrements neurofibrillaires sont constitués de protéines Tau déformées qui perturbent la structure des cellules nerveuses.

Les symptômes de la maladie d'Alzheimer progressent généralement avec le temps. Au début, les signes peuvent inclure des oublis occasionnels et une légère confusion. Cependant, à mesure que la maladie évolue, les personnes atteintes peuvent avoir des difficultés à accomplir des tâches quotidiennes, à reconnaître les proches et à communiquer efficacement. Les symptômes avancés incluent des problèmes de mobilité, une dépendance accrue et une perte progressive de la capacité à prendre soin de soi.

La maladie d'Alzheimer n'a pas de cure, mais des traitements existent pour ralentir les symptômes et améliorer la vie des patients, comprenant des médicaments pour les neurotransmetteurs et des approches non médicales comme la stimulation cognitive et la gestion comportementale

La recherche sur la maladie d'Alzheimer se poursuit activement dans le but de mieux comprendre ses mécanismes sous-jacents et de développer de nouvelles stratégies de traitement. Les efforts visent également à améliorer le diagnostic précoce, car une intervention précoce peut contribuer à ralentir la progression des symptômes.

## 3.2 Contexte

Pour commencer, j’ai dû assimiler le cadre de l’étude.

L’objectif étant de trouver d’éventuelles associations entre les biomarqueurs sanguins et les marqueurs de la maladie d’Alzheimer, il est important de bien comprendre ces deux facteurs.

Dans la base de données associée à l’étude, il y a quatre marqueurs connus de la maladie d’Alzheimer : Abeta42, Abeta40, Tau et Ptau. Les deux marqueurs principaux sont Abeta42 et Tau. Le ratio Abeta42/Abeta40 est un indicateur important.

De manière générale, il a été montré que les patients atteints de la maladie d’Alzheimer ont tendance à avoir des taux anormaux de ces marqueurs. Cependant, cela ne veut pas dire que l’inverse est toujours vrai : Des taux anormaux de ces marqueurs ne sont pas synonymes de maladie d’Alzheimer.

Les biomarqueurs sanguins recueillis dans la base de données sont : Le glucose, le triglycéride et le cholestérol.

L’objectif est de voir si chaque biomarqueur a un lien significatif ou non avec un ou plusieurs des marqueurs de la maladie d’Alzheimer.

Il existe aussi d’autres facteurs connus qui ont tendance à favoriser la maladie d’Alzheimer ou affectent le taux de ces marqueurs :

* L’âge est un facteur qui favorise l'apparition et l’évolution de tous types de démence avec le temps.
* Le sexe, les femmes ont tendance à avoir plus de problèmes cognitifs que l’homme.
* L’APOE est définie par un couple d'allèles. Il existe trois types d’allèles : E2, E3 et E4.
* L’allèle E4 est associé à une augmentation du risque de la maladie d’Alzheimer, particulièrement si cet allèle est présente en deux exemplaires.
* La mesure du MMSE est un score sur 30 qui mesure la présence de troubles cognitifs.
* Les traitements pris par les patients peuvent impactés les taux de de biomarqueurs sanguins ou des marqueurs de la maladie d’Alzheimer et donc peuvent biaiser les analyses s’ils ne sont pas pris en compte.
* IMC (?)
* Niveau d’étude (?), en général, les personnes ayant fait des études supérieures ont tendance à présenter des compétences cognitives plus développées par rapport à celles qui n'ont pas suivi d'études.

## 3.3 Ma mission

Ma mission est de produire un compte rendu ergonomique, ne nécessitant pas de mode d’emplois pour être lu.

Le compte rendu est doté des analyses et subtilités qui m’ont été demandés.

Le code du compte rendu doit pouvoir supporter les mises à jour de la base de données en entrée à fin qu’il puisse être réutilisable au fil du temps.

# 4. RÉALISATION

## 4.1 Compte rendu : Analyses univariées et bivariées

L’objectif de ce rapport est dans un premier temps de mettre avant les statistiques de chaque variable pour vérifier leur pertinence. Dans un second temps d’observer de premier signe d’association. Dans un dernier temps d’observer les effets des variables, citées dans le contexte d’étude, sur les différents marqueurs de la maladie d’Alzheimer.

### 4.1.1 Analyses univariées

La réalisation de cette première partie se divise en trois sous parties et présente des résultats différents en fonction du format de la variable analysée : qualitative, quantitative ou cas atypique.

L'intérêt de ces analyses est de faire ressortir en un coup d'œil la tendance de la variable et son nombre de valeurs manquantes. Une variable avec un trop haut taux de valeur manquante réduira fortement la population dans les futurs modèles et donc sa puissance. Avec ces informations, le chercheur pourra tout de suite savoir si telle variable est exploitable ou non.

#### 4.1.1.1 Variables qualitatives

Pour les variables qualitatives, on retrouve une table montrant le nombre et pourcentage de valeurs manquantes de la variable et la distribution des modalités parmi les valeurs apparentes.

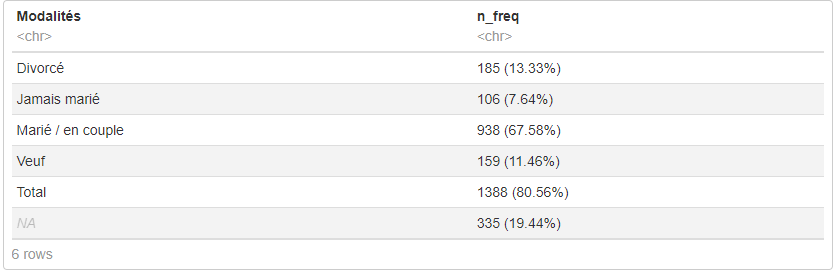


Figure 1, Table de la répartition de la variable “Statut Marital”

Les fonctions de base de R ne permettant pas d’avoir toutes ces informations résumées en un tableau, une fonction a été créée pour le permettre (annexe I).

On retrouve aussi une représentation graphique en bar.

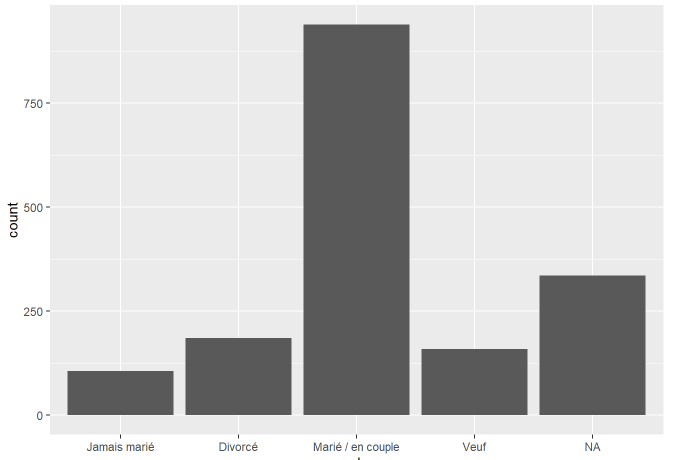


Figure 2, Représentation graphique de la variable “Statut Marital”

Pour certaines variables, un recodage des modalités est nécessaire lorsque ces dernières ne sont pas intuitives (Exemple : Sexe : 0 ou 1, devient “Homme” ou “Femme”) ou si l’ensemble de définition des modalités ne sont pas les plus pertinentes.

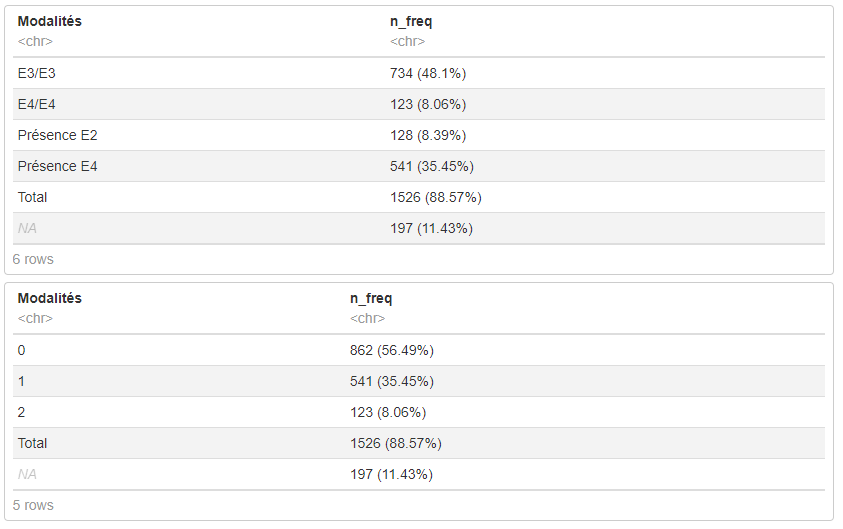


Figure 3, Recodage des modalités de la variable “APOE” en fonction du nombre d’allèle E4, allèle favorisant la maladie d’Alzheimer.

#### 4.1.1.2 Variables quantitatives

Pour les variables quantitatives, un sommaire de la variable est dressé donnant des informations comme la moyenne, le minimum, le maximum, les quartiles, l’écart type, le nombre de valeurs manquantes.

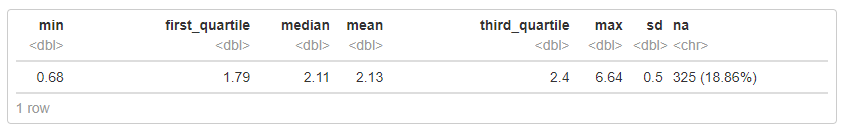


Figure 4, Sommaire de la variable “Cholestérol Total”

Les fonctions de base de R ne permettant pas d’avoir toutes ces informations résumées en un tableau, une fonction a été créée pour le permettre (annexe II)

On retrouve aussi un boxplot et un histogramme en guise de représentation graphique.

Pour certaines variables quantitatives, des seuils peuvent être identifiés pour créer des modalités et donc transformer la variable en qualitative. Cette transformation pour les marqueurs de la maladie d’Alzheimer permet d’ouvrir une nouvelle dimension d’analyse en utilisant la régression logistique.

#### 4.1.1.3 Variables atypiques

##### 4.1.1.3.1 Variable MMSE

En fonction de leur signification et de leur fonctionnement, certaines variables ont besoin d’analyses particulières.

Pour le MMSE, il m’a été demandé d’approfondir les analyses, de créer de nouveaux indicateurs à partir des variables mises à disposition (Nombre de mesure du MMSE, leur date, date de la ponction lombaire). Ainsi, les variables suivantes ont étés créés :

* Une variable correspondant au score MMSE ayant la date de réalisation la plus proche de celui de la ponction lombaire. Un patient peut avoir effectué plusieurs scores MMSE lors de son suivi médical. Cependant le score le plus intéressant est celui qui est le plus proche de la date de ponction lombaire, pour que ça soit cohérent avec les taux de protéines et des marqueurs lors des analyses.
* Une variable sur la date du dernier score MMSE réalisé en jours.
* Une variable sur la différence de date entre le premier et le dernier score MMSE réalisé en jours.
* Une variable sur le nombre de scores MMSE effectués avant la ponction lombaire.
* Une variable sur le nombre de scores MMSE effectuée après la ponction lombaire.

Pour chacune de ces variables, qui sont quantitatives, nous avons un sommaire, un boxplot et un histogramme.

##### 4.1.1.3.2 Variable sur les traitements

La variable sur les traitements pris par les patients a été conçue en plusieurs étapes.

Pour chaque patients, ses médicaments sont rentrés manuellement dans la base de données par le médecin traitant. Autrement dit, des erreurs de saisies sont attendues et la syntaxe complexe des différents médicaments accroît le taux d’erreur.

Dans un premier temps, afin de réduire le taux d’erreur humaine, un masque (Voir annexe III) a été appliqué sur toutes les modalités des traitements.

Par exemple, il est courant de se tromper entre “I” et “Y”, d’oublier un “L” ou d’en rajouter un, ou alors d’oublier un “E” à la fin.

Ci-dessous, le fonctionnement du masque sur le mot “FILLE”

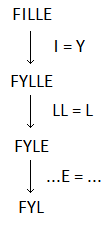


Figure 5, Schéma d’application du masque sur le mot “FILLE”

N’importe quelle variance mineure dans l’orthographe du mot “Fille” n’aura finalement aucune conséquence sur sa sortie après le masque, qui sera “Fyl”.

D’autres subtilités comme des nombres, des espaces etc… ont été pris en compte afin d’uniformiser le plus possible les modalités.

La prochaine étape est d’identifier les traitements les plus utilisés par les patients de notre base de données, afin de créer plusieurs listes de traitement en fonction de leur principe actif (diabète, haute tension artérielle, cholestérol…).

Pour chaque liste, une variable en dichotomique est créée :

* 0, si le patient ne prend pas un des médicaments de la liste derrière la variable.
* 1, si c’est le cas.

Une fois ces variables créées, nous avons le tableau de fréquence suivant :

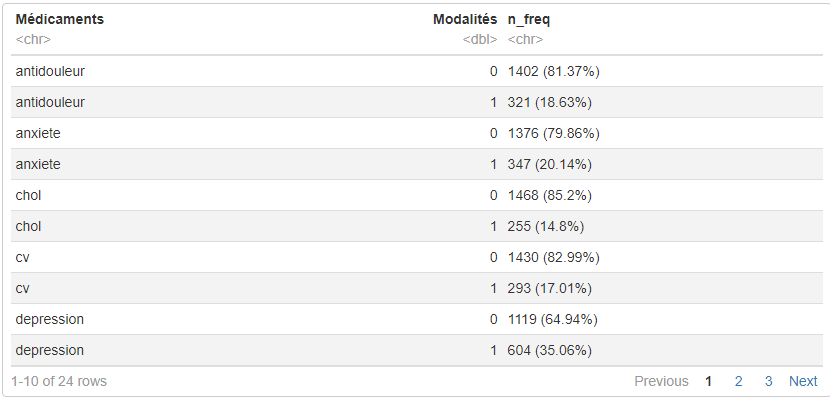


Figure 6, Tableau de fréquence des différents types de traitements pris par les patients

##### 4.1.1.3.3 Variables sur les marqueurs de la maladie d’Alzheimer

Chaque marqueur de la maladie d'Alzheimer est exprimé en pg/mL, donc en continu.

Comme dit plus tôt, il est aussi intéressant d’avoir ces marqueurs sous forme de variable qualitative. Pour cela, il faut se fier à des seuils.

Des seuils existent mais ils sont différents en fonction de la date de la ponction lombaire.

Autrement dit, le taux du marqueur n’a pas la même signification en fonction de la période.

Une fois ce critère respecté, on obtient le tableau suivant :

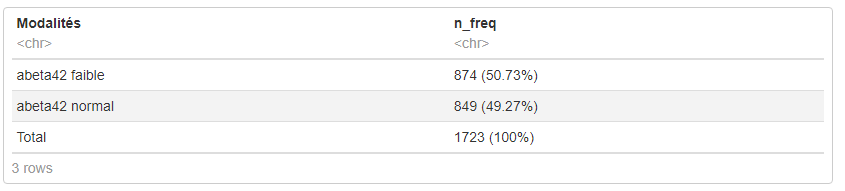


Figure 7, Tableau de fréquence des modalités de la variable “abeta42”

Afin de faciliter l’interprétation des modèles, les marqueurs ont été standardisés.

Pour avoir une standardisation sans biais à cause des différentes périodes, les marqueurs ont été découpés en plusieurs sous-groupes correspondant chacun à une période et standardisés sur leur logarithme (afin de lisser la distribution) par groupe avant d’être réassemblés.

Le taux des marqueurs Abeta42 et Abeta40 sont anormaux lorsqu’ils sont faibles.

Toujours dans le but de faciliter l’interprétation des modèles, les variables associés à ses marqueurs ont été multipliées par -1 afin d’associer l’augmentation du taux à un risque.

### 4.1.2 Études d'association

L’objectif de cette partie est d’avoir une idée globale des associations entre les marqueurs de la maladie d’Alzheimer et les biomarqueurs avec des modèles linéaires et logistiques.

Les biomarqueurs sanguins que nous testons sont le triglycéride, le glucose et le cholestérol total, en continue.

Pour chaque étude d’association entre un biomarqueur et un marqueur de la maladie d’Alzheimer, les modèles suivants décrivant les variables explicatives seront dressés :

* Modèle 1 : Biomarqueur + âge
* Modèle 2 : Biomarqueur + âge + sexe
* Modèle 3 : Biomarqueur + âge + sexe + APOE
* Modèle 4 : Biomarqueur + âge + sexe + APOE + Traitements
* Modèle 5 : Biomarqueur + âge + sexe + APOE + Traitements + MMSE
* Modèle 6 : Biomarqueur + âge + sexe + APOE + Traitements + MMSE + IMC

Ces différents modèles nous permettent de suivre l’évolution des indicateurs au fur et à mesure que l’on ajoute des variables explicatives au modèle et donc de déterminer à quelle étape le modèle perd le plus de puissance.

Pour pouvoir comparer ces modèles sans biais, une base de données a été spécialement créée, comportant aucune valeur manquante sur l’ensemble des variables vouées à être explicatives.

Autrement dit, la base de données en entrée du modèle 1 à 6 est la même, donc la même population étudiée.

(Expliquer comment marche un modèle ?)

### 4.1.3 Conception du compte rendu

Pour rendre le document ergonomique et faciliter sa navigation, un sommaire a été ajouté pour

Une de mes missions est de rendre le compte rendu lisible et ergonomique.

Pour faciliter la navigation dans le document, j’ai ajouté un sommaire dynamique, en bord de page, afin que l’on puisse toujours avoir un œil sur notre position dans le document

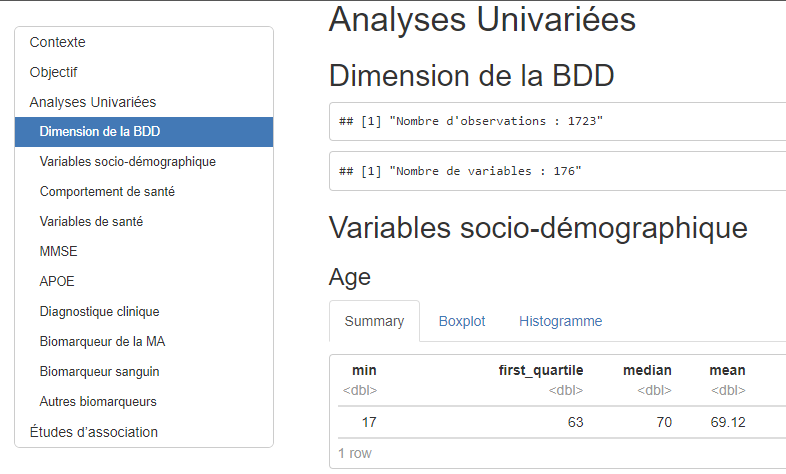


Figure 8, Sommaire du compte rendu

Les variables sont classées en différentes catégories en fonction de leurs significations afin que le défilement des variables soit cohérent. (Exemple : les variables socio-démographiques, variables de santé etc…)

Pour chaque variable, ses analyses sont rangées en onglet pour maximiser l’ergonomie du document et faire apparaitre seulement les analyses souhaitées par l’utilisateur.

### 4.1.4 Conclusion

Ce compte rendu a permis de mettre en évidence certaines relations notamment entre le cholestérol et les marqueurs de la maladie d’Alzheimer.

## 4.2 Compte rendu : Analyses centrées sur le cholestérol

### 4.2.1 Fonctionnement du compte rendu

Sur ce compte rendu, nous allons étudier plus en détails l’effet du cholestérol sur les marqueurs de la maladie d’Alzheimer. Plus précisément, l’abeta42, le ratio abeta42/abeta40 et le ptau.

*Le cholestérol total correspond à la somme du cholestérol hdl et ldl.*

*Le cholestérol hdl est ce qu’on appelle “le bon cholestérol ” et le cholestérol ldl le “mauvais cholestérol ”*

*Il est donc intéressant de voir la forme de l’association entre les biomarqueurs de la maladie d’Alzheimer et les différents types de cholestérol.*

Des sous populations ont été créées avec comme critère un facteur connu de la maladie d’Alzheimer afin d’étudier le comportement du cholestérol et identifier des changements.

Plusieurs de ces sous populations ont comme critère un diagnostic émis par le médecin traitant du patient. La variable associée au diagnostic du patient est une variable qualitative avec saisie contrôlée définit par les modalités suivantes :

* MA (Maladie d’Alzheimer)
* MCI amnésique (Troubles cognitifs légers amnésique)
* MCI autre
* Démence mixte
* Démence vasculaire
* DFT (Dégénérescence fronto-temporale)
* Lewy
* Atrophie corticale postérieure
* Démence alcoolique
* Aphasie progressive primaire
* Trouble psychiatrique
* VIH
* NSP
* Autre

La modalité “MA” est intéressante parce qu’elle nous permet d’avoir des observations “références”.

Les modalités sur le “MCI” le sont aussi parce qu’elles désignent des patients avec des troubles cognitifs. Les patients diagnostiqués “MCI” ont tendance à être des cas prématurés d’Alzheimer.

Les sous populations créées sont les suivantes :

* En fonction de la présence d’allèle E4 (APOE)
* En fonction de la présence d’allèle E4 chez les patients diagnostiqués “MA”
* En fonction du diagnostic “MA”
* En fonction du diagnostic “MA” ou “MCI amnésique”
* En fonction du diagnostic en plusieurs groupes (“MA”, “MCI amnésique”, “MCI autres”, autres)

Pour chacun de ces cas, une variable dichotomique (sauf le dernier) a été créée pour savoir si les patients remplissent les conditions du cas ou non.

Soit “1” pour les patients remplissant les conditions et “0” pour le contraire.

Cette variable permet de couper la base de données en deux, ayant d’un côté, une base de données avec les patients remplissant les conditions et de l’autre côté ceux qui ne les remplissent pas.

Dresser les mêmes modèles sur ces deux sous-populations permet de voir l’influence des conditions sur la population totale.

Exemple :

Ci-dessous, les résultats des modèles visant à expliquer le taux d’abeta42 en fonction du cholestérol total opposant les patients diagnostiqués “MA” et ceux qui ne le sont pas.

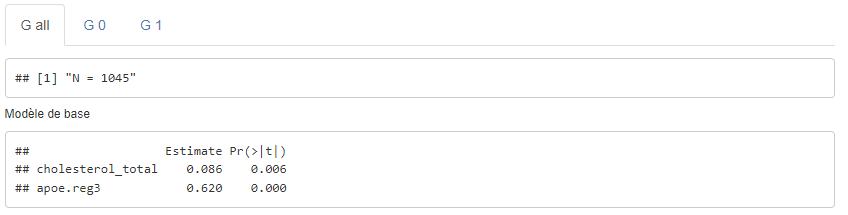


Figure 9, Résultats du modèle avec tous les patients

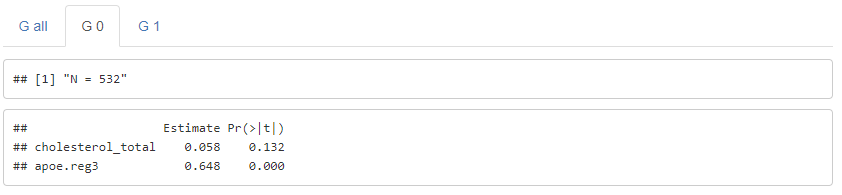


Figure 10, Résultats du modèle avec les patients qui ne sont pas diagnostiqués “MA”

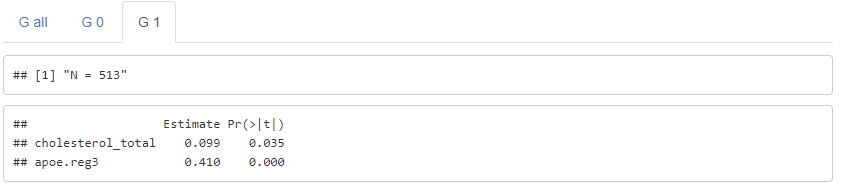


Figure 11, Résultats du modèle avec les patients qui sont diagnostiqués “MA”

La première sortie nous sert de référence, à savoir quels sont les coefficients sur la population totale. On remarque que l’effet du cholestérol est positif (0.086) et est significatif au seuil 5% (0.006).

On peut remarquer sur les résultats du modèle avec les patients non diagnostiqués “MA”, que l’effet du cholestérol n’est pas significatif au seuil 5% (0.132).

Contrairement aux résultats du modèle avec les patients diagnostiqués “MA”, où l’effet du cholestérol est significatif (0.035).

Donc, en se basant sur ces résultats, il semblerait que les patients diagnostiqués “MA” ont un taux de cholestérol lié à leur taux d’abeta42 (Interprétation à revoir).

Nous avons vu que le cholestérol total était significatif sur le sous-groupe diagnostiqués “MA”

Or le cholestérol total est la somme du cholestérol hdl et ldl.

Il serait donc intéressant de savoir lequel de ces deux types du cholestérol influence l’association significative du cholestérol total.

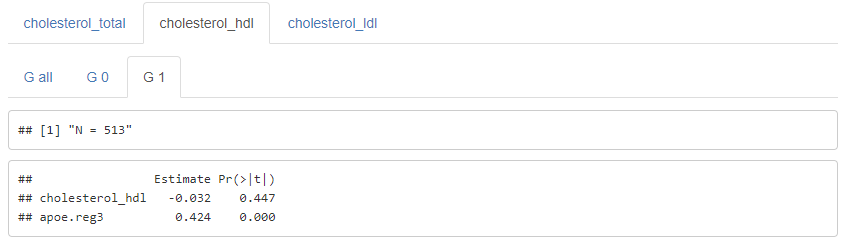


Figure 12, Résultats du modèle sur le cholestérol hdl avec les patients qui sont diagnostiqués “MA”



Figure 13, Résultats du modèle sur le cholestérol ldl avec les patients qui sont diagnostiqués “MA”

D’après ces deux sorties, nous voyons que le cholestérol hdl n’est pas significatif contrairement au cholestérol ldl.

Nous pouvons donc être plus précis dans notre interprétation en précisant que les patients diagnostiqués “MA” ont un taux de cholestérol ldl lié à leur taux d’abeta42 (Interprétation à revoir).

### 4.2.2 Conception du compte rendu

Ce document comporte presque 80 modèles différents. Pour réaliser ce document en minimisant le temps d’écriture du code, il a fallu “automatiser” le code le plus possible.

Autrement dit, concentrer les changements entre chaque partie derrière des variables.

Pour cela, il faut bien comprendre quelle structure donner au document.

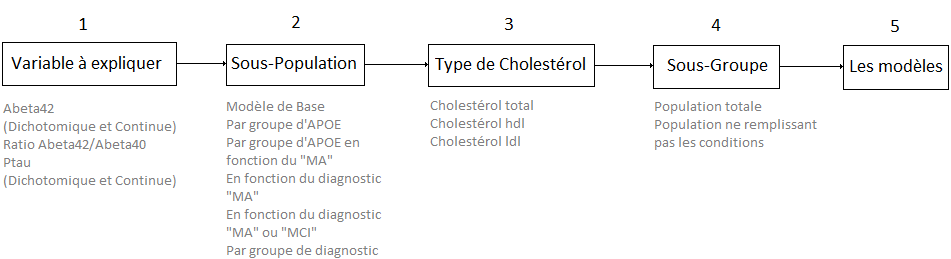


Figure 14, Schéma de la conception du document

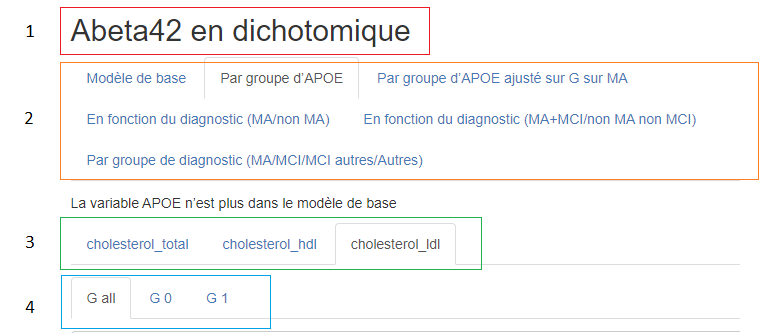


Figure 15, Aperçu du document

Chaque étape correspond à un paramètre dans le code du modèle.

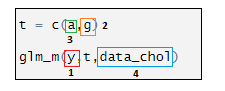


Figure 16, Code d’un modèle

Pour minimiser les changements entre chaque modèle, chaque étape correspond à une variable. Les fonctions de base de Rstudio ne permettant pas de supporter ce format en entrée et de faire la sortie attendue, la fonction “glm\_m” a été créée (voir Annexe IV)

### 4.2.3 Conclusion

Ce compte rendu nous à permis de remarquer une association significative entre le cholestérol ldl et le taux d’Abeta42. Cette association se remarque notamment chez les patients atteints de troubles cognitifs. Les résultats divergent en fonction de si le marqueur est en continue ou en dichotomique. L’association entre le cholestérol et les marqueurs n’est pas forcément linéaire. Il serait donc intéressant d’analysant la forme de cette association.

## 4.3 Compte rendu : Analyses exploratoires sur le cholestérol

### 4.3.1 Fonctionnement du compte rendu

Le dernier compte rendu a mis en évidence une liaison significative entre l’abeta42 et le cholestérol chez les patients diagnostiqués “MA” ou “MCI”.

Nous nous intéressons maintenant à la forme de cette association. Lors de nos analyses, nous nous limitions à un modèle linéaire pour expliquer les marqueurs de la maladie d’Alzheimer en continue.

#### 4.3.1.1 Les différents modèles

L’objectif de ce document est de modéliser la liaison entre l’abeta42 et les différents types de cholestérol d’une population spécifique en utilisant plusieurs méthodes :

* Un spline
* Un modèle cubique
* Un modèle quadratique
* Un modèle linéaire
* Un modèle linéaire par morceau

Chaque méthode a ses avantages et ses inconvénients :

* Le spline permet de contrôler la souplesse de la courbe en ajustant le nombre et la position des nœuds. Cette méthode est particulièrement utile lorsqu’on attend une courbe lisse. Cependant le choix du nombre et de la position des nœuds est subjectif et peut grandement influencer les résultats.
* Le modèle linéaire est le plus simple à interpréter et est moins impacté par les valeurs aberrantes que les modèles plus complexes. Il est totalement adapté si la relation entre les variables est linéaire. Cependant ce modèle ne peut capturer que des relations linéaires, ce n’est pas souvent le cas dans le domaine médical et sa flexibilité est nulle comparé aux modèles plus complexes.
* Le modèle linéaire par morceau permet de capturer des relations non linéaires en divisant l’espace des variables en segments linéaires distincts. Il est plus flexible qu’un modèle linéaire simple, tout en conservant sa simplicité. Néanmoins la détermination du ou des points de rupture est subjective.
* Le modèle quadratique peut saisir des tendances de courbes simples dans les données et est plus flexible qu’un modèle linéaire. Cependant ce modèle est sensible aux valeurs aberrantes et n’est pas suffisamment flexible pour capturer les variations les plus complexes.
* Le modèle cubique peut capturer des variations plus complexes que le modèle linéaire et quadratique et est plutôt adapté pour ajuster des relations non linéaires. Cependant, ce modèle est sensible aux valeurs aberrantes ce qui peut influencer les coefficients estimés et les interprétations de ces derniers est plus compliqués que pour un modèle linéaire ou quadratique.

Pour résumer en un tableau :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Spline** | **Linéaire** | **Linéaire par morceau** | **Quadratique** | **Cubique** |
| **Flexibilité** | Elevée | Faible | Moyenne | Moyenne | Elevée |
| **Sensibilité aux valeurs aberrantes** | Moyenne | Faible | Moyenne | Elevée | Elevée |
| **Interprétation** | Difficile | Très facile | Facile | Moyenne | Difficile |
| **Subjectivité** | Oui | Non | Oui | Non | Non |

Figure 17, Comparaison des méthodes de modélisation : Spline, Modèles Polynomiaux et Linéaires par Morceau

Ci-dessous, la représentation graphique des 5 modèles modélisant la relation entre l’abeta42 et le cholestérol total chez les patients diagnostiqués “MA” :

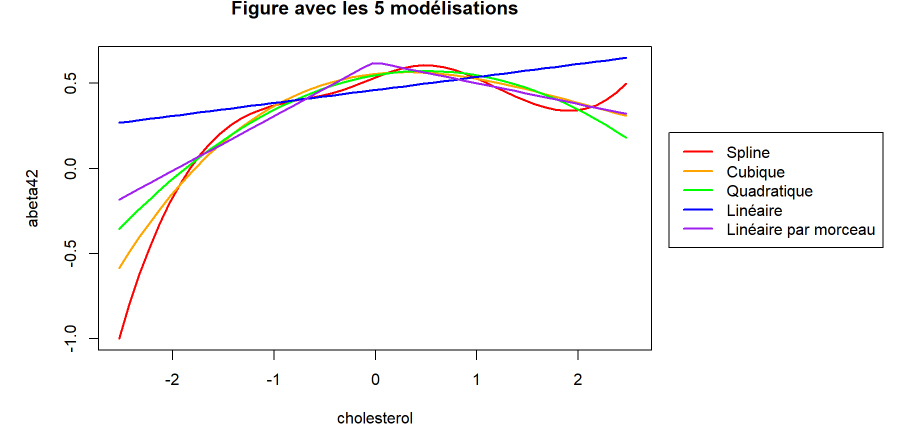


Figure 18, Représentation graphique des 5 modèles modélisant la relation entre l’abeta42 et le cholestérol total chez les patients diagnostiqués “MA”.

Pour obtenir des représentations graphiques correctes et interprétables, une nouvelle base de données a été créée, retenant seulement les variables présentes dans les modèles.  
Le nombre d’observations correspond aux nombres de valeurs du cholestérol entre son minimum et son maximum avec un pas de 0,1. Ce choix de pas est la valeur standard mais reste subjectif.  
Les variables explicatives sont initialisées sur leur modalité par défaut (“0” pour une variable dichotomique, la modalité neutre pour une variable qualitative). La variable du biomarqueur est manquante.

Pour définir les valeurs du biomarqueur, nous utilisons le modèle pour prédire ces valeurs en l’appliquant sur cette nouvelle base de données. La représentation graphique sera dressée avec ces prédictions.

Ci-dessous la représentation graphique du modèle sur la base de données originale et la représentation graphique du modèle sur la nouvelle base de données.

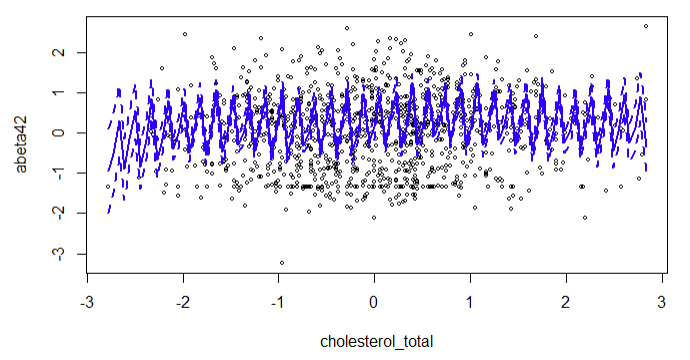


Figure 19, la représentation graphique du modèle spline sur la base de données originale

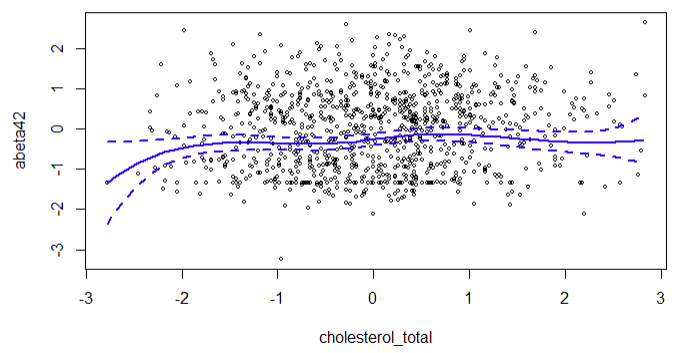


Figure 20, la représentation graphique du modèle spline sur la nouvelle base de données

#### 4.3.1.2 Test de vraisemblance

Une représentation graphique ne suffit pas à choisir quelle méthode est la meilleure.

En se basant sur les coefficients des modèles, le test de vraisemblance permet de comparer un premier modèle à un deuxième, en évaluant la probabilité d'observer les données et en déterminant si la différence de vraisemblance entre les modèles est significative pour justifier le choix du second modèle.

Exemple :

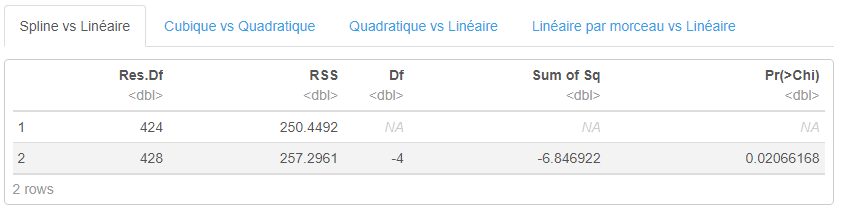


Figure 21, Test de vraisemblance opposant le modèle Spline au modèle Linéaire.

On peut apercevoir sur ce test de vraisemblance une somme des carrés des résidus (RSS), qui mesure la quantité de variation non expliquée par le modèle, plus faible pour le modèle spline que le modèle linéaire, donc le modèle spline explique mieux la variation que le modèle linéaire

Or le modèle linéaire a 4 paramètres (Df) de moins que le modèle spline, donc il est moins complexe.

La p-valeur Pr(>Chi) est significative au seuil 5%. Donc en prenant en compte ces coefficients, le modèle linéaire est meilleur que le modèle spline.

#### 4.3.1.3 Critères d’information

Il existe d’autres méthodes pour comparer la qualité des modèles.

L'AIC (Critère d'Information Akaike) est une mesure qui évalue la qualité de l'ajustement d'un modèle aux données tout en pénalisant la complexité du modèle. Le BIC (Critère d'Information Bayésien) est similaire à l'AIC, mais il pénalise davantage les modèles plus complexes, favorisant ainsi la parcimonie.

Plus la mesure est faible, plus le modèle est de qualité.

Exemple :

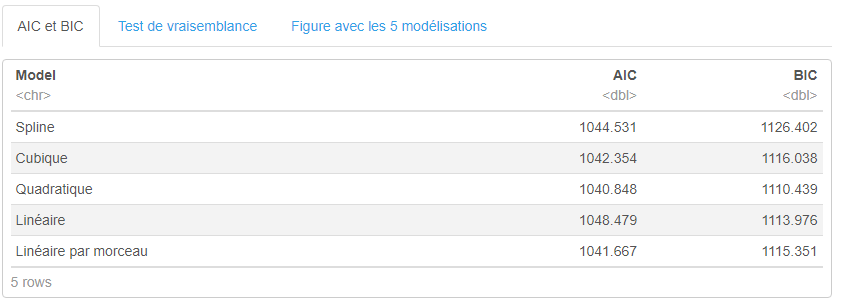


Figure 22, Comparaison de l’AIC et le BIC des 5 modèles.

Ce tableau montre que le modèle quadratique semble être le meilleur en se basant sur l’AIC et le BIC.

Le modèle linéaire par morceau a une mesure AIC proche du modèle quadratique mais une mauvaise mesure BIC.

### 4.3.2 Conception du compte rendu

Comme pour le compte rendu précédent, toutes les modifications entre les différentes analyses ont été regroupées derrière des variables.

Cette fois-ci, le but de ce compte rendu n’est pas seulement visuel mais aussi interactif.

Le code doit être facilement manipulable par une personne ne connaissant pas spécialement le langage R, pour le modifier et obtenir les résultats souhaités.

Les facteurs qui peuvent changer sont la variable à expliquer (le marqueur de la maladie d’Alzheimer) et la population étudiées (exemple : les patients seulement diagnostiqués “MA”)

Tous ces changements sont résumés dans la boîte de code suivante :

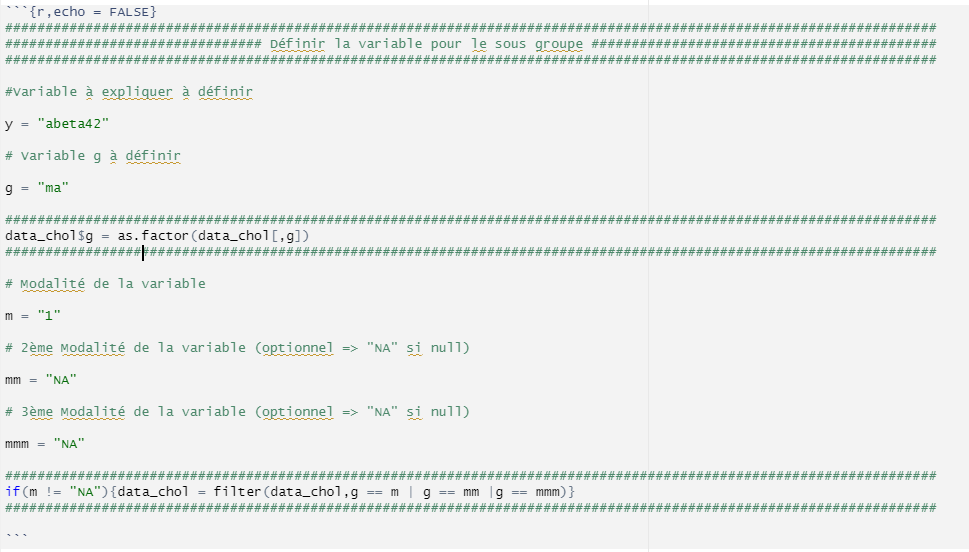


Figure 23, Code interactif

La variable “y” correspond à la variable à expliquer, ici abeta42.

La variable “g” correspond à la variable qui filtrera la population.

La variable “m” correspond à la modalité souhaitée de la variable “g”.

Les variables “mm” et “mmm” sont présentes dans le cas où nous voudrions filtrer sur plusieurs modalités de la variable “g”.

En nous fiant à ce code, nous savons que les résultats seront sur l’abeta42 et la population correspond aux patients diagnostiqués “MA”.

Pour que cela soit clair, ces spécificités sont précisées dans le compte rendu.



Figure 24, Spécificités de population étudiée dans le compte rendu

### 4.3.3 Conception du compte rendu

Comme pour le compte rendu précédent, toutes les modifications entre les différentes analyses ont été regroupées derrière des variables.

Cette fois-ci, le but de ce compte rendu n’est pas seulement visuel mais aussi interactif.

Le code doit être facilement manipulable par une personne ne connaissant pas spécialement le langage R, pour le modifier et obtenir les résultats souhaités.

Les facteurs qui peuvent changer sont la variable à expliquer (le marqueur de la maladie d’Alzheimer) et la population étudiées (exemple : les patients seulement diagnostiqués “MA”)

Tous ces changements sont résumés dans la boîte de code suivante :

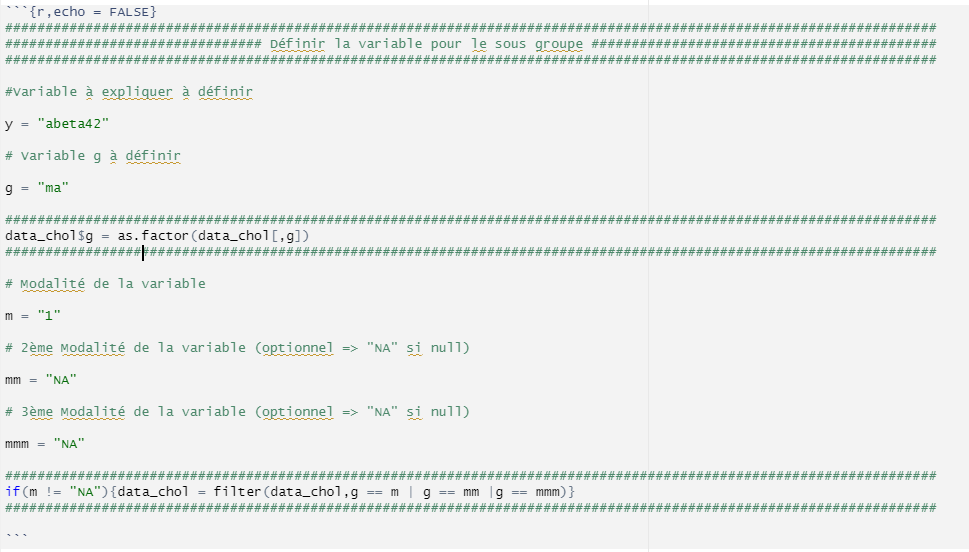


Figure 25, Code interactif

La variable “y” correspond à la variable à expliquer, ici abeta42.

La variable “g” correspond à la variable qui filtrera la population.

La variable “m” correspond à la modalité souhaitée de la variable “g”.

Les variables “mm” et “mmm” sont présentes dans le cas où nous voudrions filtrer sur plusieurs modalités de la variable “g”.

En nous fiant à ce code, nous savons que les résultats seront sur l’abeta42 et la population correspond aux patients diagnostiqués “MA”.

Pour que cela soit clair, ces spécificités sont précisées dans le compte rendu.



Figure 26, Spécificités de population étudiée dans le compte rendu

# 5. CONCLUSION GÉNÉRALE

Cette alternance m'a permis de créer des comptes rendus d’analyses sur une base de données inexplorée, du traitement des données jusqu’à des analyses exploratoires, qui contribueront à la recherche sur la maladie d’Alzheimer.

J’ai évolué dans cette alternance au côté d’un chercheur et d’une ingénieure statisticienne, j’étais donc guidé dans mes démarches afin de répondre aux demandes du chercheur.  
  
D’ici la fin de mon alternance, il me reste encore à peaufiner le compte rendu sur les analyses exploratoires, à identifier les différents cas d’Alzheimer (...) et de sortir des analyses sur ces sous-populations, et enfin, écrire une procédure d’utilisation de cette dernière pour que mon équipe puisse l’utiliser une fois mon alternance terminé.

Cette expérience d’alternance fut très constructive. Elle m’a permis de comprendre comment les équipes de recherche médicale opèrent pour exploiter leurs données et les interpréter.

# 6. TABLE DES ILLUSTRATIONS

[Figure 1, Table de la répartition de la variable “Statut Marital” 8](#_Toc143778543)

[Figure 2, Représentation graphique de la variable “Statut Marital” 9](#_Toc143778544)

[Figure 3, Recodage des modalités de la variable “APOE” en fonction du nombre d’allèle E4, allèle favorisant la maladie d’Alzheimer. 10](#_Toc143778545)

[Figure 4, Sommaire de la variable “Cholestérol Total” 10](#_Toc143778546)

[Figure 5, Schéma d’application du masque sur le mot “FILLE” 12](#_Toc143778547)

[Figure 6, Tableau de fréquence des différents types de traitements pris par les patients 12](#_Toc143778548)

[Figure 7, Tableau de fréquence des modalités de la variable “abeta42” 13](#_Toc143778549)

[Figure 8, Sommaire du compte rendu 14](#_Toc143778550)

[Figure 9, Résultats du modèle avec tous les patients 16](#_Toc143778551)

[Figure 10, Résultats du modèle avec les patients qui ne sont pas diagnostiqués “MA” 16](#_Toc143778552)

[Figure 11, Résultats du modèle avec les patients qui sont diagnostiqués “MA” 17](#_Toc143778553)

[Figure 12, Résultats du modèle sur le cholestérol hdl avec les patients qui sont diagnostiqués “MA” 17](#_Toc143778554)

[Figure 13, Résultats du modèle sur le cholestérol ldl avec les patients qui sont diagnostiqués “MA” 17](#_Toc143778555)

[Figure 14, Schéma de la conception du document 23](#_Toc143778556)

[Figure 15, Aperçu du document 23](#_Toc143778557)

[Figure 16, Code d’un modèle 23](#_Toc143778558)

[Figure 17, Comparaison des méthodes de modélisation : Spline, Modèles Polynomiaux et Linéaires par Morceau 19](#_Toc143778559)

[Figure 18, Représentation graphique des 5 modèles modélisant la relation entre l’abeta42 et le cholestérol total chez les patients diagnostiqués “MA”. 19](#_Toc143778560)

[Figure 19, Test de vraisemblance opposant le modèle Spline au modèle Linéaire. 20](#_Toc143778561)

[Figure 20, Comparaison de l’AIC et le BIC des 5 modèles. 20](#_Toc143778562)

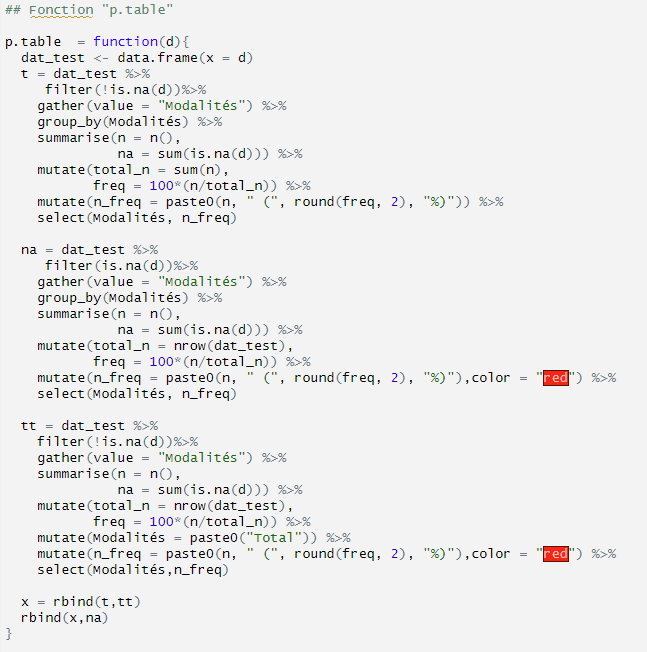
[Figure 21, Code interactif 21](#_Toc143778563)

[Figure 22, Spécificités de population étudiée dans le compte rendu 22](#_Toc143778564)

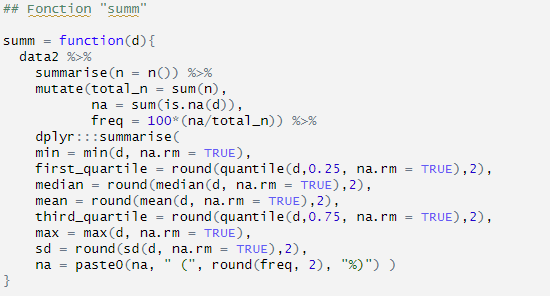
# 

# 7.ANNEXES

**Annexe I :** Fonction p.table



**Annexe II :** Fonction summ



**Annexe III :** Code du Masque



**Annexe IV :** Fonction glm\_m



1. Une étude de cohorte est un type d'enquête qui consiste à suivre dans le temps une population définie (la cohorte) et de comparer l'incidence d'un évènement [↑](#footnote-ref-1)