

Mémoire présenté le : 22 juin 2021

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA
et l'admission à l'Institut des Actuaires**

Par : HUYNH Sandrine

Titre Open data et Assurance santé : l'union fait la force ?

Confidentialité : NON OUI (Durée : 1 an 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus.

*Membres présents du jury de l'Institut
des Actuaires*

Signature

Entreprise :

Nom : ACTUELIA

Signature :

Eugenie POYET

Directeur de mémoire en entreprise :

Nicolas MARESCAUX

Nom : Frank BOUKOBZA

Signature :

Membres présents du jury de l'ISFA

Invité :


Nom :

Signature :


Yahia SALHI

***Autorisation de publication et de mise en
ligne sur un site de diffusion de documents
actuariels (après expiration de l'éventuel
délai de confidentialité)***

Signature du responsable entreprise



Signature du candidat



Open data et Assurance santé : l'union fait la force ?

Mémoire d'Actuariat pour l'obtention du Diplôme Universitaire d'actuariat de l'Institut de Science Financière et d'Assurances et l'admission à l'Institut des Actuaire



HUYNH Sandrine

2021

Résumé

Mots clés : Assurance santé, Open data, Open DAMIR, tarification, enrichissement de base de données, MLG, données agrégées, prime pure, machine virtuelle.

Il était une fois... Une mutuelle qui souhaitait étendre le tarif de quelques-uns de ses produits d'**assurance santé** préexistants à un périmètre national. En effet, bien qu'éclairée sur les prix pratiqués dans sa région de localisation, elle n'avait cependant pas cette expertise sur les régions voisines, et encore moins sur le reste de la France. Son budget était cependant limité et elle souhaitait faire le plus possible d'économies : envisager l'achat de données n'était donc pas son choix favori. Elle décida alors de s'orienter vers les « **Open data** », données publiques facilement accessibles et qui plus est, gratuites. Elle y trouva alors une opportunité. Après état des lieux des alternatives et informations légales obtenues, elle sélectionna dans la vastitude des possibilités l'**Open DAMIR**, par sa complétude. Équipée de la puissance de calculs de son nouvel outil que constitue la **machine virtuelle** nouvellement acquise, elle s'attaqua alors à une démarche de **tarification** dans un but ultime d'extension de **primes pures** : détermination des différents segments tarifaires, études et retraitements de la matière première que sont les données à sa disposition (internes et externes), mise en cohérence de jeux de données, choix de méthodes de tarification (**modèles linéaires généralisés** ou calculs statistiques directs) et enfin, construction de coefficients de déformation de tarifs pour modéliser l'effet « région ». Le chemin sera cependant semé d'embûches avec un adversaire de taille : le traitement des **données agrégées**.

Abstract

Keywords: Health insurance, Open data, Open DAMIR, pricing, data enrichment, GLM, aggregate data, net premium, virtual machine.

Once upon a time... There was a mutual insurance company which wanted to extend the prices of some of its existing **health insurance** products national-wide. While the company was familiar with pricing business in the region of France that it has been operating in for years, it had little experience when it came to pricing for neighbouring areas and even less when it came to the rest of France. As its budget was limited and savings were to be optimized, buying databases was not its first choice. The company then decided to use "**Open data**" because they are easily accessible and completely free data sources. It thus found an opportunity. It chose the **Open DAMIR** database because of its comprehensiveness after performing some legal and situational analysis of the possible database alternatives. Equipped with the computational power of its newly-found tool – a **virtual machine** – it took the following **pricing** steps to determine the **net premiums** for its business expansion, which included: identifying various pricing segments; studying and altering the raw materials which are the external and internal data at its disposal; matching databases; choosing the pricing methods (**generalized linear regressions** or straight statistical calculations); and building the adjustment factors to model the "area" effect on premiums. Nevertheless, it was not an easy task and the path was strewn with pitfalls with a challenging opponent: how to process **aggregate data**.

Remerciements

Je tenais tout d'abord à remercier l'ensemble de mes collègues du cabinet de ACTUELIA qui se sont rendus disponibles pour échanger tout au long de la réalisation de ce mémoire, que ce soit en tant que soutien, pour me donner leur avis sur les divers entraînements de présentation, ou pour avoir eu assez de curiosité pour se mettre à jour sur l'avancement de mon travail. Cela m'a ainsi demandé d'expliquer mon sujet sous toutes ses coutures. Plus particulièrement :

* Victoire PIAT et Frank BOUKOBZA, mes tuteurs d'entreprise, qui m'ont suivie tout le long de cette aventure, de la première étincelle d'idée à sa finalisation ; qui m'ont encouragée, supportée, conseillée et qui ont dû relire à plusieurs reprises les écrits, temporaires ou finaux, en situations diverses et variées (congrès, week-end, au bord de la piscine, ...). Je vous remercie sincèrement pour votre implication, je n'aurais pas pu avoir mieux.

* Jean-Nicolas MARRILLIET, qui s'est retrouvé soudainement embarqué dans ce projet, pour son expertise et son expérience, pour ses éclaircissements et ses idées, pour sa relecture ciblée et enfin, pour avoir généreusement accepté de subir une partie de la classification fastidieuse qui sera présentée par la suite dans ce mémoire.

Je tenais de même à remercier l'équipe pédagogique de l'Institut de Science Financière et d'Assurances (ISFA) pour leur pédagogie malgré un temps de crise sanitaire et pour s'être rendue réceptive aux questions que j'avais à poser. Plus particulièrement :

* Didier RULLIERE, mon tuteur pédagogique, qui m'a aussi suivi tout au long du mémoire malgré son changement d'école. Bien que le sujet de ce mémoire ne soit pas son domaine de prédilection, ses avis, sa bienveillance et ses aides d'accès aux ressources n'ont pas été négligeables.

* Anne MARION qui m'a fait réaliser en cours d'Assurance Santé que mon sujet initial fonçait droit dans un mur et qui n'a pas hésité à me fournir de la lecture complémentaire.

* Xavier MILHAUD pour s'être rendu disponible et réactif pour répondre à des questions sur ses cours de tarification, y compris après ma sortie d'école.

Petite mention spéciale aussi à mes camarades de l'ISFA, notamment François-Xavier CHAMOULAUD pour son support constant et son partage de connaissance dans le domaine de l'informatique (programmation) et de la machine virtuelle. De même, à Arnold MEKONTSO et Alban GARNIER, qui ont aussi abordé dans leur mémoire l'Open DAMIR et qui ont accepté de répondre aux quelques questions initiales que j'avais ; Stella LUGIERY et Thomas COLONNEAUX sur qui je me tournais en cas de questions sur SAS comme j'ai dû apprendre aussi à coder sur ce programme au fur et à mesure.

Enfin, après cette liste de remerciement non exhaustive, je souhaiterais juste terminer sur :

Si je devais dédicacer cet écrit qui représente une culmination d'années d'études, cela serait à mes parents, HUYNH Thi Khanh Van et HUYNH Phuoc Kha, qui ont toujours été là pour moi et qui, je sais, m'auraient aidé à le faire s'ils le pouvaient. Merci de m'avoir offert le meilleur environnement de travail sur la fin où je n'ai eu pratiquement qu'à manger, dormir, étudier et rien d'autres à faire lorsque je rentrais à la maison pour travailler à 100 % sur ce projet et qui n'osaient même pas parler de peur de me déconcentrer. Ce que je fais est pour moi mais aussi pour vous.

Sommaire

Introduction	8
Guide de lecture.....	9
Chapitre 1 - Quand le monde réel de l'assurance santé s'entrelace avec le monde virtuel de l'open data ...	10
Section 1 - D'un côté, le monde de l'assurance santé.....	11
1.1.1. Les régimes obligatoires	11
1.1.2. Les régimes complémentaires	13
1.1.3. Les notions de base	14
1.1.4. Nomenclatures des actes	17
1.1.5. Un point sur la législation : Santé	19
Section 2 - ... De l'autre, le monde de l'open data.....	22
1.2.1. Un second point sur la législation : Open Data	22
1.2.2. Un petit état de l'art de l'Open Data en matière de santé	23
1.2.3. Un focus sur l'Open DAMIR	28
1.2.4. Les machines virtuelles : une solution logistique à la pointe de la technologie	29
Pour faire le point.....	32
Chapitre 2 – « Si vous essayez de bâtir pour le futur, il faut couler des fondations solides »	33
Section 1 – La description et le traitement des bases de données	34
2.1.1. Des généralités sur les bases de données de VirtuaMut'	34
2.1.2. Traitements préliminaires des données de VirtuaMut'	36
2.1.3. Des généralités sur les bases de données de l'Open DAMIR	43
2.1.4. Traitements préliminaires des données de l'Open DAMIR	47
Section 2 - La segmentation retenue	52
2.2.1. Segmentation des prestations présentes dans l'Open DAMIR	52
2.2.2. Segmentation des prestations de la mutuelle VirtuaMut'	55
2.2.3. Segmentation retenue pour la tarification	55
Section 3 – Une étude descriptive comparative des deux bases de données	59
2.3.1. Les adhérents	59
2.3.2. Les dépenses	63
Pour faire le point.....	68
Chapitre 3 – « Le prix d'une chose, c'est l'idée qu'on y attache »	69
Section 1 – Rappels théoriques	70
3.1.1. Rappel sur la tarification	70

3.1.2. Les méthodes de tarification retenue par segment de santé	74
Section 2 – Tarification	76
3.2.1. Quelques réflexions préliminaires	76
3.2.2. Un exemple de cas pour la fréquence	77
3.2.3. Un exemple de cas pour le coût	85
3.2.4. Un exemple de cas pour la consommation (version GLM)	89
3.2.5. Un exemple de cas pour la consommation (version statistique)	92
3.2.6. Résultats	93
Section 3 – Extension du tarif	99
3.3.1. Détermination des coefficients d’ajustement	99
3.3.2. Résultats finaux	101
Section 4 – Tests et sensibilités	106
3.4.1. Une question de segmentation	106
3.4.2. Utilisation des modèles optimisés	107
3.4.3. Le seed	108
3.4.4. Modèles basés sur la consommation uniquement	110
3.4.5. Déformation du tarif par l’inflation et les 100 % santé	112
Pour faire le point	113
Conclusion	114
Liste des abréviations	117
Liste d’infographies	120
Bibliographie	123
Annexes	128

Introduction

« Il était une fois ... »

La mutuelle dénommée VirtuaMut'¹ est une petite mutuelle installée dans une région bien précise² en France qui possède un portefeuille très concentré géographiquement puisque 93 % de ses assurés y habitent. Bonne connaisseuse des pratiques et tarifs de son territoire, elle souhaite cependant étendre quelques-uns de ses produits préexistants à une échelle nationale (« France entière », y compris la Corse) et donc, connaître leur tarif s'ils devaient être commercialisés sur d'autres régions françaises. Malheureusement, son budget est très limité et s'avère peu compatible avec l'achat de données. Amatrice cependant de technologies et détentrice d'une équipe efficace en veille informationnelle, elle décide alors de se tourner vers les données publiques (appelées aussi *Open data*) qui elles, sont complètement gratuites et présentent un vaste champ de possibilités et d'applications.

Il s'agira donc pour VirtuaMut' dans un premier temps de mobiliser son équipe dans la recherche de bases de données accessibles et utiles pour une mission de tarification. Cette recherche s'accompagnera d'un petit détour du côté du monde de l'assurance santé et du virtuel afin de se remémorer les notions principales et assurer sa conformité avec le cadre législatif.

Une fois les données sélectionnées, la mutuelle devra ensuite intégrer ces données nationales dans son portefeuille jusqu'à lors restreint. Elle devra alors effectuer une mise en harmonie des deux bases de données (format, types, sélection de données, ...), tout en adaptant ces dernières aux travaux qu'elle entrevoit d'entreprendre. Une attention particulière sera portée à la compréhension et à l'analyse des données, qui représentent en fin de compte la matière première de ses travaux, notamment en ce qui concerne le périmètre couvert par chaque variable. Elle devra aussi trouver une solution aux problèmes qui surgiront lors de la mise en cohérence et qui, parfois, auront pour cause la volonté d'anonymisation des bases de données publiques (niveau d'agrégation des données).

Il s'agira ainsi pour elle dans un second temps d'enrichir une base de données préexistante à partir d'une base en Open data plus générale, qui auront été toutes deux retraitées au préalable.

Enfin, VirtuaMut' pourra se baser sur ces données nettoyées afin d'élaborer des cotisations cohérentes, compétitives et surtout, adaptées aux autres régions pour les produits qu'elle aura choisis. Pour cela, une réflexion doit être faite pour, d'une part, déterminer la classification des actes de soin en segments tarifaires et, d'autre part, élaborer les coefficients de déformation de tarifs afin de tenir compte de l'impact des différentes régions. Un risque d'échec demeure cependant et il faudra alors réfléchir sur la réelle utilité des données publiques retenues dans une telle quête de tarif et garder un esprit critique.

Il s'agira donc dans un dernier temps d'utiliser des méthodes de tarification assez traditionnelles (calcul statistique direct ou modèles linéaires généralisés) afin d'évaluer de nouveaux tarifs.

En résumé, les objectifs principaux tiennent alors en trois mots : enrichir, tarifer, critiquer. Le tout, en abordant les aspects nouveaux des données publiques et du big data et en utilisant des moyens logistiques qui jusqu'à récemment n'auraient pas été envisageables pour le traitement des données sans les progrès technologiques de notre temps.

¹ Ce nom inventé permet de garder l'anonymat de l'organisme originel dont nous avons utilisé les données. Toute ressemblance avec le nom d'un autre organisme est fortuite. Nous nous placerons dans sa structure et exposerons nos travaux comme si nous étions missionnés sur place et faisons partie de l'organisme, afin de rendre les études plus immersives. Ce n'est qu'un choix d'écriture en soi et l'organisme en question ne s'est pas directement impliquée dans les travaux. En réalité, la mutuelle n'ayant pas d'actuaire interne, la mission a été confiée au cabinet de conseil indépendant en actuariat ACTUELIA.

² Pour des raisons d'anonymat, il y a volonté ici de ne pas identifier la région.

Guide de lecture

Ce mémoire est segmenté en trois chapitres. Chaque chapitre est segmenté en sections puis sous-sections (ou sous-parties). Il sera trouvable en chaque début de chapitre, une page introductive et en fin de chapitre, une page récapitulative des principaux points abordés.

En ce qui concerne les annexes, la mention d'une proposition de lecture ne sera pas souvent explicitée en corps de texte. *A contrario*, le lecteur pourra trouver, quand la lecture d'une annexe est pertinente, en marge de page, un signet similaire à celui ci-dessous :



Ce signet est une balise qui, en lecture au format numérique, mènera directement à la page d'intérêt (celle qui contient l'annexe à éventuellement lire pour des informations complémentaires). Pour cela, il suffira de cliquer dessus.

Pour un lecteur de la version en support papier, la balise contient en gras le numéro de l'annexe, puis en dessous, la partie d'intérêt de l'annexe concernée (quand l'annexe est elle-même divisée en parties).

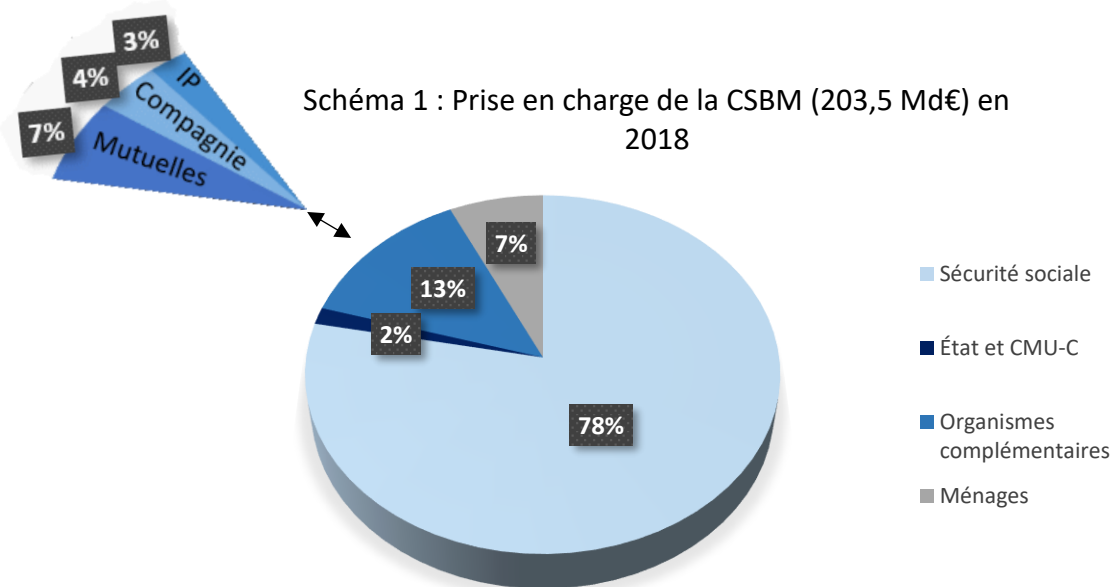
Pour faciliter les allers-retours entre corps de mémoire et annexes, il sera aussi trouvable en marge de pages des annexes, une balise de retour :



Similairement à la balise précédente, il suffira de cliquer dessus pour reprendre la lecture à l'endroit où elle serait *a priori* laissée. Pour la version en support papier, cette balise mentionne la page exacte de retour de lecture.

Chapitre 1 - Quand le monde réel de l'assurance santé s'entrelace avec le monde virtuel de l'open data ...

La France est très connue pour son système de santé et sa couverture sociale. Contrairement à certains pays comme les États-Unis où la facture peut être très élevée à la sortie de l'hôpital, un patient Français n'a souvent qu'à se soucier de sa maladie et de sa vie, et non de son compte bancaire. Si cette sérénité d'esprit face à l'accès au soin est rendue possible, cela ne veut pas pour autant dire que les soins sont administrés gratuitement. Pour preuve, en 2018, la consommation de soins et de biens médicaux (CSBM) s'élève à 203,5 Md€, soit 8,6 % du PIB français et le reste à charge en santé des ménages atteint 7,0 % de la CSBM en 2018, soit environ 210€ par habitant par an (c'est-à-dire, le forfait mobile annuel de la plupart de ces ménages). Mais si ce ne sont pas ces derniers qui paient, il faut pourtant bien que quelqu'un d'autre paie dans l'histoire : la part de la CSBM prise en charge par la Sécurité Sociale s'élève à 78,1 % en 2018 et celle des organismes complémentaires à 13,4 %. Les 1,5 % restants sont à la charge de l'État dans le cadre de la Couverture Maladie Universelle Complémentaire (CMU-C).



Il s'agit alors dans ce chapitre de présenter le monde de la santé, ses acteurs, ses données et ses notions clés afin de replacer ce mémoire et ses travaux dans leur contexte et les ancrer dans une réalité tangible. Ses sections permettront ainsi aux lecteurs non spécialisés dans le domaine de la santé de pouvoir aborder avec plus de clarté les chapitres suivants et aux lecteurs plus aguerris de parcourir leurs souvenirs.

Section 1 - D'un côté, le monde de l'assurance santé...

1.1.1. Les régimes obligatoires

« *La Sécu c'est l'assurance contre le risque de devenir un mauvais risque* » - Anne MARION (Actuarielles)

1.1.1.1. Création et définition

Après une tentative avortée par Jean Jaurès à la suite de son assassinat en 1914 et de la Seconde Guerre Mondiale qui a chamboulé la planète entière, le système de protection sociale de la France a enfin pu voir un jour nouveau : avec les ordonnances d'octobre 1945, la Sécurité Sociale est née sous le gouvernement du Général De Gaulle. À son fondement, trois objectifs principaux furent poursuivis : unicité, universalité, uniformité.

Annexe 1
Histoire

Annexe 1
Les 3 objectifs
initiaux

La Sécurité Sociale peut être définie comme étant un ensemble d'institutions qui ont pour fonction de protéger les individus des conséquences de divers événements ou situations, généralement qualifiés de risques sociaux (ex : maladie, vieillesse, maternité, invalidité, accident du travail, décès, ...). Elle est ainsi destinée à assister financièrement ses bénéficiaires qui rencontrent différents événements coûteux de la vie.

Annexe 1
Organisation

1.1.1.2. Les différents régimes de la Sécurité Sociale

La Sécurité Sociale présente quatre régimes principaux (i.e. un ensemble de droits et d'obligations réciproques des employés et leurs ayants droit, des patrons, et d'une caisse de Sécurité Sociale) qui sont intimement liés à la catégorie socio-professionnelle des individus :

- Le régime général de la Sécurité Sociale couvre l'ensemble des salariés et travailleurs assimilés à des salariés ainsi que leurs ayants droit et compte plus de 62 millions de bénéficiaires, soit plus de 92 % de la population française [1] ;
- Le régime des travailleurs non-salariés non agricoles ;
- Le régime agricole ;
- Les régimes spéciaux (dont le régime local Alsace Moselle).

Annexe 1
Les quatre régimes
principaux

1.1.1.3. Les branches de la Sécurité Sociale

D'un point de vue fonctionnel, la Sécurité Sociale se ramifie en cinq branches qui se chargent de couvrir des risques différents :

- La branche Maladie (maladie, maternité, invalidité, décès, ...) ;
- La branche Accidents du travail et maladies professionnelles ;
- La branche Vieillesse et veuvage (retraite) ;
- La branche Famille (dont logement, RSA, handicap, ...) ;
- La branche Cotisation et recouvrement.

Annexe 1
Les autres
branches

En particulier, la branche Maladie assure la prise en charge des dépenses en santé des assurés et garantit l'accès aux soins, notamment aux personnes démunies. Elle mène aussi des actions de prévention et participe à la régulation du système de santé français. L'un de ses objectifs principaux est d'améliorer l'état de santé de la population française tout en maîtrisant l'évolution des dépenses de santé qui ne cessent d'augmenter avec l'accroissement et le vieillissement de la population.

Elle est gérée par la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS), généralement abrégée en Caisse Nationale d'Assurance Maladie (CNAM), mais aussi par son réseau composé des Caisses Primaires d'Assurance Maladie (CPAM), des Caisses Générales de Sécurité Sociale (CGSS) pour les départements d'outre-mer, des Directions Régionales du Service Médical (DRSM), des Caisses d'Assurance Retraite et de la Santé au Travail (CARSAT) et enfin, des Unions de Gestion des Établissements de Caisse d'Assurance Maladie (UGECAM).

Les prestations de la branche Maladie peuvent être de deux types :

- Les prestations en nature (frais de soin de santé sous condition que le soin soit dispensé par un praticien habilité et soit présent dans la liste des soins remboursables) ;
- Les prestations en espèces (indemnités journalières pour un assuré qui se retrouverait dans l'incapacité physique constatée par son médecin traitant de continuer ou reprendre le travail).

Le principe appliqué ici est le principe indemnitaire : seule la prestation et uniquement cette dernière est remboursée ; il ne peut y avoir d'enrichissement de l'assuré.

Toute personne qui réside ou travaille en France de manière stable et régulière a droit à la prise en charge de ses frais de santé à titre personnel et de manière continue tout au long de sa vie. Pour ce qui est des prestations en espèces, le versement est sous conditions d'heures de travail et de cotisations définies selon la durée de l'arrêt de travail et la situation de l'assuré.

Par ailleurs, une cinquième branche de risque (car la branche Cotisation et recouvrement n'est pas une branche associée à un risque en tant que tel) pour la dépendance³ est depuis une dizaine d'années sujette à réflexion et semble enfin devenir concrète : en effet, dans la nuit du lundi 15 juin 2020, l'Assemblée Nationale a voté favorablement au projet de loi destiné à instaurer cette nouvelle branche de la Sécurité Sociale dans le futur [2]. Des débats sont cependant en cours quant à son financement.

À titre indicatif, les travaux effectués pour ce mémoire se focaliseront uniquement sur la branche maladie et sur le régime général de la Sécurité Sociale.

1.1.1.4. Qu'est-ce que la Sécurité Sociale vaut en tant qu'assureur ?

Le mécanisme de pensées des assurés et donc, leur comportement, est perturbé par de nombreux biais cognitifs que la Sécurité Sociale parvient à contourner.

Par exemple :

- Le biais de la temporalité : c'est le fait de reporter à demain les décisions désagréables. La Sécurité Sociale ne demande pas l'avis de l'assuré actif qui lui doit des cotisations sociales et qui doit être couvert par un contrat collectif d'assurance qui le place sous la couverture de cette première. Même un étudiant a l'obligation d'être affilié au régime général de la Sécurité Sociale.
- Le biais de la loi du petit nombre : les assurés sont plus vigilants aux garanties non anxiogènes (optique, pharmacie, ...) qu'aux garanties anxiogènes (hospitalisation). La Sécurité Sociale se concentre sur le contraire, i.e. sur les priorités, en proposant des prestations pauvres pour l'optique par exemple et élevées pour l'hospitalisation.

La Sécurité Sociale se présente alors comme un bon assureur en dépassant ces biais et en faisant outre des besoins perçus pour privilégier les besoins réels.

³ Pour rappel, la dépendance est définie comme la difficulté voire l'impossibilité d'effectuer soi-même sans aide extérieure au moins trois des cinq actes de la vie quotidienne (s'alimenter, s'habiller, se déplacer, se laver et aller aux toilettes, se coucher et se lever).

Cependant, les prestations versées par la Sécurité Sociale seule sont parfois insuffisantes, notamment en ce qui concerne les prothèses auditives dont le prix moyen jusqu'au 31 décembre 2018 était autour de 1 500 € avec une prise en charge de seulement près de 120 € de la Sécurité Sociale. Par ailleurs, il est observé au fil du temps un désengagement de plus en plus important de la Sécurité Sociale. C'est dans ce contexte-là que les organismes complémentaires prennent sens et justifient leur utilité.

1.1.2. Les régimes complémentaires

« Rien de tel qu'un accident pour nous faire déchiffrer les passages les moins lisibles de notre police d'assurances. » - Anonyme

1.1.2.1. Les différents acteurs

Voici ci-dessous un aperçu des particularités des différents organismes :

Organismes	Réglementation	Mode de fonctionnement	Domaines d'intervention privilégiés
Compagnies d'assurance	Code des Assurances	Société anonyme Société d'assurance mutuelle	Tous les domaines
Mutuelles	Code de la Mutualité	Groupements à but non lucratif, gouvernés par leurs adhérents	Historiquement, les frais de santé mais plus récemment, la prévoyance lourde (décès et arrêt de travail) avec la généralisation de la complémentaire santé en 2016 et l'épargne retraite
Institutions de prévoyance	Livre IX du Code de la Sécurité Sociale	Personnes morales de droit privé à but non lucratif, administrées paritairement par des membres adhérents et des membres participants	Prévoyance collective (notamment grâce aux anciennes clauses de désignations des conventions collectives)

Tableau 1 : Quelques caractéristiques des organismes assureurs

Bien qu'ils soient régis par des Codes différents, ils sont tout de même soumis à des réglementations communes notamment en matière technique telle que Solvabilité II.

L'intérêt de ces organismes est qu'ils versent des prestations qui viennent en complément de celles versées par la Sécurité Sociale et dans certains cas, ils proposent des garanties sur des actes non remboursables par cette dernière.

D'après le registre des organismes d'assurance publié par l'Autorité de Contrôle Prudentiel et de Résolution (ACPR), au 1^{er} janvier 2020, 272 sociétés d'assurances (dont 163 ayant la branche d'agrément 2 - maladie), 370 mutuelles (dont 240 ayant la branche d'agrément 2) et 32 institutions de prévoyance (dont 30 ayant la branche d'agrément 2) ont été dénombrées. Cependant, un phénomène de concentration des mutuelles (fusions, absorptions, faillites) est observé et conséquemment, un nombre de mutuelles qui décline au fil des années (près de 1 200 mutuelles en 2006). Selon l'ACPR encore, cela serait lié à l'entrée en vigueur de nouvelles contraintes réglementaires (généralisation de la complémentaire santé, Solvabilité II).

De plus, les bancassureurs prennent de plus en plus d'importance et de parts du marché : le classement des 20 premiers groupes d'assurance en France par leur chiffre d'affaires de 2018 [4] place Crédit Agricole comme top 1, et CNP et BNP Paribas Cardif ne sont pas loin derrière.

1.1.2.2. Les différents types de contrats d'assurance maladie complémentaire

Un contrat d'assurance maladie complémentaire, souvent abrégé par abus de langage en « complémentaire santé » peut être collectif ou individuel :

- Un contrat collectif couvre un groupe de salariés dans une entreprise et si le contrat le permet, leurs ayants droit. C'est un contrat souscrit par l'employeur pour l'ensemble de ses salariés ou pour une catégorie objective de salariés (ex : les cadres). Depuis le 1^{er} janvier 2016 avec l'entrée en vigueur de la généralisation de la complémentaire santé, il y a obligation de couverture de ses salariés par l'entreprise ;
- Un contrat individuel couvre un particulier et parfois aussi ses ayants droit. Il est à souscription libre et à accès personnel. Selon une étude⁴ de la Direction de la Recherche, des Etudes de l'Evaluation et des Statistiques (DREES) en 2018 portant sur l'année 2014, 95 % de la population française est couverte par une complémentaire santé.

Un contrat d'assurance maladie complémentaire collectif peut aussi être à adhésion obligatoire ou facultative. Il existe des cas de dispenses d'affiliation sous certaines conditions bien précises (ex : être déjà couvert par son conjoint qui a un contrat type famille obligatoire).

En 2018, d'après les chiffres publiés par la Fédération Nationale de la Mutualité Française (FNMF), plus de 47 % des contrats en santé et prévoyance sont des contrats collectifs [5].

Enfin, un contrat d'assurance maladie complémentaire peut être uniforme (même niveau de garantie partout i.e. gamme bâton) ou modulaire (des niveaux de garanties plus élevés que d'autres dans un même contrat). Il est par ailleurs aussi possible de définir des régimes « uniques » (mêmes garanties pour tout le monde) ou des régimes à option (des garanties différentes selon les individus du groupe). Les contrats modulaires sont cependant sources d'anti-sélection et les personnes choisissant de tels types de contrats auront tendance à consommer plus que les autres sur les postes à garanties élevées. Le mémoire de Laetitia FENET, intitulé *Le risque dentaire en assurance complémentaire santé*, aborde ce sujet dans le cas des régimes à options dans sa première partie.

Dans notre cas, ne seront traités que des contrats individuels uniques.

1.1.3. Les notions de base

« *La faiblesse de vocabulaire signifie la faiblesse de penser.* » - Jean-Pierre Raffarin

1.1.3.1. Le remboursement

Considérons un individu âgé de 70 ans qui décide qu'il serait enfin temps de remédier à son problème d'audition de l'oreille gauche. Il aimerait donc s'équiper d'un appareil auditif et après passage chez un audioprothésiste sur prescription d'un médecin spécialiste, il reçoit un devis où il est indiqué qu'une audioprothèse vaut 1 800 € (pour une oreille).

Par ailleurs, l'individu a souscrit à un contrat d'assurance santé individuel où il est indiqué qu'il peut bénéficier d'un remboursement s'élevant à « 40 % de la BRSS + 80 € ». Le régime général de la Sécurité Sociale couvre quant à elle « 60 % de la BRSS ». La BRSS vaut ici 350 € (en 2020). L'individu n'a pas souscrit à une surcomplémentaire⁵.

⁴ DREES (2018), "La complémentaire santé en 2014 [...]", *Etudes et Résultats*, numéro 1048

⁵ C'est un deuxième contrat d'assurance santé qui vient en renfort à la première.

Que devra réellement payer l'individu s'il veut s'équiper ? Répondre à cette question permettra de se familiariser avec le vocabulaire de l'assurance santé.

On appelle **Frais Réel (FR)** ou encore, **dépense engagée**, le prix « brut » du soin ou de l'équipement, avant toute intervention d'un quelconque acteur de la santé. C'est le prix que l'audioprothésiste exige pour monter l'audioprothèse, le prix d'achat de l'équipement.

Ici : dépense engagée = 1 800 €.

La Sécurité Sociale intervient ensuite pour couvrir une partie de ce prix. Elle associe à chaque soin ou équipement une **Base de Remboursement de la Sécurité Sociale (BRSS)**, souvent abrégée en BR) et un **Taux de Remboursement de la Sécurité Sociale (TRSS)**.

Ce que la Sécurité Sociale prendra à sa charge, noté **Montant RO** (pour Montant pris en charge par le Régime Obligatoire), est égal à la BR de l'équipement multipliée par le TRSS de l'équipement.

Ici, montant RO = BR x TRSS = 350 € x 60 % = 210 €.

Le **ticket modérateur (TM)** est le montant restant de la BR après déduction du montant couvert par la Sécurité Sociale.

Soit : 350 – 210 = 140 €.

À ne pas confondre avec ce qui est appelé **dépassement d'honoraire** qui correspond à la dépense engagée déduite de la BR.

C'est-à-dire : Dépassement honoraire = 1 800 – 350 = 1 450 €

L'assureur de l'individu intervient ensuite pour prendre à sa charge une partie du prix restant. Selon le contrat souscrit et les grilles de garantie du contrat, le montant de remboursement de l'assureur (**montant RC**) varie. Ici, le contrat souscrit couvre « 40 % de la BRSS + 80€ ». Cela signifie que le remboursement présente une partie variable (40 % de la BR) et une partie fixe qui constitue un forfait de 80€ (peu importe l'audioprothèse achetée).

Donc : montant RC = BR x taux de remboursement de l'assureur + forfait de l'assureur
= 350 x 40 % + 80 € = 140 € + 80 € = 220 €.

Par ailleurs, l'assureur intervient en second lieu et par principe indemnitaire, il ne peut y avoir d'enrichissement de l'assuré : l'assureur ne peut rembourser plus que le montant restant après passage de la Sécurité Sociale (1 800 – 210 = 1 590 €).

Ainsi :

montant RC final = min(dépense engagée – montant RO ; montant RC précédemment calculé)
= min(1 590 ; 220) = 220 €

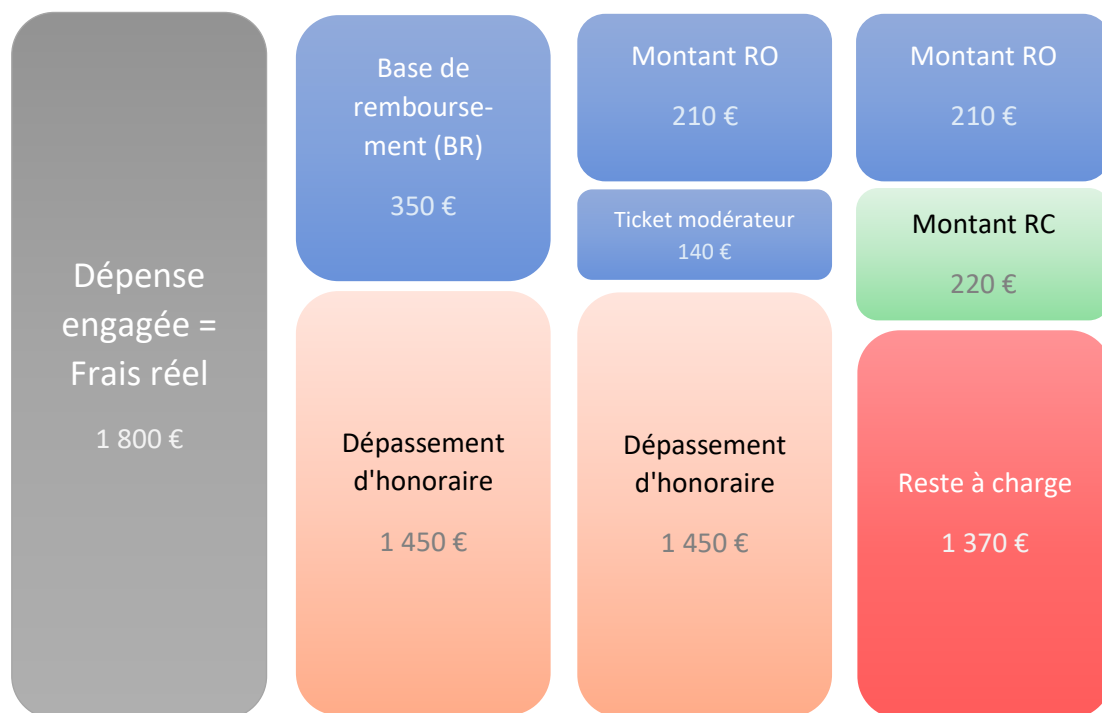
Pour récapituler :

- Dépense engagée = 1 800 €
- Montant RO = 210 €
- Montant RC = 220 €

Après intervention de la Sécurité Sociale et de l'organisme complémentaire, il reste du prix de l'audioprothèse 1 370 €. Ce montant sera à la charge de l'individu (sauf s'il possède une surcomplémentaire). Il est appelé **reste à charge (RAC)**. C'est la réponse à la question initiale, c'est ce qu'il devra réellement payer de son portefeuille personnel.⁶

⁶ Pour certains actes de soin, il existe aussi une participation forfaitaire (cf. 1.1.5.1).

Schéma 2 : Récapitulatif des différents montants



Le montant de la BR à 350 € paraissait ici aléatoire ou arbitraire, mais c'est en fait un montant règlementé et fixé. La sous-section suivante permettra de comprendre en partie son origine.

1.1.3.2. Les grilles de garanties

Lorsqu'une personne souscrit à un contrat d'assurance santé, l'une des premières choses à analyser est la grille de garanties qui lui indique toutes les prestations dont il aura droit de la part de son organisme d'assurance.

Cette grille se présente sous la forme d'une liste organisée de prestations avec leur remboursement associé et ces prestations sont pour la plupart du temps classées selon des grands postes de soin. Une grille se présente donc souvent sous la forme de tableau en blocs. Les organismes sont libres sur la mise en forme et sur la segmentation de leurs grilles tout comme sur l'expression des garanties mais une uniformisation des grilles tend à se faire de plus en plus (contrat responsable, réforme 100 % santé, ...). En effet, avec la réforme 100 % santé par exemple, l'UNOCAM (l'Union Nationale des Organismes Complémentaires d'Assurance Maladie qui rassemble les différentes familles de complémentaires santé) et ses fédérations s'engagent à améliorer et à faciliter la lisibilité des garanties santé par le biais d'intitulés communs et harmonisés sur les principaux postes de prestations à partir du 1^{er} janvier 2020. Il s'agit cependant seulement de recommandations et non d'obligations mais il est quand même observé que les mutuelles sont fortement encouragées à les respecter.

Les grands postes les plus courants sont :

- Les **soins courants** aussi appelés « frais médicaux » ou « soins de ville » constituent l'un des postes de dépenses les plus importants. Il comprend notamment les consultations chez le médecin ou le spécialiste ou encore les soins effectués par les auxiliaires médicaux (kinésithérapie, soins infirmiers, ...);
- L'**hospitalisation** est un poste principalement constitué des prestations liées à une visite ou opération à l'hôpital : frais de séjour, forfaits journaliers, chambre particulière...;

- L'**optique** avec les montures, les verres, les lentilles, la chirurgie des yeux... ;
- Le **dentaire** avec les prothèses dentaires, l'orthodontie et les soins dentaires ;
- L'**aide auditive** avec les audioprothèses et autres appareillages auditifs.

Une segmentation plus fine des postes de soins est proposée plus tard en Section 2 du Chapitre 2. Cette dernière constitue un élément central et impactant dans les travaux effectués.

Enfin, l'expression des garanties peut prendre différentes formes :

- En pourcentage de la BR y compris la part prise en charge par la Sécurité Sociale ;
- En pourcentage de la BR sans la part prise en charge par la Sécurité Sociale ;
- En pourcentage des frais réels ;
- En pourcentage du plafond mensuel de la Sécurité Sociale (PMSS). Celui de 2019 s'élève à 3 428 €.
- Via un forfait, qui peut être appliqué acte par acte, annuellement, semestriellement, ...

Il peut y avoir un mélange d'expressions comme dans l'exemple de la partie précédente où un forfait est couplé à un pourcentage de la BR.

1.1.4. Nomenclatures des actes

Les actes pris en charge par la Sécurité Sociale suivent une codification particulière qui permet, entre autres, l'identification des actes dans les outils de gestion et une généralisation des tarifs appliqués selon la nature de l'acte en définissant notamment des BR normalisées. Plusieurs nomenclatures existent et ne seront abordées ici que les principales. La fiche Annexe 2 pourra être lue en complément de cette sous-section pour plus de détails.

1.1.4.1. La Nomenclature Générale des Actes Professionnels (NGAP)

La Nomenclature Générale des Actes Professionnels (NGAP) de 1972 définit principalement les honoraires des praticiens du secteur libéral, des sages-femmes, des dentistes et des auxiliaires médicaux (infirmiers, orthodontistes, orthoptistes, kinésithérapeutes, ...). Chaque acte est associé à une lettre « clé » et à un coefficient qui permettent de définir la cotation de l'acte. La lettre « clé » est affectée à une base de remboursement unitaire et le coefficient définit le coefficient multiplicatif à appliquer à cette base de remboursement unitaire.

Par exemple, le code acte B5 se décompose en la lettre « clé » B pour « Biologie » dont la base de remboursement unitaire associée est 0,27 €⁷ et du coefficient 5. Ainsi :

BR pour le code B5 = 0,27 € x 5 = 1,35 €.

Par ailleurs, certains actes peuvent être complétés par un forfait en plus de leur cotation propre.

1.1.4.2. La Classification Commune des Actes Médicaux (CCAM)

La Classification Commune des Actes Médicaux (CCAM) définit en France la grande majorité des actes techniques médicaux, des honoraires des médecins du secteur libéral, des prestations en établissements de soin et depuis 2014⁸ des actes bucco-dentaires. Elle recense actuellement plus de 7 500 codes actes sur 19 chapitres organisés selon des zones anatomiques du corps humain. L'objectif fut d'affiner la tarification des actes par la considération de nouveaux critères tels que la durée et la difficulté de l'acte, le stress du praticien et le matériel utilisé.

⁷ La lettre B est associée au montant unitaire de 0,27 € en France métropolitaine, à 0,31 € aux Antilles et à 0,33 € en Guyane et en Réunion.

⁸ Avenant n°3 à la convention nationale des chirurgiens-dentistes - Journal officiel du 30 novembre 2013.

Une première version de la classification a été publiée le 1^{er} janvier 2002 et fut le fruit de travail de centaines d'experts. Ce n'est qu'au 13 août 2004 avec la loi de réforme de l'Assurance Maladie (art L.162-1-7 du code de la Sécurité Sociale) qu'elle devient une référence obligatoire pour les actes techniques médicaux. Elle succède au Catalogue des Actes médicaux (CdAM) et à la NGAP qui reste cependant en vigueur pour les actes précédemment cités.

Les codes actes sous la nomenclature de la CCAM sont assez précis et plus techniques que ceux de la NGAP et se composent de quatre lettres suivies par trois chiffres :

- La première lettre désigne un appareil anatomique (c'est-à-dire un ensemble d'organes qui concourt à la réalisation d'une tâche commune complexe) tels que l'appareil digestif, l'appareil cardiovasculaire, l'appareil respiratoire, ... ;
- La deuxième lettre désigne l'organe concerné qui fait partie de l'appareil en question ;
- La troisième lettre désigne l'action effectuée sur la partie ciblée du corps ;
- La quatrième lettre précise la technique utilisée ;
- Les trois chiffres permettent ensuite de peaufiner la différenciation.

Par exemple, le code « DBAF004 » désigne une « dilatation intraluminale de l'orifice atrioventriculaire gauche, par voie veineuse transcutanée avec perforation du septum interatrial » (chapitre 4).

À chaque code est associé un tarif de base. Ces codes peuvent être complétés par des codes « modificateurs » qui viennent majorer le tarif sur la base d'un critère particulier pour la réalisation de l'acte ou d'une valorisation jugée nécessaire compte tenu du contexte de réalisation de l'acte (urgence, jour férié, ...). Il existe aussi des codes de remboursement exceptionnel (pour les actes exceptionnellement pris en charge).

Cette nomenclature présente cependant des codes bien trop fins : la chirurgie faisait autrefois partie de la NGAP mais il n'était pas indiqué la partie du corps concernée par l'opération, or, à présent, il est même possible de savoir la technique de chirurgie utilisée. Ainsi, par ces codes, il est tout à fait envisageable de déterminer la pathologie d'un patient, ce qui est contraire au principe de confidentialité et de secret médical. De ce fait, les organismes complémentaires disposent alors plutôt de codes de regroupement qui donnent des informations plus générales.

Annexe 2
Exemple de calcul

1.1.4.3. Tarification à l'Activité (T2A)

La Tarification à l'Activité (T2A) concerne principalement les prestations dans les établissements hospitaliers et est un élément essentiel du Plan Hôpital 2007. Elle constitue une refonte des modes de financement et de tarification de l'environnement hospitalier et fut mise en place en deux temps : immédiatement dès mars 2005 pour le secteur privé et progressivement pour le secteur public. Elle a permis d'harmoniser les principes de facturation entre le secteur public et le secteur privé en rendant comparables les coûts, ce qui a permis d'éviter les difficultés de facturation provenant notamment d'un transfert d'un patient d'un établissement public à un établissement privé (ou vice-versa).

La T2A introduit les notions de Groupe Homogène de malade (GHM) et Groupe Homogène de séjour (GHS), notions clés de la tarification des prestations hospitalières. En fait, lorsqu'un patient est à l'hôpital, son séjour est classifié dans un GHM qui est identifié par un code alphanumérique (ex : GHM 08M04W) combiné à un libellé (ex : « fracture de la hanche et du bassin avec CMA (CoMorbidity Associée) » pour le code précédent). Ce GHM est par la suite associé à un GHS (défini par l'Assurance Maladie) qui est associé quant à lui à un tarif donné. À chaque type de séjour correspond un seul et unique GHS.

Enfin, dans certains cas particuliers, des compléments peuvent venir majorer le tarif comme par exemple pour le cas d'une prise en charge d'un acte particulièrement lourd nécessitant des unités de soin très spécialisées.

À savoir que les éléments de ces nomenclatures sont sujets à changer dans le temps et que quelques nomenclatures se voulaient être évolutives : certains codes actes peuvent alors être amenés à disparaître et il y a actuellement plus de cinquante versions de la CCAM.

Annexe 2
GHS
Les autres
nomenclatures

Il est possible de trouver plus de détails sur les nomenclatures dans le mémoire de Élodie PAGET (*Amélioration de l'outil de tarification santé d'Actélior à partir des actes codés avec la Classification Communes des Actes Médicaux, 2008*) où elle s'attarde par exemple plus sur les prédécesseurs de la CCAM et propose aussi des statistiques sur les codes actes dans son chapitre 2.

1.1.5. Un point sur la législation : Santé

1.1.5.1. Contrat responsable

La loi de Douste-Blazy du 13 août 2004 et un décret du 29 septembre 2005 ont instauré les participations forfaitaires laissées à la charge des assurés par le régime général, le parcours de soins avec un médecin traitant ainsi que le dossier médical personnel.

Annexe 3
Dossier médical
personnel

Depuis le 1^{er} janvier 2005, il existe une franchise de 1 € pour tout acte, consultation ou visite réalisée par un médecin, tout examen de radiologie et tout acte de biologie médicale. Un plafond annuel de 50 € est fixé (par année civile et par bénéficiaire) avec un plafond journalier de 4 € (en cas de plusieurs actes réalisés par le même praticien).

Depuis le 1^{er} janvier 2008, il existe aussi une franchise de 0,50 € par boîte de médicaments ou acte paramédical et de 2 € par transport sanitaire. Le plafond est aussi fixé à 50 € par année civile par bénéficiaire avec un plafond journalier de 2 € pour les actes paramédicaux et 4 € pour les transports.

Quant à la notion de parcours de soins, une personne est dite « hors parcours de soins » lorsqu'elle n'a pas de médecin traitant ou qu'elle consulte un autre médecin sans passer par son médecin traitant. Cette notion a été introduite pour limiter la surconsommation des consultations de spécialistes et pour assurer un suivi personnalisé auprès d'un médecin.

En ce qui concerne la caractéristique de « responsable », notion en vigueur depuis le 1^{er} janvier 2006, un contrat d'assurance maladie est dit « responsable » lorsqu'il satisfait différentes conditions qui permettront au souscripteur de bénéficier d'aides sociales et fiscales. Les conditions sont contenues dans un cahier des charges qui a été redéfini par un décret du 18 novembre 2014 et qui a récemment évolué pour inclure la réforme 100 % Santé en critère.

Ces contrats doivent exclure de leurs garanties la prise en charge de :

- La participation forfaitaire de 1 € ou toute autre participation ;
- Les dépassements d'honoraires sur les actes cliniques et techniques hors parcours de soins ;
- La majoration de participation sanctionnant l'absence de choix ou de recours au médecin traitant.

A contrario, ces contrats doivent inclure la prise en charge :

- Des 7 actes de prévention fixés par arrêté interministériel du 8 juin 2006 ;
- Du forfait journalier hospitalier de manière illimitée ;

- Du ticket modérateur (à l'exception des médicaments remboursés à 15 % et 30 % dont l'homéopathie et les cures thermales) ;
- Si le contrat prend en charge les dépassements d'honoraires pratiqués par les médecins, il y a obligation de différencier la prise en charge selon que le médecin est adhérent ou non à l'OPTAM (Option de pratique tarifaire maîtrisée) i.e. l'ex-CAS (Contrat d'Accès aux Soins).

À titre indicatif, le CAS est le fruit d'une négociation nationale entre l'Assurance maladie et les syndicats médicaux. Il est proposé à quelques médecins du secteur 1 et aux médecins de secteur 2⁹ et doit favoriser l'accès aux soins pour les patients en leur permettant d'être mieux remboursés. Le médecin s'engage pour trois ans mais peut changer d'avis tous les ans à la date d'anniversaire du contrat. Il s'engage à figer ses tarifs et à limiter son dépassement d'honoraire contre certains avantages. Au 1^{er} janvier 2017, le CAS est remplacé par l'Option de pratique tarifaire maîtrisée (OPTAM) qui le simplifie et qui se différencie entre autres par des paiements plus rapides et des sorties de contrat à tout moment.

1.1.5.2. 100 % Santé

Lors de sa campagne présidentielle de 2017, Emmanuel Macron avait promis une réforme du système de santé français avec l'ajout du « reste à charge zéro » (connu à présent sous le nom de « 100 % Santé » pour des raisons de communication). Promesse tenue puisque fin octobre 2018, l'article 33 de la Loi de Financement de la Sécurité Sociale (LFSS) 2019 est adopté par l'Assemblée générale et la mise en œuvre progressive du 100 % Santé dès 2019 est mise en marche.

L'objectif premier de la réforme est de lutter contre le renoncement aux soins pour des raisons financières en proposant à tous les Français couverts par un organisme d'assurance (et ayant un contrat dit « responsable ») l'accès à des paniers de soins dont au moins un se démarque par son reste à charge à 0. Trois postes sont concernés par la réforme : l'optique, l'audio et le dentaire. L'assuré n'est bien entendu pas contraint à choisir le panier « zéro » (celui qui aboutit à un reste à charge nul) et reste libre de choisir des paniers à tarifs libres qui ne le restreindraient pas sur les matériaux et designs de ses équipements.

Pour réaliser cet objectif, la réforme s'appuie sur trois leviers : une modification des bases de remboursement de la Sécurité Sociale (tendance haussière pour une majorité des équipements), des prix limites de vente des équipements et le reste à charge nul.

Cette réforme a impacté le monde de la santé dans son intégralité : les assurés qui se voient ouvrir une nouvelle alternative de soins sans frais, la Sécurité Sociale et les assureurs qui doivent prendre à leur charge plus que ce qu'ils ont l'habitude de payer et les professionnels de santé qui se voient obligés de proposer ces paniers « zéros » dans leur devis avec des prix limites de vente pratiqués et de changer leur manière de communiquer.

1.1.5.3. Les dispositifs de solidarité

Il existe des alternatives qui sont possibles grâce au principe de solidarité afin de donner l'accès aux soins aux individus ne disposant pas de complémentaire santé et/ou ceux n'ayant ni le statut, ni une activité professionnelle leur permettant de cotiser et d'accéder à une couverture sociale normale :

- La Protection Universelle Maladie (PUMA) : entrée en vigueur le 1^{er} janvier 2016, elle permet à toute personne qui travaille ou réside en France de manière stable et régulière, un droit à la

⁹ Les médecins du secteur 1 appliquent des tarifs dits de convention car fixés par convention avec l'assurance maladie obligatoire. Ils ne pratiquent pas de dépassement d'honoraires sauf en cas de demande du patient. Les médecins du secteur 2 appliquent des tarifs dits d'autorité car libres et pratiquent des dépassements d'honoraires.

prise en charge de ses frais de santé à titre personnel et de manière continue tout au long de la vie ;

- Aide Médicale d'État (AME) : elle permet aux ressortissants étrangers en situation irrégulière sous conditions de résidence stable et de ressources la prise en charge de certains soins et actes ;
- La Couverture Maladie Universelle Complémentaire (CMU-C) qui est une sorte de complémentaire santé pour les personnes ayant un revenu modeste ;
- L'Aide à l'acquisition d'une Complémentaire Santé (ACS) qui est une aide financière attribuée par le régime général de santé lorsque le bénéficiaire dispose de revenus faibles mais pas suffisamment pour obtenir la CMU-C.

Annexe 3
CMU-C
ACS

La CMU-C et l'ACS ont cependant fusionné le 1^{er} novembre 2019 pour devenir la Complémentaire Santé Solidaire (CSS) et bien que gratuite pour les bénéficiaires de la CMU-C, elle demande une participation financière pour les anciens bénéficiaires de l'ACS. Les garanties de la CMU-C sont gardées.

1.1.5.4. Autres textes sur l'assurance maladie

Il existe d'autres textes fondateurs comme :

- La loi Evin du 31 décembre 1989 qui définit les règles de déontologie communes à tous les acteurs du marché de l'assurance collective. Dans les articles phares, il est possible de citer l'article 7 sur le maintien des droits en cas de résiliation (il oblige aussi les assureurs à se constituer une provision pour sinistres à payer) ; l'article 7.1 sur le maintien de la garantie décès en cas d'arrêt de travail ; l'article 2 sur la non-sélection médicale et les prestations non différenciées et l'article 4 sur la portabilité des droits pour les anciens salariés ;
- L'accord National Interprofessionnel du 11 janvier 2013 et la loi relative à la sécurisation de l'emploi de 2013 qui généralisent la complémentaire santé et la rendent obligatoire pour tous les salariés en entreprise. L'ANI permet aussi de conserver le bénéfice des garanties prévues au contrat collectif en vigueur dans l'ancienne entreprise en matière de frais de santé et de prévoyance pendant une période limitée à 12 mois ;
- La loi Fillon de 2003 sur notamment l'exonération des charges sociales et fiscales ;
- ...

mais pour éviter de rendre trop exhaustive cette sous-section, nous nous contentons de les mentionner sans plus de détails.

Section 2 - ... De l'autre, le monde de l'open data.

1.2.1. Un second point sur la législation : Open Data

Le domaine de la santé connaît un véritable élan digital où l'*Open data* serait un vecteur de transparence et un levier vers une meilleure gestion médicale des patients. Tous les jours, une quantité astronomique de données est générée puisque la santé est un sujet qui nous suit quotidiennement : le marché mondial de la donnée est estimé à 6 milliards de dollars en 2020 [6]. En effet, l'information a une valeur monétaire et constitue donc un nouveau type de patrimoine. Avec cette émergence digitale viennent aussi ses propres acteurs tels que les *start-ups* spécialisées en santé ou les GAFAs (Google, Apple, Facebook et Amazon) qui s'introduisent dans le secteur.

Face à cette euphorie et à ces mutations, un cadre réglementaire a dû se créer pour pouvoir réguler les changements de l'environnement digital, éviter les écarts et protéger le secret médical, gage de confiance entre patient et praticien, en garantissant une sécurisation des données. Ce cadre s'est cependant construit progressivement puisque le sujet est sensible et concerne tout le monde : les données touchent à la vie privée des citoyens qui peuvent se sentir attaqués dans leur intimité et elles sont aussi convoitées car inestimables pour de nombreux acteurs. Il était donc nécessaire de concilier *Open data* et droit de vie privée, innovation et éthique, utilité et sécurité.

Tout d'abord, en novembre 2003, l'Union Européenne a invité ses États membres à réfléchir sur le sujet et éventuellement à permettre l'utilisation des données publiques. Dix ans après, en juin 2013, cela devient une obligation pour les acteurs publics (l'État, les collectivités territoriales et toutes autres extensions dont la CNAM). Une consultation publique sur l'ouverture des données de santé a néanmoins été lancée en novembre 2013 par le Ministère des Affaires sociales afin de recueillir l'avis de la population qui reste en somme plutôt méfiante. Un des arguments avancés à l'époque en faveur de l'ouverture est le scandale du médicament Mediator (clôturé en 2009) qui consistait en un détournement d'utilisation qui aurait potentiellement pu prendre moins d'ampleur si les données (anonymisées) à ce sujet avaient été rendues disponibles pour des études. Certains experts [7] étaient aussi d'avis que des conditions d'ouverture devaient être posées.

Ainsi, plusieurs textes de loi ont vu le jour. Parmi eux, il est possible de citer la loi de santé et « Open Data », ou encore Loi n° 2016-41 du 26 janvier 2016 de modernisation du système de santé français, qui crée deux groupements. D'une part, le Système National des Données de Santé (SNDS) dont les objectifs sont divers :

- Centraliser les données des bases existantes en matière sanitaire et médico-sociale, c'est-à-dire, des données par exemple issues des systèmes d'information des établissements de santé ou encore de l'Assurance maladie, des informations transmises par les mutuelles, des données sur les causes de décès, ... ;
- Les conserver sous un format permettant l'anonymat pendant une durée de 20 ans (absence du nom des patients, de leur adresse ou de leur numéro de Sécurité Sociale).

La gestion des bases est confiée à la CNAM qui est aussi responsable du traitement des données.

D'autre part, l'Institut National des Données de Santé (INDS) qui se porte garant des usages de l'*Open data*, qui contrôle le SNDS et qui veille sur la qualité des données.

Cette loi différencie aussi les données où aucune identification n'est possible et auquel cas, elles peuvent être accessibles et réutilisables par tous de manière gratuite ; des données potentiellement identifiables dont l'utilisation est encadrée et limitée à, par exemple, des utilisations pour de la recherche, des études ou des évaluations sur la santé. Pour ces données, il faut nécessairement obtenir l'autorisation de la

Commission Nationale de l'Informatique et des Libertés (CNIL)¹⁰ et soumettre le fruit de ses résultats à l'INDS. La loi interdit aussi l'utilisation des données à certaines fins comme pour la modification des cotisations pour un groupe d'individus partageant le même risque.

Deuxième texte important qui devait être mentionné : le règlement européen du 14 avril 2016 qui pose un cadre uniforme et identique sur la protection des données pour l'ensemble du territoire de l'Union Européenne. Plus communément connu sous le nom de RGPD (Règlement Général sur la Protection des Données), ses directives sont entrées en applications récemment (le 25 mai 2018) et visent une responsabilisation dans le traitement des données personnelles de la part des organismes publics et privés (dont font partie les organismes assureurs). Voici ci-dessous quelques principes clés de ce règlement :

- Un cadre harmonisé est posé par ce règlement qui s'applique directement dans tous les États membres. Il n'y a pas de nécessité de transposition contrairement à son prédécesseur, la directive sur la protection des données personnelles de 1995 ;
- Les individus se doivent d'être plus informés (objectif de transparence). Ils doivent donner explicitement leur consentement quant à l'utilisation de leurs données ;
- Ils doivent être informés en cas de fuite de données. Les entreprises doivent aussi notifier sous 72h à l'autorité compétente tout piratage de données à caractère personnel ;
- Il y a possibilité de désigner un délégué à la protection des données (DPO, pour *Data Protection Officer*) au sein des organisations. Il s'assure notamment du respect du règlement ;
- Un Comité européen de la protection des données a été créé. Il a autorité sur tout ce qui concerne l'interprétation du RGPD ;
- En cas de non-respect, des sanctions graduelles (avertissement, injonction, limitation ou suspension temporaire, définitive) sont prévues.

À savoir que la CNIL et l'Agence des Systèmes d'Information Partagés de Santé (ASIP Santé) ont pour mission de protéger les données de santé et que toute diffusion illicite de ces dernières expose le coupable à une peine de 5 ans de prison et 300 000 € d'amende.

1.2.2. Un petit état de l'art de l'Open Data en matière de santé

« *La France est le pays qui a la plus grande base de données médico-économique au monde.* » - Simon CHIGNARD, chargé de mission à Etalab [8].

1.2.2.1. Un nombre considérable de jeux de données en Santé

Il existe de très nombreuses bases de données publiques sur le domaine de la santé. Le site du gouvernement (data.gouv.fr), administré par Etalab, une administration publique française qui apparaît comme étant un *Chief Data Officer* (CDO) pour l'État, rassemble les données publiques produites ou reçues dans le cadre d'une mission de service public ainsi que différentes restitutions anonymes de données du SNDS. Elle laisse à la disposition du public, au 14 juin 2020, 569 bases qui vont du ciblé (ex : localisation des toilettes publiques) au global (ex : Comptes nationaux de la santé) et il convient de choisir celles adaptées à son étude. Ce site centralise de nombreuses bases qui proviennent pour certaines d'autres plateformes comme celle de l'Assurance Maladie (*ameli*) ou encore, celle de Eco-santé de la DREES.

Les séries de données sont par ailleurs fournies par des instances publiques voire gouvernementales telles que le Ministère des Solidarités et de la Santé, la Caisse Nationale d'Assurance Maladie (CNAM)

¹⁰ Autorité indépendante compétente en matière de protection des données personnelles.

et la Haute Autorité de Santé (HAS, qui avait commandité par ailleurs l'établissement de la nomenclature CCAM).

Les thèmes couverts sont très divers : consommation de soins, efficacité du système santé, médicament, l'offre de santé, etc. Il existe par ailleurs une cartographie des bases de données publiques en santé¹¹ qui avait été élaborée dans le cadre d'une mission de l'Etalab faisant suite aux débats sur l'ouverture des données publiques de santé lancée par le Ministère des Affaires Sociales et de la Santé en novembre 2013. La mission entreprenait de recenser l'ensemble des bases de données publiques disponibles mais ce recensement est désormais devenu obsolète puisque datant de 2014. À l'époque, plus de 260 bases ou jeux de données avaient pu être dénombrés et chacune d'entre elles faisait l'objet d'une évaluation de « son niveau d'ouverture » déterminé par 4 critères (la liberté d'accès, le coût d'accès, le format de mise à disposition et les conditions juridiques de réutilisation) et de sa granularité (niveau granulaire ou agrégé). Ce niveau d'ouverture permet ainsi de très vite évaluer la pertinence d'une base vis-à-vis de ses objectifs d'étude et la cartographie en elle-même permet de centraliser l'existence des bases et de rendre toute recherche hasardeuse moins chronophage : elle aurait été pertinente pour la mutuelle VirtuaMut' qui n'a pas d'idée précise initialement de ce qu'elle souhaite utiliser. Malgré son obsolescence, elle reste tout de même un bon point de départ de recherche.

1.2.2.2. *Le SNIIRAM, la plus grosse base de données en Santé*

Le Système National Inter Régimes d'Assurance Maladie (SNIIRAM) est une base de données nationale créée en 1999 avec la loi de financement de la Sécurité Sociale (LFSS) et apparaît comme étant la plus grosse base de données de santé en France et une des plus importantes au monde. Elle est complète, détaillée, anonyme et renseigne sur le parcours des patients et l'organisation du système de soins. Son enrichissement et sa construction sont par ailleurs progressifs : par exemple, les dates de décès n'ont été intégrées que dix ans plus tard en 2009. Elle est un réel investissement puisqu'en plus d'avoir pris 10 ans à la CNAM pour la construire, plus de 200 personnes sont nécessaires pour la gérer et la maintenir à jour.

Ses objectifs sont quadruples et inscrits dans l'article L161-28-1 du Code de la Sécurité Sociale :

- Contribuer à une meilleure gestion de l'Assurance Maladie ;
- Contribuer à une meilleure gestion des politiques de santé ;
- Améliorer la qualité des soins ;
- Transmettre aux professionnels de santé les informations pertinentes sur leur activité.

Ses modalités (périmètre, finalités, alimentation et accès) sont définies dans un arrêté du Ministère des Affaires Sociales et de la Santé. La CNAM est chargée de sa gestion et se présente comme étant la responsable du système d'information aux yeux de la CNIL.

Sa procédure d'anonymisation passe par le contrôle et le cryptage des flux de données de santé entrants et par une anonymisation de façon irréversible des numéros de Sécurité Sociale des bénéficiaires. Pour autant, il a été estimé qu'il restait possible de reconnaître certains individus dans les données.

Le SNIIRAM est constitué des informations issues des remboursements effectués par l'ensemble des régimes d'assurance maladie pour les soins du secteur libéral. Des informations sur les séjours hospitaliers sont également disponibles. Cette base s'organise en :

- Un ensemble de 15 bases de données agrégées ou individualisées, par thématiques, appelées *datamarts*, orientées vers un objectif bien précis : suivi des dépenses (avec l'Open DAMIR), biologie, pharmacie, ... ;

¹¹ <https://www.data.gouv.fr/fr/datasets/cartographie-des-bases-de-donnees-publiques-en-sante/>

- Un Échantillon Général des Bénéficiaires (EGB) qui permet entre autres d'analyser le parcours individuel de près de 660 000 bénéficiaires en ville et à l'hôpital ;
- Une base de données individuelles des bénéficiaires (DCIR) pour réaliser des études sur la consommation des soins.

L'historique accessible des données est variable (20 ans pour l'EGB, sans limites pour les *datamarts*) et leur archivage dure 10 ans. Les archives ne sont consultables qu'après une autorisation émise par la CNIL.

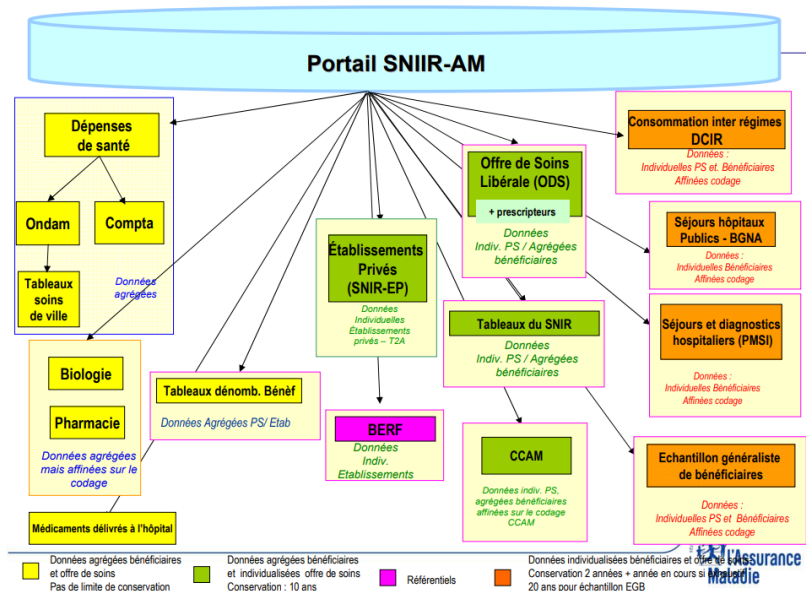


Schéma 3 : Structure du SNIIRAM [11]

En termes de contenu, sont disponibles à la restitution aux utilisateurs :

- Des données sur les patients telles que l'âge, le sexe, s'il est bénéficiaire de la couverture maladie universelle complémentaire (CMU-C) ou non, sa localisation (commune et département de résidence), la date de décès, ... ;
- Toutes les prestations remboursées dans le cadre des soins réalisés en médecine de ville : des informations sur le prestataire de soins et éventuellement le prescripteur (spécialité, mode d'exercice, sexe, âge, département d'implantation), le codage détaillé, la date des soins et les montants remboursés par l'Assurance maladie et payés par les patients ;
- Des données sur la consommation de soins en établissement : centralisation des données relatives aux séjours facturés directement à l'Assurance maladie, des données sur les pathologies traitées.

Un dictionnaire (*wiki-sniiram*) est par ailleurs mis à disposition afin de faciliter l'appropriation de la base. Il énumère les variables du SNIIRAM et documente les données et les règles de gestion associées. Cette base présente cependant des limites. En effet, malgré sa complétude et son exhaustivité, il y a peu d'informations sur les affections psychiatriques et dermatologiques alors que ce sont des effets secondaires fréquents des médicaments ; sur les données sociales (hors la mention de la CMU-C) ; sur les antécédents personnels, ...

Accéder à cette base de données serait une véritable aubaine pour la mutuelle VirtuaMut' (et pour les mutuelles en général) puisque qu'une quantité astronomique de données y est inscrite (une volumétrie de plus de 18 téraoctets).

Malheureusement, son accès est règlementé, sur mot de passe, tracé, et les utilisateurs doivent être formellement et nominativement identifiés. Un arrêté du 19 juillet 2013 liste les organismes autorisés à y accéder (parmi eux, les régimes d'assurance maladie et de nombreux partenaires - services

ministériels, agences sanitaires, organismes publics de recherche). Des démarches de demande d'accès sont aussi à prévoir et une autorisation d'instances compétentes telles que la CNIL est nécessaire. Les organismes poursuivant un but lucratif sont exclus de cette liste, ainsi que les organismes assureurs. Il faudra donc s'orienter vers un autre jeu de données.

En revanche, une partie du SNIIRAM est disponible en tant qu'*Open data* : l'EGB et l'Open DAMIR par exemple.

1.2.2.3. Un diaporama des bases de données en santé

Pour éviter une sur-exhaustivité et car nous ne prétendons pas faire un travail de cartographie aussi abouti que la cartographie précédemment mentionnée en partie 1.2.2.1., nous nous contenterons dans la suite de ne mentionner que certaines bases ou jeux de données publics qui nous ont paru intéressants pour la mutuelle VirtuaMut' ou pour une visée de tarification de manière plus générale. Nous essayerons de même d'émettre notre avis sur la pertinence de ces bases.

Titre :	Compte de santé de la DREES : prise en charge des différents postes de la consommation de soins et biens médicaux
Source :	www.ecosante.fr Irdes d'après données Comptes de la santé de la Drees
Contenu :	Financement de la dépense de soins par la Sécurité Sociale, les ménages et les organismes complémentaires : montant des prestations prises en charge par les trois acteurs de 2006 à 2014 sur les grands postes de soin (médicament, dépenses courantes de santé, soins ambulatoires, soins hospitaliers, ...)
-	Pas spécifiquement utile pour une tarification ou un enrichissement de base. Ne couvre que de 2006 à 2014 donc est désormais obsolète.
+	Permet d'avoir une idée globale de la part prise en charge par la Sécurité Sociale, les organismes complémentaires et le reste à charge des ménages (étude macroscopique).

Titre :	Open DAMIR : base complète sur les dépenses d'assurance maladie inter régimes
Source :	http://open-data-assurance-maladie.ameli.fr/depenses/index.php
Contenu :	C'est une extraction du SNIIRAM qui est mise à disposition du public. L'ensemble des prestations prises en charge par l'Assurance Maladie obligatoire y compris les prestations hospitalières facturées directement à l'Assurance Maladie pour l'ensemble des régimes avec des informations sur le patient, son prescripteur, le praticien de son soin et l'établissement de soin.
-	Fichiers de données trop lourds (besoin de matériel logistique particulier pour exploiter), difficilement exploitable. Données agrégées avec méconnaissance du nombre d'individus par ligne. Mise à jour tous les ans en juin.
+	55 variables, bases très complètes qui donnent beaucoup d'informations que ce soit sur les assurés, les praticiens, les prescripteurs ou les établissements de soin. Permet un enrichissement de bases préexistantes. Est une des bases de données les plus complètes dans le domaine de la santé.

Titre :	Dépenses d'assurance maladie hors prestations hospitalières (données nationales)
Source :	http://open-data-assurance-maladie.ameli.fr/depenses/index.php
Contenu :	L'ensemble des remboursements mensuels effectués par le régime général de l'Assurance Maladie (hors prestations hospitalières) par type de prestations, type d'exécutant et par type de prescripteurs. Extraction moins détaillée de l'Open DAMIR (28 variables, dont la plupart en communes avec l'Open DAMIR).

-	Les informations sont déjà incluses dans l'Open DAMIR. Pas de données hospitalières. Peu d'informations sur les patients (âge, sexe, région, ...).
+	Est suffisant pour certaines études. Mise à jour mensuellement donc l'année 2020 est déjà disponible en partie. Est plus explicite au niveau de la nature des prestations.

Titre :	Dépenses d'assurance maladie hors prestations hospitalières par caisse primaire/département
Source :	http://open-data-assurance-maladie.ameli.fr/dépenses/index.php
Contenu :	L'ensemble des remboursements mensuels effectués par le régime général de l'Assurance Maladie (hors prestations hospitalières) par caisse primaire/département, par type de prestations, par type d'exécutant et par type de prescripteurs. Extraction moins détaillée de l'Open DAMIR mais plus riche que la base de dépenses d'assurance maladie hors prestations hospitalières (37 variables).
-	Pas de données hospitalières. Donne plus d'informations que la base précédente notamment des précisions sur la localisation des patients mais rien sur leur âge (qui est une variable discriminante dans les tarifs en santé).
+	Est largement suffisant pour certaines études. Mise à jour mensuellement donc l'année 2020 est déjà disponible en partie. Est plus explicite au niveau de la nature des prestations. Donne plus de précisions que l'Open DAMIR sur la localisation des individus (ici, découpage en départements disponible).

Titre :	Dépenses annuelles d'assurance maladie
Source :	https://www.data.gouv.fr/fr/datasets/dépenses-annuelles-d-assurance-maladie
Contenu :	Rétrospective annuelle de l'ensemble des remboursements du régime général de l'Assurance Maladie effectués en France métropolitaine, par type de risque : maladie, maternité, invalidité et décès, accident du travail et maladie professionnelle. Pour chaque type de risque, les dépenses sont présentées par catégorie de professionnels de santé et, pour chaque catégorie, par acte ou par groupe d'actes.
-	Il n'y a surtout que des données récapitulatives qui peuvent être déduites de l'Open DAMIR et cela ne couvre que de 2010 à 2016.
+	Donne une vision très globale en cas d'impossibilité d'exploitation de l'Open DAMIR.

Tableau 2 : Aperçu de l'état de l'art sur les *Open data* en santé

De nombreuses autres bases existent cependant, parmi elles :

- Des bases de données relatives à l'actualité et aux catastrophes récentes comme par exemple les bases de données sur le COVID-19 présentant des données hospitalières ou concernant les tests de dépistage. Par principe, ces bases se doivent d'être créées assez rapidement et tenues régulièrement à jour pour être pertinentes et servir de fondation d'étude dans un temps d'urgence et d'appréhension.
- Des bases de données sur les médicaments avec la base de données publique des médicaments, l'Open Medic, l'Open PHMEV ;
- Des bases de données sur les renoncements aux soins ou relatives aux décès ;
- Des bases un peu plus atypiques voire surprenantes comme la localisation des toilettes publiques ou encore les profils et trajectoires des personnes ayant des idées suicidaires.

1.2.3. Un focus sur l'Open DAMIR

La mutuelle VirtuaMut', après avoir fait son état de lieu des bases de données existantes concernant la santé, décide de partir sur l'Open DAMIR car c'est la base la plus complète adaptée à l'étude visée, ouverte au public, gratuite, assez explicite grâce à son lexique mis à disposition et que la mutuelle se place dans une réflexion portée sur le long terme : se familiariser avec des jeux de données aussi massifs lui servira sûrement dans le futur pour d'autres problématiques.

L'Open DAMIR est en fait une extraction du SNIIRAM (encadré jaune « Dépense de santé » du schéma 3 de la page 25) et fait suite à l'*Hackaton* organisé en janvier 2015 par la CNAM dans sa démarche d'ouverture des données de santé. À la différence du SNIIRAM, elle est accessible sans demande d'autorisation préalable y compris par les organismes d'assurance. Elle vient en complément des bases préalablement mises en ligne de :

- Dépenses d'assurance maladie hors prestations hospitalières (données nationales) ;
- Dépenses d'assurance maladie hors prestations hospitalières par caisse primaire/département ;
- Dépenses annuelles d'assurance maladie ;

Et permet ainsi d'aborder des axes complémentaires d'étude.

En libre accès depuis le 26 janvier 2015 et gérée par la CNAM, l'Open DAMIR contient l'ensemble des prestations prises en charge par l'Assurance Maladie obligatoire (à l'exception d'une grande majorité des prestations hospitalières du secteur public) et est restituée sous la forme de fichiers Excel .csv (un fichier par mois). Pour des raisons d'anonymat des patients mais aussi des professionnels de santé (anonymat qui pour rappel est imposé par la réglementation), les données sont proposées sous forme agrégées : une ligne d'un fichier est donc une somme des actes et des montants associés à un ensemble d'individus ayant une quarantaine de caractéristiques communes (i.e. variables catégorielles telles que l'âge, le type d'acte, le sexe, ...).

La base présente de plus des données sous différents axes :

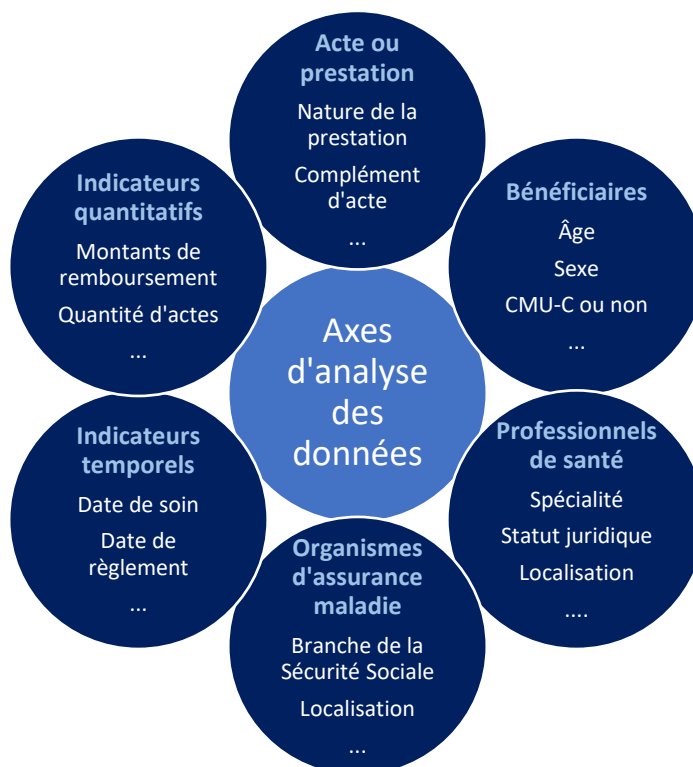


Schéma 4 : Les axes d'analyse des données de l'Open DAMIR

Les axes géographiques sont cependant limités par souci d'anonymat à 9 zones géographiques (regroupements de régions administratives) de 2009 à 2014 ; puis à 13 zones géographiques (proches des grandes régions administratives créées par la réforme territoriale de 2014) à partir de 2015.

De manière plus précise, un fichier, donc un mois, présente 55 variables (la liste complète des variables est présente en Annexe 7), une trentaine de millions de lignes et a une taille de plus de 4 gigaoctets (Go). Sur un périmètre de deux ans, cela reviendrait à devoir travailler avec plus de 720 millions de lignes, une centaine de gigaoctets de données et il faut savoir que l'historique actuel de la base de données est de 10 ans (de 2009 à 2019), donc à éventuellement multiplier par 5 en cas d'étude sur l'historique complet. La base se met à jour annuellement approximativement au mois de juin, ce qui revient à avoir les données de 2020 en juin 2021.

Annexe 7

Le monde de la santé est cependant en constante évolution : prenez la crise sanitaire du COVID-19 qui ravage le monde au moment de la rédaction de ce mémoire ; pensez-vous qu'il serait pertinent d'attendre juin 2021 pour pouvoir étudier les conséquences et l'évolution de la pandémie ? En fait, dès le 21 mars 2020, le portail de données publiques data.gouv.fr proposait déjà des jeux de données sur la pandémie issus de l'Agence nationale de santé publique (organisme gouvernemental chargé d'une mission de surveillance épistémologique) [10]. Dans un contexte d'urgence et dans un monde de données en flux constants, cette base ne semble alors pas adaptée pour des études de faits d'actualité mais plus pour des études de faits du passé. Son homologue, la base de Dépenses d'assurance maladie hors prestations hospitalières par caisse primaire/département semble alors mieux convenir puisqu'elle est mensuellement mise à jour. Pour autant, cette dernière ne présente pas assez d'information sur l'assuré, notamment sur son âge. De plus, dans une optique de tarification, des données très récentes ne sont pas primordiales pour garantir des résultats pertinents. Ainsi, choisir l'Open DAMIR paraît donc plus adapté pour la visée choisie.

Cependant, par sa taille conséquente et ses lignes nombreuses, la mutuelle VirtuaMut' s'est retrouvée devant un problème majeur quand elle a souhaité traiter et manipuler les fichiers de données : ouvrir un seul fichier de manière complète ne fut pas envisageable sur les outils et programmes à sa possession. À titre indicatif, par exemple, l'ouverture n'était pas réalisable sur Excel qui est limité par un peu plus d'un million de lignes sur une feuille ou encore, sur R dont l'importation de la base en table de données (*dataframe*) s'est soldée par un échec et cette opération s'est même avérée être nuisible pour les ordinateurs sensibles et peu récents. Ainsi, la mutuelle VirtuaMut' se devait de chercher une solution logistique de substitution qui lui permettrait de travailler avec l'Open DAMIR. C'est de cela que traite la sous-section suivante.

La difficulté de traitement sans logistique adaptée mais peu souvent présente chez les petites mutuelles, couplée à la latence de mise à jour et à d'autres facteurs, fait donc émerger une seconde problématique au mémoire (dont la réponse ne pourra être apportée qu'à la fin) : l'Open DAMIR a-t-elle une utilité réelle pour un organisme d'assurance ?

[1.2.4. Les machines virtuelles : une solution logistique à la pointe de la technologie](#)

« *Un ordinateur créé à l'intérieur d'un ordinateur* » - Définition de Microsoft Azure¹².

Cette section a pour objectif de vous présenter la solution logistique retenue ainsi que la démarche jusqu'à sa mise en marche opérationnelle. Il s'agira aussi de répondre aux questionnements qu'un futur utilisateur pourrait être amené à se poser en se basant sur notre propre vécu et l'ensemble des échanges

¹² <https://azure.microsoft.com/fr-fr/overview/what-is-a-virtual-machine/>

avec les différents services (*Microsoft*, *SAS*, professeurs, ...) que nous avons réalisés - échanges qui étaient pour la plupart en anglais.

La première idée naïve fut de chercher un moyen d'améliorer les capacités de son propre ordinateur mais après test, il s'avère que même un ordinateur pour joueurs de jeux vidéo (« *gamers* ») qui est optimisé pour les jeux dynamiques en temps réel et qui a donc de bonnes performances n'est pas forcément adapté pour les traitements et calculs envisagés (du moins, une gamme ordinaire avec un budget moyen). Avec des connaissances limitées en composantes d'ordinateur et en montage, cette solution semblait risquée en coût et très peu malléable une fois les composants achetés (par exemple, si la mémoire vive initialement considérée se révèle au final être insuffisante).

La deuxième idée fut d'utiliser une machine virtuelle et c'est elle qui a été retenue, en partie pour sa flexibilité et la possibilité de changer de configuration en cas de nécessité. Une machine virtuelle est une sorte d'émulation, de logiciel informatique « image » qui se comporte comme un ordinateur réel. Plusieurs machines peuvent être utilisées en même temps sur un même ordinateur physique. Une machine virtuelle marche indépendamment de l'ordinateur physique : celle-ci est en quelque sorte « isolée », si bien que les logiciels installés et les actions effectuées dans la machine ne modifient *a priori* pas l'ordinateur hôte.

Sur le marché, il existe différents fournisseurs de machines virtuelles dont *Amazon* avec sa branche *Amazon Web Services (AWS)* ou *Microsoft* avec *Microsoft Azure*. C'est ce dernier qui a été choisi pour principalement sa compatibilité avec le système d'exploitation utilisé en entreprise et la familiarité de l'interface d'utilisation (*Windows 10*).

La première étape cruciale lors du montage¹³ d'une machine virtuelle est de savoir quelle configuration choisir et les choix sont nombreux :

- Le système d'exploitation : *Linux* ou *Windows*, les deux sont possibles mais pour des raisons de simplicité et d'habitude, *Windows* a été sélectionné dans notre cas ;
- L'emplacement de la machine virtuelle : *a priori* France Centre. Exceptionnellement, le début du mémoire a été réalisé avec une machine en Europe de l'Ouest car la crise sanitaire de la COVID-19 restreignait les possibilités de montage et les machines en France Centre étaient principalement réservées aux chercheurs et professionnels de santé dans le but de faire des avancées sur la pandémie. Il a donc fallu faire une demande d'accès au serveur français. La localisation a son importance avant tout pour optimiser les performances mais ne pas choisir la bonne n'est pas drastique.
- Les caractéristiques de la machine : *Azure* propose de nombreuses combinaisons de puissance de calculs, de mémoire vive et de nombre de processeurs selon les tâches à réaliser avec la machine. Par exemple, la gamme de machines D est optimisée pour l'usage général alors que la gamme F pour les calculs. Nous avons retenu la configuration de machine suivante : Standard F16s - 16 processeurs virtuels et 32 Go de RAM (mémoire vive).
- La taille des disques de stockage : il existe des disques de stockage « temporaires » et des disques de stockage « persistants ». Ces derniers permettent de retenir les données même une fois la machine éteinte. Toute sauvegarde sur un disque temporaire disparaît au moment d'éteindre la machine (et l'arrêt est fréquent pour limiter les coûts). Nous avons retenu dans un premier temps un disque dur persistant de 1 To. Les disques permanents ont des coûts fixes mensuels qui dépendent de la capacité de stockage.

¹³ Le terme « monter » la machine virtuelle sera utilisé et correspond aussi au terme utilisé par le portail de *Microsoft Azure* mais il s'agira concrètement plutôt de la configurer : tout est virtuel, il n'y a rien de physique, il n'y a pas de pièces détachées d'ordinateur à manipuler. Une fois que la machine est prête et activée, son interface est la même que l'interface habituelle de *Windows 10* : c'est exactement comme utiliser son propre ordinateur sauf qu'ici, les performances de calculs et de traitement seront beaucoup plus puissantes.

Ensuite, il a fallu se renseigner sur comment configurer la machine pour que l'accès à distance puisse être possible. Techniquement, si le montage devait être fait de manière optimale pour un réseau d'entreprise, la formation prendrait son temps et des compétences informatiques pourraient être prérequis pour limiter le temps d'apprentissage. Le point sensible étant la sécurisation du système monté. Dans le cas d'une étude non confidentielle et brève, nous conseillons de ne pas s'attarder sur les notions informatiques trop complexes : la plupart des paramètres par défaut lors du montage sont suffisants.

Une fois la machine configurée et prête, il faut s'atteler à l'installation des programmes utiles : il faut s'imaginer être devant un ordinateur tout neuf qui n'a rien dessus et faire le nécessaire pour installer ce qu'il faut. Pour rappel, pour qu'un programme soit installé durablement, il faut l'installer dans la mémoire permanente de la machine. Il est aussi possible de transférer des fichiers de son ordinateur personnel à sa machine virtuelle.

À titre indicatif, il n'y a pas besoin de savoir coder quoi que ce soit pour monter la machine virtuelle : tout peut se faire via le portail de *Microsoft Azure*. Des guides sont par ailleurs disponibles et mis à disposition par *Microsoft* afin de guider pas à pas l'utilisateur dans le montage de sa machine, un programme de e-learning (*Microsoft LEARN*) existe aussi et fonctionne par modules pour permettre à l'utilisateur d'avancer et d'apprendre à son rythme. De plus, à la souscription aux services d'*Azure*, une séance de présentation est organisée gratuitement sur souscription volontaire et permet d'avoir une vue d'ensemble des possibilités.

Ainsi, bien qu'au premier abord, un tel montage peut paraître insurmontable, il est au final réalisable pour une petite structure n'ayant pas de service IT et n'ayant pas de connaissances au préalable en informatique. Par la suite, les travaux réalisés ont été effectués sous *SAS* pour le traitement des données et sous *R* pour l'exploitation des données et le tout, sous machine virtuelle d'*Azure* (Standard F16s - 16 processeurs et 32 Go de RAM – 1 To de stockage).

Pour faire le point

Le monde de l'assurance santé est constitué de nombreux acteurs qui coexistent et interagissent entre eux :

- La Sécurité Sociale avec ses différents régimes et ses différentes branches, intervient en premier lors d'un remboursement d'une prestation santé. Pour rappel, seul le régime général et la branche maladie (frais de soin) seront étudiés par la suite ;
- Les organismes complémentaires compensent en partie l'insuffisance de remboursement de la Sécurité Sociale ;
- L'assuré (ou le patient) prend à sa charge le résiduel restant ;
- Les autres intervenants (professionnels de santé, gestionnaires, ...) ont chacun leur propre rôle.

Les prestations sont classées dans des postes de soin et suivent pour la plupart du temps une nomenclature générale (CCAM par exemple) qui permet de leur associer une base de remboursement, montant sur lequel la Sécurité Sociale et les organismes complémentaires se basent en majeure partie pour déterminer leur part de prise en charge.

Le monde de la santé a ses propres lois et quand ce dernier rencontre le monde du virtuel et de l'innovation avec les données publiques (*Open data*), de nouvelles règles s'ajoutent afin de pouvoir concilier protection de la vie privée des constituants et amélioration du système de santé.

De nos jours, les bases de données en matière de santé sont abondantes, le SNIIRAM étant la plus fournie mais demeure inaccessible pour une mutuelle. Son extraction, l'Open DAMIR, est une *Open data* et sera choisie comme élément d'étude pour cet écrit. Cependant, étant tout de même massive, une nouvelle solution logistique doit être considérée : une machine virtuelle, solution de notre temps.

Chapitre 2 – « Si vous essayez de bâtir pour le futur, il faut couler des fondations solides »¹⁴

Comme mentionné en introduction, l'objectif opérationnel de ce mémoire est double. D'une part, VirtuaMut' souhaite enrichir ses bases de données de prestations et d'adhérents avec des données publiques nationales. D'autre part, elle souhaite ensuite exploiter son nouveau portefeuille ainsi plus riche afin de pouvoir élaborer un tarif pour certains de ses produits sur d'autres régions de la France que celle de sa localisation. Pour des raisons de commodité et de confidentialité, nous appellerons par la suite sa région de localisation « *Dreamland* », notée *D*. Les autres régions seront nommées *Otherland i* (avec $i \in \{1, 2, \dots, 12\}$), notées O1, ..., O12.

Pour ces travaux, nous avons pour rappel décidé de retenir l'Open DAMIR en tant que base de données publique.

Les jeux de données à disposition provenant de la mutuelle sont les suivants :

- Une base de données sur les sinistres survenus en 2018 avec les prestations versées ;
- Une base de données sur les sinistres survenus en 2019 avec les prestations versées ;
- Une base des adhérents de la mutuelle avec leurs caractéristiques.

Une description plus détaillée de ces jeux de données sera effectuée dans la suite du mémoire.

Seules les garanties¹⁵ A et B¹⁶ feront l'objet d'études. Ce sont les deux garanties les plus souscrites par les adhérents de VirtuaMut' puisque parmi ceux affiliés au régime général de la Sécurité Sociale, environ 15 % d'entre eux présents en 2018 ou en 2019 ont choisi d'adhérer à la garantie A, 79 % à la garantie B et 6 % du reste se répartissent dans les autres garanties proposées par la mutuelle.

Un tableau synthétique des prestations proposées par la mutuelle dans le cadre de ces garanties est présent en Annexe 5. Y sont aussi présentes les hypothèses de prestations garanties par la Sécurité Sociale pour quelques types d'actes de soin.

Annexe 5
Prestations
VirtuaMut'

Par ailleurs, la garantie A est une garantie d'entrée de gamme, la garantie B est une garantie de milieu de gamme. Pour rappel, la sous-partie 1.1.3. donne des informations basiques et générales sur les grilles de garanties et renseigne sur comment ces grilles sont utilisées pour déduire un montant de remboursement de la mutuelle.

Il s'agira dans ce chapitre de décrire et retraiter ces différentes bases de données afin de les faire converger vers un format identique et les rendre prêtes à l'exploitation dans une démarche de tarification. Une classification des différents actes de soins et leur répartition dans des segments tarifaires seront aussi abordées.

¹⁴ Ridley SCOTT, *Robin des bois*, Robin Longstride

¹⁵ Ceci est un abus de langage. Il faudrait techniquement parler de contrat, produit ou police d'assurance.

¹⁶ Noms arbitraires par souci d'anonymat.

Section 1 – La description et le traitement des bases de données

2.1.1. Des généralités sur les bases de données de VirtuaMut'

2.1.1.1. Bases de sinistres

Initialement, les deux bases de sinistres de VirtuaMut' de 2018 et 2019 présentaient toutes deux dix variables informatives :

Nom de la variable	Description
REF_PERSONNE	(Variable qualitative nominale) Identifiant unique de l'adhérent dans les bases de données de VirtuaMut'
CODE_ACTE	(Variable qualitative nominale) Code acte du soin selon une nomenclature propre à la mutuelle (cf. Section 2 du présent chapitre)
LIBELLE_ACTE	(Variable qualitative nominale) Libellé associé au code acte qui traduit en langage intelligible la nature de l'acte de soin
DATE_SOINS	(Variable qualitative ordinale) Date de délivrance du soin i.e. date à laquelle le soin est effectivement exécuté
DATE_REGLT	(Variable qualitative ordinale) Date de règlement du soin (remboursement de la mutuelle à son adhérent)
DEPENSE	(Variable quantitative continue) Dépense engagée i.e. frais réel
MT_RO	(Variable quantitative continue) Montant RO
MT_RC	(Variable quantitative continue) Montant RC
QTE_ACTE	(Variable quantitative discrète) Quantité d'actes
OPTION	(Variable qualitative nominale) Garantie souscrite par l'adhérent

Tableau 3 : Les dix variables initiales des jeux de données des prestations de VirtuaMut'

Les bases de prestations sont par ailleurs des bases par année de survenance (ou encore, par année de soin), c'est-à-dire que la base de 2018, par exemple, recense l'ensemble des prestations versées aux adhérents de la mutuelle au titre des actes de soin délivrés en 2018. Ces soins peuvent être réglés par la mutuelle à une date ultérieure de celle de délivrance. Les prestations réglées couvertes par les bases de VirtuaMut' sont de manière exhaustive :

- Les prestations remboursées en 2018 au titre des soins effectués en 2018 ;
- Les prestations remboursées en 2019 au titre des soins effectués en 2018 ;
- Les prestations remboursées en 2020 au titre des soins effectués en 2018 ;
- Les prestations remboursées en 2019 au titre des soins effectués en 2019 ;
- Les prestations remboursées en 2020 au titre des soins effectués en 2019 ;
- Les prestations remboursées en 2021 au titre des soins effectués en 2019 ;

L'Open DAMIR est cependant une base par année de règlement, de ce fait, les bases de 2018 et 2019 couvrent a minima :

- Les prestations remboursées en 2018 au titre des soins effectués en 2016 ;
- Les prestations remboursées en 2018 au titre des soins effectués en 2017 ;
- Les prestations remboursées en 2018 au titre des soins effectués en 2018 ;
- Les prestations remboursées en 2019 au titre des soins effectués en 2017 ;
- Les prestations remboursées en 2019 au titre des soins effectués en 2018 ;
- Les prestations remboursées en 2019 au titre des soins effectués en 2019 ;

En cas d'années de soin antérieures à 2016, les lignes qui s'y réfèrent sont supprimées.

Il est d'usage de travailler avec des bases par année de soin (technique) plutôt que par année de règlement (comptable). L'optimal aurait été donc d'avoir pour l'Open DAMIR des informations sur les règlements en 2020 et 2021, ce qui n'est pas possible sans attendre 2022 ; et de supprimer les lignes relatives aux années de soin antérieures à 2018. Nous avons décidé que faute d'information et afin de garder un périmètre cohérent d'étude, il était préférable de considérer une étude en année de règlement (sur la base de l'Open DAMIR, dont nous avons souhaité éviter de poser des hypothèses non informées à son sujet). Pour cela, l'hypothèse retenue pour les bases de données de VirtuaMut' est celle de stabilité des prestations, équivalente à considérer que le futur est représentatif du passé. Cela est justifié dans le sens où le portefeuille de VirtuaMut' n'a que très peu évolué au fil du temps. Ainsi, par exemple, les **prestations de 2019 réglées en 2020** (dans les bases de VirtuaMut') sont considérées comme étant suffisamment équivalentes à des **prestations de 2017 réglées en 2018**. Le reste étant négligeable. Il faudra cependant être conscient qu'un biais en naît. Ceci soulève alors une des premières limites de l'Open DAMIR en tant que base pour une visée de tarification : sans hypothèse simplificatrice supplémentaire, travailler en années de soin implique de ne pas travailler avec l'historique d'années le plus récent.

Par la suite, les deux jeux de données de VirtuaMut' ont été fusionnés et ont été filtrés selon les deux garanties retenues puis séparés de nouveau : un fichier pour la garantie A et un autre pour la garantie B. À ce stade, les fichiers étant différenciés en garantie, la variable OPTION devint obsolète et a été supprimée. Chaque jeu possède donc désormais 9 variables.

	A	B	Autres	Total
Base 2018	76 520	543 348	18 176	638 044
Base 2019	92 225	641 398	21 055	754 678
Total	168 745	1 184 746	39 231	1 392 722

Tableau 4 : Dénombrement de lignes pour chacune des garanties

Comme il est possible de l'observer dans le tableau ci-dessus, la grande majorité des lignes sont générées par des individus ayant souscrit aux garanties A et B, ce qui confirme d'autant plus l'intérêt de se limiter à ces deux dernières pour nos travaux (prendre l'une des dizaines autres garanties revient à prendre le risque de travailler avec relativement peu de données et donc, d'aboutir à des résultats peu concluants). Il a été préférable de travailler dans un premier temps sur la garantie A qui présente environ 7 fois moins de lignes que la garantie B afin de tester et de faire tourner les algorithmes plus rapidement. Le traitement de la base associée à la garantie A est le même que celui appliqué à la garantie B, de même pour la démarche. Choisir deux garanties permettra ultérieurement de les comparer.

2.1.1.2. Bases des adhérents

En ce qui concerne la base des adhérents, elle présente dix variables permettant de caractériser chaque adhérent de la mutuelle.

Nom de la variable	Description
REF_PERSONNE	(Variable qualitative nominale) Identifiant unique de l'adhérent dans les bases de données de VirtuaMut'
DATE_NAIS	(Variable qualitative ordinale) Date de naissance de l'adhérent
DEPT	(Variable qualitative nominale) Département de résidence de l'adhérent
QUALITE	(Variable qualitative nominale) Qualité de l'adhérent, c'est-à-dire s'il est le souscripteur ou s'il est un ayant droit (conjoint ou enfant)
SEXE	(Variable qualitative nominale) Sexe de l'adhérent
OPTION	(Variable qualitative nominale) Garantie souscrite par l'adhérent

DATE_DEBUT	(Variable qualitative ordinale) Date de début du contrat de l'adhérent
DATE_FIN	(Variable qualitative ordinale) Date de fin du contrat de l'adhérent
REGIME	(Variable qualitative nominale) Régime de la Sécurité Sociale (cf. partie 1.1.1., par exemple, régime général ou local)
TNS	(Variable qualitative nominale) Indicateur binaire de statut de Travailleur Non Salarié ¹⁷ (1 = TNS, 0 = Non TNS)

Tableau 5 : Les dix variables initiales du jeu de données des adhérents de VirtuaMut'

La plupart de ces variables seront utilisées au profit de la création de nouvelles variables et devront subir des changements de format afin d'être cohérentes avec les variables constituant l'axe du bénéficiaire de l'Open DAMIR (cf. schéma 4, sous-section 1.2.3.).

De plus, la variable **TNS** ne sera pas utilisée dans les travaux bien qu'elle permet d'affiner les tarifs car elle est trop peu fiable : un adhérent de VirtuaMut' ayant le statut de TNS est effectivement un TNS mais un adhérent n'ayant pas ce statut n'est pas pour autant un non TNS ; l'information n'est pas forcément bien déclarée et la mutuelle est consciente de cela.

À savoir qu'un assuré peut apparaître plusieurs fois (donc être associé à plusieurs lignes) dans cette base des adhérents. Plusieurs possibilités peuvent expliquer cela : l'adhérent change de garantie en cours d'année, il change de statut marital (cette information n'est pas disponible dans nos extractions de données), il change d'adresse postale, ou encore, il résilie son contrat et se souscrit à nouveau par la suite (avec possibilité d'avoir un laps de temps où il n'est plus couvert par VirtuaMut'). En soi, tout changement de gestion peut amener la mutuelle à recodifier des adhérents. C'est le cas pour environ 66 adhérents du régime général.

2.1.2. Traitements préliminaires des données de VirtuaMut'

Pour rappel et par souci de redondance, sans mention supplémentaire, nous présenterons la démarche effectuée sur le fichier de prestations associé à la garantie A mais elle sera identiquement applicable pour la garantie B. L'ensemble des traitements est fait sous R.

Par ailleurs, au vu du nombre conséquent de retraitements faits, seuls les retraitements principaux ou ayant un impact notable seront mentionnés. Le reste, ainsi que des détails complémentaires, pourront être trouvés en Annexe 6.

2.1.2.1. Identification de l'acte

Dans un premier temps, il a été associé à chaque ligne de la base des prestations pour la garantie étudiée sa classification qui sera présentée en section suivante (sous-famille de l'acte associée au **CODE_ACTE** du soin, famille d'actes et grand poste de soin) afin de pouvoir identifier l'acte et la ligne dans la grille de garanties qui fut appliquée à ce dernier. Trois variables supplémentaires sont alors créées : **SS_FAM**, **FAM** et **POSTE** pour respectivement la sous-famille d'acte, la famille d'actes et le grand poste associé. C'est à partir de ces nouvelles variables que la segmentation pour la tarification sera créée.

2.1.2.2. Rajout des informations sur les adhérents

Dans un second temps, la base des adhérents a été fusionnée à la base des prestations via la variable clé

¹⁷ Un Travailleur Non Salarié (TNS) est un chef d'entreprise qui a un statut différent de celui d'un salarié et qui bénéficie d'un régime ayant, par exemple, des cotisations moindres par rapport au régime général de la Sécurité Sociale tout en ayant les mêmes prestations.

d'identification **REF_PERSONNE** commune aux deux jeux de données et la variable **OPTION**. Le principe est de pouvoir associer à chacune des lignes présentes dans la base de sinistres plus d'informations sur le patient : sa date de naissance, son département, sa qualité, son sexe et son régime de Sécurité Sociale. Les données ont ensuite été filtrées sur la variable **REGIME** afin de ne garder que les adhérents affiliés au régime général de la Sécurité Sociale (périmètre d'étude).

Annexe 6.1

2.1.2.3. Création des variables **AGE** et **CLASSE_AGE**

La variable **AGE** (variable quantitative discrète entière) a été créée grâce à la date de naissance de l'adhérent (**DATE_NAIS**) et à la date de soin de l'acte (**DATE_SOINS**) en calcul exact (tout comme pour l'Open DAMIR), arrondie au plus proche. **DATE_NAIS** a ensuite été supprimée pour cause de redondance d'information. En effet, nous travaillerons avec des bases de données lourdes (Open DAMIR) et dans cette optique, il est nécessaire de limiter les données au strict minimum pertinent afin de pouvoir alléger les temps de calcul des algorithmes.

La variable **CLASSE_AGE** a ensuite été créée conformément au format de son homologue dans l'Open DAMIR sur la base de la variable **AGE**, selon la table de correspondance ci-dessous :

CLASSE_AGE	Tranche d'âge du bénéficiaire au moment des soins
0	0-19 ans
20	20 - 29 ans
30	30 - 39 ans
40	40 - 49 ans
50	50 - 59 ans
60	60 - 69 ans
70	70 - 79 ans
80	80 ans et +
99	Âge inconnu

Tableau 6 : Table de correspondance des âges de l'Open DAMIR

Il est à noter qu'il n'y a pas de classe d'âge « 99 » dans les données de VirtuaMut'.

Annexe 6.2
Variable SEXE

2.1.2.4. Les vérifications traditionnelles

Par la suite, des vérifications traditionnelles ont été réalisées et ont amené à la suppression des lignes :

- Arborant une quantité ou des montants de remboursement négatifs car ces valeurs sont sans doute liées à des corrections ou des régularisations effectuées dans le cas par exemple d'un sur-remboursement détecté par la Sécurité Sociale ou par la mutuelle ;
- Affichant à la fois des montants de dépenses engagées et de remboursement de la mutuelle à 0 ;
- Dont la somme des montants RO et RC dépasse les dépenses engagées (sauf pour le cas des compléments et suppléments d'actes) ;
- Dont les montants de remboursement de la mutuelle n'étaient pas cohérents avec les grilles de garanties afférentes (il n'y en avait pas) ;
- Où l'adhérent avait un âge jugé impossible (par exemple, 130 ans, en sachant que la doyenne de l'humanité i.e. la personne détenant le record de longévité humaine est Jeanne Calment avec ses 122 ans¹⁸) (il n'y en avait pas).

¹⁸ Selon l'article *Doyens de l'humanité : l'homme et la femme les plus vieux du monde*, datant du 24 juillet 2020 et publié sur le site <https://www.notre-planete.info/actualites/2780-homme-femme-plus-vieux-monde>.

2.1.2.5. Création de la variable REGION

La variable **DEPT** renseignant sur le département de résidence de l'assuré présente les valeurs suivantes :

- 1 à 95 pour les départements de la France métropolitaine tels que nous les connaissons (i.e. le département codé 75 correspond bien à Paris et celui codé 31 à la Haute-Garonne) ;
- 97 pour les DOM-TOM ;
- L pour Luxembourg et B pour Belgique.

Cette variable a été utilisée pour pouvoir créer la variable **REGION** qui est en adéquation avec les valeurs prises par son équivalent dans l'Open DAMIR. Pour cela, chaque département a d'abord été associé à une région puis cette région a été associée à sa codification dans l'Open DAMIR. Un tableau de correspondance est présent en Annexe 9. Dans le cas du Luxembourg ou de la Belgique, la variable **REGION** prend la valeur « 99 » (valeur inconnue) et les lignes associées ne seront pas prises en compte dans les travaux car elles sont considérées comme hors périmètre d'étude.

Annexe 9

2.1.2.6. Retraitement de la variable QTE_ACTE

La quantité d'actes renseignée par VirtuaMut' correspond au coefficient dit « global » qui permet de définir la cotation de l'acte dans la nomenclature NGAP (cf. partie 1.1.4.). C'est donc une quantité déduite du code acte de la ligne et du montant de dépense engagée. Par exemple, pour une ligne donnée, si la dépense engagée s'élève à 4,73 € et que le code acte de la ligne est « AMI » (Acte Médico-Infirmier comme les perfusions ou les prises de sang) avec une base de remboursement unitaire associée à 3,15 €, alors la quantité d'actes s'élèvera à $4,73/3,15$, soit environ 1,5.

Cela est cependant un élément de divergence avec l'Open DAMIR qui propose une variable de quantité d'actes à valeur entière. Cette dernière prend supposément une vision purement réaliste dans le sens où un acte de soin n'est *a priori* pas réalisé à moitié (pour être dans l'excès illustratif, l'infirmière prendra par exemple sa prise de sang de manière complète et ne s'arrêtera pas en plein milieu de son opération).

Toutes les quantités d'actes ne sont pas systématiquement déterminées de la même façon : quand les éléments sont télétransmis, la codification est automatique entre les flux Noémie et l'outil de gestion ; cependant, en cas de saisie manuelle de l'acte par un gestionnaire, des valeurs aberrantes peuvent apparaître. Ce sont avant tout les quantités à virgule ou les quantités supérieures à 1 qui sont concernées par une éventuelle mise en cohérence avec l'Open DAMIR (elles n'étaient pas suffisamment nombreuses pour qu'un traitement manuel soit impossible).

Le retraitement s'est effectué sur la considération des prix moyens pour les différents actes de soin et de leur nature. Il a été ainsi déduit une quantité d'actes plus cohérente avec l'Open DAMIR. Une incertitude sur la fréquence naît de ce traitement mais il était bien trop chronophage de tenter une stratégie plus précise que celle décrite ci-dessus. De plus, il a fallu mettre à zéro la quantité d'actes pour tous les compléments et suppléments d'actes qui ne sont pas des actes en tant que tels puisqu'ils sont rattachés à ces derniers (c'est le cas par exemple des majorations pour travail de nuit ou pour travail en jour férié, etc.). Ceci évite de majorer de manière injustifiée les fréquences d'actes de soin.

Annexe 6.3
Exemple

2.1.2.7. Traitement des quantités nulles

Il fallait ensuite rattacher ces compléments¹⁹ d'actes à leur acte de référence et donc, sommer leurs montants de dépense (**DEPENSE**), de remboursement par la Sécurité Sociale (**MT_RO**), de

¹⁹ Par abus de langage, dans la suite, le terme « compléments » désignera à la fois les compléments d'acte, les suppléments d'acte et les majorations.

remboursement par la mutuelle (**MT_RC**) à celle de leur acte de référence. La raison d'une telle procédure est que cela n'est pas dans les objectifs de tarifier séparément un acte de soin et son complément d'acte. Mais comment identifier pour chaque complément leur acte de référence ?

Le choix optimal de traitement aurait été d'associer à chacun des codes actes des compléments, le code acte de leur acte de référence. Par exemple, pour l'acte de pharmacie codé PH4, le complément associé est codé HD4 (pour honoraire de dispensation 4). Puis, de sommer les montants selon la classification ainsi réalisée en considérant la date de soin et de règlement et l'identifiant de l'assuré (**REF_PERSONNE**) *a priori* communes entre le complément et son acte de référence. Cela nécessite cependant d'effectuer une classification presque aussi chronophage que celle présentée en section 2 de ce chapitre (classification des codes actes en sous-famille d'actes, familles d'actes et postes de soin).

L'autre éventualité considérée fut d'ignorer ces compléments d'actes. Or, comme ils représentent entre 3 % et 8 % (selon la garantie, selon l'année considérée) des montants de remboursement de la mutuelle, cela ne fut pas envisageable.

La solution retenue fut alors de :

- Créer une clé de sommation (stockant les informations suivantes : l'identifiant de l'adhérent, la date de soin et de règlement, sa qualité, son sexe, son âge, sa région de localisation, la famille de l'acte et le poste de soin associé à l'acte) ;
- Rechercher pour chaque ligne de quantité nulle (complément) les lignes de quantités non nulles ayant une même clé de sommation (ce sont les lignes « candidates » d'actes de référence) ;
- Choisir parmi les lignes candidates celles dont le montant RC est le plus petit (« le plus petit » est le critère choisi pour deux raisons : la plupart des compléments d'acte concernent de la pharmacie et les montants pour les actes de pharmacie sont généralement plus petits que pour les autres actes ; nous avons donc plus de chance de sommer un complément de pharmacie avec un acte de pharmacie ; de plus, s'il y a deux actes de référence quasiment identiques le même jour, le premier complément sera sommé à l'un et le second à l'autre) ;
- Sommer la ligne (**MT_RC**, **MT_RO**, **DEPENSE**) du complément à l'acte de référence ainsi sélectionné ;
- Supprimer ensuite la ligne de complément pour ne pas la double-compter.

Suite à ce traitement, il a été observé, lors de la vérification du code *R* créé, une perte de moins de 1 % des montants de dépense, des montants RO et RC, pour chacune des deux garanties. Deux raisons ont pu être identifiées : soit, aucun acte de référence n'existait pour un complément donné (une erreur éventuelle des bases de données ou de la date de soin), soit, le complément a été réglé un autre jour que celui de son acte de référence. La perte étant négligeable, nous avons procédé avec cette dernière dans la suite des travaux.

2.1.2.8. Retraitement de la variable QUALITE

La variable de qualité de l'assuré prend les valeurs d'assuré souscripteur, d'ayant droit conjoint ou d'ayant droit enfant. Elle a été retraitée de façon à prendre les valeurs utilisées dans l'Open DAMIR pour les mêmes modalités. Ainsi, d'après le lexique de l'Open DAMIR :

- 1 correspond au souscripteur ;
- 2 correspond au conjoint ou assimilé ;
- 3 correspond aux enfants.

Dans le cadre de nos travaux, la tarification élaborée n'admet pas de différence de tarifs entre un souscripteur et son conjoint. Ainsi, la modalité « 2 » sera regroupée dans la modalité « 1 » (souscripteur).

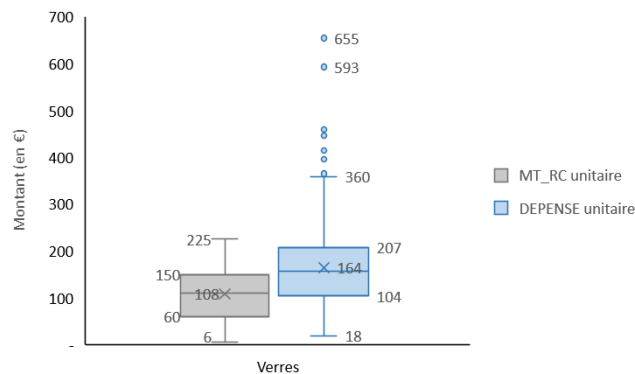
2.1.2.9. Traitement des données manquantes

Il n'y a pas de données manquantes à signaler. À noter qu'il est important de faire cette vérification avant toute agrégation de lignes puisqu'un tel procédé tant à masquer l'absence de données, ce qui peut être source de biais.

2.1.2.10. Traitement des données aberrantes

Le principe de cette étape est de pouvoir exclure les lignes associées à des données aberrantes (communément appelées « *outliers* »). Pour VirtuaMut', ce serait par exemple un remboursement pour une prothèse auditive de l'ordre de 100 000 € (qui serait alors peut-être une erreur de saisie dans la base de gestion). Cette procédure permet de ne pas biaiser faussement la tarification qui sera ensuite réalisée.

Pour cela, nous avons analysé via des boîtes à moustaches, pour chaque sous-famille d'actes, les montants minimum et maximum de remboursement de la mutuelle et des dépenses engagées. Il faut au préalable rendre unitaires les montants associés à chaque ligne afin de les rendre comparables. Les références de comparaison sont les prix moyens du marché par équipement ou acte de soin.



Graphique 1 : Exemple de boîte à moustache pour le cas des verres de la garantie B

Cette analyse fut cependant complexifiée par la nature diverse des actes contenus dans certaines sous-familles d'actes. En effet, par exemple, quand il s'agit des verres optiques, nous avons été initialement étonnés de voir des verres à 600 € l'unité, mais cela s'avère être une possibilité pour les verres progressifs. De même, pour les vaccins, les prix unitaires varient fortement selon le type de maladie.

Au final, nous n'avons pas détecté de valeurs aberrantes. Néanmoins, le cas des actes de transport a retenu notre attention notamment avec le cas où un individu aurait par exemple parcouru environ 1 000 km pour la garantie B. Cela pourrait s'expliquer par un rapatriement ou encore, par un acte de soin effectué en Corse alors que la personne habite à Paris. La sous-famille « Transport » présentant une myriade de possibilités sur les comportements des individus et des doutes sur son dénombrement, elle ne sera par la suite pas considérée dans notre tarification (via la méthode « GLM »).

2.1.2.11. Agrégation des lignes selon l'Open DAMIR

La base Open DAMIR est une base de données agrégées alors que les bases de données de VirtuaMut' présentent des données tête par tête. Dans un contexte de fusion de ces deux bases, il faut nécessairement les harmoniser. Or, désagréger l'Open DAMIR est un sujet complexe dont la faisabilité reste à étudier. Il est plus raisonnable de partir sur l'agrégation des données de VirtuaMut' et de travailler en vision moyenne. Il ne faudra cependant pas oublier qu'agréger, c'est perdre de l'information.

Pour cela, la variable **REF_PERSONNE** (qui assure que les lignes renseignent des données individuelles) est délaissée. La base ainsi obtenue est une base agrégée semblable à l'Open DAMIR où une ligne indique des montants de dépense, de remboursements et des quantités d'actes, pour un type d'acte donné (selon le code acte et le mois de délivrance du soin), pour un agrégat d'individus ayant des caractéristiques communes.

Au final, la base de la garantie A dénombre 18 347 lignes (contre 168 745 au départ) et celle de la garantie B dénombre 62 142 lignes (contre 1 184 746 au départ).

2.1.2.12. Création de la variable EXPO_TOTALE

L'exposition est la durée normalisée pendant laquelle un assuré est exposé aux risques au sein du portefeuille de la mutuelle pendant l'année de soin au titre de la souscription d'un contrat. Pour rappel, un contrat d'assurance santé est très souvent annuel à tacite reconduction. Par conséquent, un assuré couvert pendant l'année entière aura une exposition évaluée à 1. S'il est présent dans le portefeuille du 01/01/2018 au 31/06/2018 soit, la moitié de l'année, son exposition sera de 0,5. Il est donc effectué une sorte de pro rata sur l'année.

Pour créer la variable d'exposition **EXPO_TOTALE** de la base de données de VirtuaMut' qui associe à chaque ligne de prestations agrégées le nombre d'adhérents exposés (non sinistrés ou sinistrés et générant une partie du montant de prestations de la ligne) dont il sera nécessaire pour le calcul de la fréquence lors de la démarche de tarification en Chapitre 3, un travail sur la base de données initiales des adhérents a dû être effectué.

Nous travaillons avec deux années d'observation pour les bases de VirtuaMut' (année de soin 2018 et année de soin 2019). Ainsi, pour chaque adhérent présent dans la base des adhérents, la valeur d'exposition pour une année donnée sera toujours entre 0 et 1. Chaque adhérent présentera alors une variable **EXPO_2018** et une variable **EXPO_2019** pour respectivement son exposition en 2018 et son exposition en 2019. Pour déterminer la valeur prise par ces deux nouvelles variables, il a fallu tout d'abord retraiter les différentes dates de début et de fin de contrat de chaque individu pour éliminer les cas de discontinuité de dates et obtenir une date de début et de fin de contrat unique pour chaque adhérent, sur une seule ligne. Les cas de discontinuité sont directement liés aux 66 adhérents qui apparaissent plusieurs fois dans la base des adhérents de la mutuelle (cf. 2.1.1.2. Bases des adhérents). La méthodologie de retraitement selon les situations rencontrées est explicitée en Annexe 6.5.

Annexe 6.5

Exemple : Selon la base des adhérents, l'adhérent Ray (code de référence 81194) a souscrit :

- À la garantie A du 01/01/2010 au 30/04/2019
- À la garantie A du 01/07/2019 à « » (contrat en cours)

Il y a discontinuité de dates. Pour déterminer la valeur prise par **EXPO_2019** :

- **DATE_DEBUT** = 01/01/2019 car la date de début de son premier contrat (01/01/2010) est moins récente que la date de début de notre année de soin (2019).
- **DATE_FIN** = 31/10/2019 car la date de fin de son dernier contrat n'est pas encore déterminée (le contrat étant encore en cours après le 31/12/2019) mais comme il y a une discontinuité de dates entre les deux contrats, afin de garder une présence de Ray à 10 mois sur 12 dans l'année 2019 sans discontinuité, il y a un décalage de deux mois (**DATE_FIN** ne prend donc pas la valeur du 31/12/2019 mais celle du 31/10/2019, deux mois avant).

Par ailleurs, dans ce cas, **EXPO_2018** vaut 1.

Après ce retraitement, la variable d'exposition **EXPO_i** (avec $i = \{2018, 2019\}$) associée à une ligne d'adhérent est calculée de la manière suivante :

$$EXPO_i = \frac{\text{Nombre de jours présents dans l'année de soin } i}{365} = \frac{DATE_FIN - DATE_DEBUT}{365} \text{ (sous Excel)}$$

Les années 2018 et 2019 ne sont pas des années bissextiles donc la division par 365 comme étant le nombre de jours d'une année civile est justifiée.

Par la suite, dans la base fusionnée contenant les prestations agrégées, il suffira alors d'associer à chaque ligne la somme des expositions ainsi calculées en filtrant sur l'année de soin de la ligne et sur les caractéristiques retenues pour les individus. La variable **EXPO_TOTALE** est ainsi créée.

Une simplification a cependant été faite à ce stade (et cette simplification est aussi faite dans l'Open DAMIR) : si dans la base agrégée similairement à l'Open DAMIR, une combinaison de caractéristiques d'individus est manquante (car ces derniers n'ont eu aucun sinistre durant les deux années de soin), alors, ils seront négligés (ils n'apparaîtront pas dans la base).

2.1.2.13. Dernière agrégation

Ce dernier traitement a été réalisé à la suite d'une erreur constatée dans les résultats lors de l'étape de tarification par GLM. Il est néanmoins présenté ici pour rester cohérent avec le plan du mémoire.

La base laissée telle quelle conduira à des résultats erronés sur la détermination de la fréquence par une méthode de GLM. En effet, à ce stade, la variable **EXPO_TOTALE** associée à chaque ligne correspond en fait à un pool d'individus similaires²⁰ qui auraient pu être concernés par la prestation mensuelle reliée à un code acte d'une ligne spécifique (et seulement certains d'entre eux ont effectivement été touchés par l'acte de soin ou d'équipement). En ce sens, elle détermine bien une exposition.

Cependant, dans le cadre de nos travaux, il est primordial que la valeur de cette variable soit cohérente avec le périmètre des autres variables mais aussi, avec le périmètre de la tarification réalisée. Ce dernier consiste en un tarif **annuel** et **par segment de tarification retenu** (cf. sous-partie 2.2.3.) et non pas par code acte. Le périmètre de la variable **EXPO_TOTALE** consiste en une vision **annuelle** et un nombre **agrégé** de bénéficiaires déterminé sur la base de **trois critères** : la classe d'âge, le sexe, la région.

Pour que les résultats soient corrects, il a été constaté qu'une dernière agrégation était nécessaire pour pouvoir réellement associer l'exposition totale d'une ligne à une quantité d'actes cohérente avec cette dernière (celle qui est, en théorie, réellement associée à cette exposition) et aux bons montants de prestations. Il a ainsi fallu abandonner la donnée des codes actes puis agréger les lignes selon :

- Les segments de tarification retenus ;
- Les années de soins ;
- Les 3-uplets de caractéristiques de bénéficiaires (classe d'âge, sexe, région).

C'est à ce niveau d'agrégation qu'il faudra associer la variable **EXPO_TOTALE** de la manière décrite précédemment. Cela revient aussi à agréger les lignes selon les valeurs d'expositions totales déjà associées (d'une mauvaise manière initialement), les années de soin et les segments de tarification.

Sans cela, la fréquence qui en résulte est largement sous-estimée. Il est à noter en revanche qu'une telle agrégation conduit, pour chaque segment de tarification, à effectuer plus tard un GLM sur une trentaine de lignes. Un tel nombre de lignes est tout à fait normal puisqu'au lieu de dénombrer selon des actes effectués, il y a ici dénombrement selon des 3-uplets restreints de caractéristiques d'adhérents et selon un nombre d'années de soin.

²⁰ Dans le sens de caractéristiques communes, à savoir, la même classe d'âge, le même sexe, la même région.

2.1.2.14. Récapitulatif des variables de la base de données de VirtuaMut'

Les noms des variables ont été adaptés (si besoin) pour se conformer à l'Open DAMIR. Nous retiendrons ainsi les variables pertinentes suivantes pour la suite des travaux :

Nom de la variable	Description
DATE_SOINS	L'année de soin
DATE_RGLT	L'année de règlement
SS_FAM	La sous-famille de l'acte (cf. Section 2 de ce chapitre)
FAM	La famille de l'acte (cf. Section 2 de ce chapitre)
POSTE	Le poste de soin de l'acte (cf. Section 2 de ce chapitre)
QLT	La qualité des adhérents (souscripteur/conjoint ou enfant)
SEXE	Le sexe des adhérents (Femme ou Homme)
CLASSE_AGE	La classe d'âge des adhérents
REGION	La région de localisation des adhérents de la ligne
DEPENSE	Le montant total de dépenses engagées par les individus de la ligne
MT_RO	Le montant RO total remboursé aux individus de la ligne
MT_RC	Le montant RC total remboursé aux individus de la ligne
QTE_ACTE	La quantité d'actes totale
DENOMBREMENT	Pour VirtuaMut', cette variable prend la même valeur que celle de QTE_ACTE
EXPO_TOTALE	L'exposition totale de la ligne (i.e. le nombre de bénéficiaires exposés)

Tableau 7 : Les variables pertinentes de VirtuaMut' pour une garantie donnée

2.1.3. Des généralités sur les bases de données de l'Open DAMIR

Pour rappel (p.28) :

En libre accès depuis le 26 janvier 2015 et gérée par la CNAM, l'Open DAMIR contient l'ensemble des prestations prises en charge par l'Assurance Maladie obligatoire (à l'exception d'une grande majorité des prestations hospitalières du secteur public). Pour des raisons d'anonymat, les données sont proposées sous forme agrégées : une ligne d'un fichier est donc une somme des actes et des montants associés à un ensemble d'individus ayant une quarantaine de caractéristiques communes.

2.1.3.1. Le choix de l'historique de données

Comme mentionnée en partie 1.2.3, la profondeur d'historique de la base Open DAMIR au moment de l'écriture de ce mémoire est de 10 ans (2009 à 2019). La base s'actualise tous les ans aux environs du mois de juin. Les données de 2019 ont donc été rendues disponibles en juin 2020. Pour ces travaux, il a été retenu les fichiers de données de 2018 et 2019, soit 2 ans d'historique ou encore 24 fichiers Excel au format .csv puisque la base Open DAMIR est restituée sous la forme de fichiers mensuels.

La première raison d'un tel choix, raison qui semble être la plus évidente, est que les données utilisées provenant de la mutuelle couvrent un historique d'années jugées similaires (en prenant l'hypothèse exposée en partie 2.1.1 énonçant que le passé est équivalent au futur pour les prestations de la mutuelle). Comme une fusion de bases est envisagée afin d'enrichir virtuellement le portefeuille de VirtuaMut', une certaine cohérence est nécessaire.

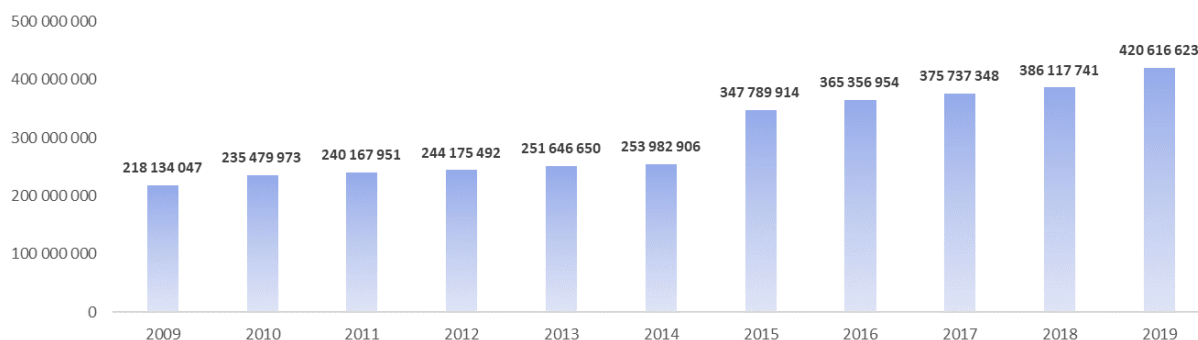
Cependant, une question pourrait amenée à subsister : pourquoi deux années d'historique ? En fait, d'une part, prendre une profondeur d'historique trop élevée n'est pas nécessairement une bonne chose et pourrait amener à des résultats biaisés par le passé car, comme nous l'avons vu antérieurement, les lois qui régissent le monde de l'assurance santé évoluent sans cesse et ce qui est normal un jour peut

être amené à différer le lendemain. Ainsi, plus concrètement, à titre d'exemple illustratif, depuis le 1^{er} mai 2017, le tarif d'une consultation de médecin généraliste de secteur 1 s'élève à 25 € contre 23 € auparavant (c'est aussi le montant de la BR). Si, pour évaluer les coûts d'un tel acte, les années de 2015 à 2019 avaient été prises, le coût moyen pour la mutuelle pour cet acte précis aurait été estimé à une valeur entre 23 et 25 multipliée par le taux de remboursement de la mutuelle (qui lui, est ici constant pour ces 5 années). Cependant, la valeur qui sera valable pour le tarif en 2021 serait 25 multipliée par le taux de remboursement de la mutuelle. Considérer le coût d'une visite chez le médecin à 25 € est ce qui est le plus conforme à la réalité. Prendre un historique de 2018 à 2019 limiterait ce biais.

D'autre part, la branche de la maladie (frais de soins) est à développement court : la plupart des prestations sont remboursées en 2 voire 3 ans (la plus grosse partie dans l'année même ou l'année suivante, éventuellement un résiduel la troisième année). Les règlements, que ce soit de la Sécurité Sociale ou de la mutuelle, sont rapides. De ce fait, 2 ans d'historique est suffisant.

2.1.3.2. Un nombre de lignes très conséquent

En ce qui concerne le décompte de lignes dans la base de l'Open DAMIR par année de règlement, nous observons une augmentation du nombre de lignes au fil du temps. Cela pourrait s'expliquer par l'évolution des nomenclatures (comme mentionnés en partie 1.1.4.) et donc, l'ajout de nouveaux codes actes, mais aussi par l'affinement et les différentes améliorations apportées à la base. Par exemple, une augmentation soudaine du nombre de lignes dans la base est enregistrée en 2015 par rapport à 2014 puisque la variable renseignant la région de localisation des adhérents a été segmentée en 13 zones au lieu de 9 (conformément à l'évolution réglementaire de la réforme territoriale de 2014). Il est donc à noter que cette *Open data* est aussi amenée à évoluer selon les nouvelles mesures réglementaires.



Graphique 2 : Dénombrement de lignes de l'Open DAMIR

Il est à préciser que nous avons repris le décompte d'Arnold MEKONTSO (2018, *L'open DAMIR : apport à la maîtrise des dépenses de santé*, p.52) pour les années 2009 à 2016, puis avons complété avec les années 2018 et 2019 issues de nos travaux. L'année 2017 a été déterminée par interpolation (moyenne entre le nombre de lignes en 2016 et 2018).

2.1.3.3. Les variables de l'Open DAMIR

Un fichier mensuel de la base Open DAMIR présente 55 variables couvrant 6 axes²¹ différents (cf. schéma 4, sous-partie 1.2.3). Il sera possible de trouver en Annexe 7 l'intégralité des variables de l'Open DAMIR. Nous ne nous contenterons de mentionner ci-après seulement que les variables principales ou celles dont une remarque d'intérêt pourrait être énoncée. Il s'agit ici de se familiariser progressivement avec ces variables qui seront par la suite utilisées dans les travaux.

Annexe 7

²¹ Nous retenons ici notre propre découpage. Le découpage proposé par le lexique de l'Open DAMIR diffère légèrement.

Dans l'axe d'analyse des prestations, il est possible de trouver les variables suivantes :

- **PRS_NAT** permet de s'enquérir sur la nature de la prestation. C'est la variable directement en lien avec le code acte du soin. Elle sera la variable centrale dans la classification des actes qui sera effectuée en section suivante.
- **PRS_REM_TAU** indique le taux de remboursement de l'acte. La valeur que prend cette variable doit être vue comme étant un pourcentage de la base de remboursement de l'acte. Par exemple, « 90 » signifie « 90 % de la BR ». Il ne faudra donc pas oublier de la multiplier par 0,01 afin de l'utiliser correctement.
- **PRS_REM_TYP** permet de déterminer le type de remboursement de l'acte, i.e. si ce dernier est un acte de référence, un complément d'acte, un ticket modérateur, ...

Dans l'axe d'analyse des bénéficiaires, il est possible de trouver les variables suivantes :

- **BEN_SEX_COD** désigne le sexe du bénéficiaire.
- **AGE_BEN_SNDS** désigne la tranche d'âge du bénéficiaire au moment de la délivrance des soins (calcul exact d'après le gestionnaire de la base de données).
- **BEN_QLT_COD** désigne la qualité du bénéficiaire (souscripteur ou types d'ayant droit).
- **BEN_RES_REG** désigne la région de résidence du bénéficiaire. Cette variable se découpe en 13 zones géographiques (proches des grandes régions administratives créées par la réforme territoriale de 2014).
- **BEN_CMU_TOP** permet d'identifier les bénéficiaires du CMU-C.

Les professionnels de santé sont les exécutants des soins qui délivrent l'acte (pour une opération de chirurgie, le chirurgien par exemple) mais aussi, les prescripteurs qui prescrivent l'acte sans pour autant les réaliser (le médecin généraliste qui redirige le patient vers le chirurgien). Pour un soin donné, l'exécutant peut *a priori* aussi être le prescripteur.

Dans l'axe d'analyse des indicateurs temporels, il est possible de trouver les variables suivantes :

- **FLX_ANN_MOI** désigne l'année et le mois de règlement de l'acte sous le format « yyyyymm ».
- **SOI_ANN** et **SOI_MOI** désignent respectivement l'année et le mois de délivrance du soin sous le format « yyyy » et « mm ».

Dans l'axe d'analyse des indicateurs quantitatifs, il est possible de trouver les variables suivantes :

- **PRS_ACT_COG** désigne le coefficient global de l'acte, c'est-à-dire, le coefficient qui intervient dans la nomenclature des actes NGAP et qui est multiplié à la base de remboursement unitaire désigné par des lettres afin de déterminer la base de remboursement de l'acte (cf. partie 1.1.4.). C'est cette variable qui correspond à la quantité d'actes initialement renseignée dans les données de VirtuaMut'.
- **PRS_ACT_NBR** désigne le dénombrement de l'acte.
- **PRS_ACT_QTE** désigne la quantité d'actes.
- **PRS_DEP_MNT** désigne le montant de dépassement de l'acte.
- **PRS_PAI_MNT** désigne le montant de dépenses engagées de l'acte.
- **PRS_REM_MNT** désigne le montant de remboursement de la Sécurité Sociale de l'acte.
- **FLT_ACT_COG**, **FLT_ACT_NBR**, **FLT_ACT_QTE**, **FLT_DEP_MNT**, **FLT_PAI_MNT**, **FLT_REM_MNT** sont les équivalents « préfiltrés » des variables précédentes dans le même ordre de présentation.
- **PRS_REM_BSE** renseigne la base de remboursement de l'acte.

Arnold MEKONTSO dans son mémoire *L'open DAMIR : apport à la maîtrise des dépenses de santé* (2018, page 51 et 52) nous explique que les indicateurs quantitatifs préfiltrés au sens de l'Open DAMIR ont pour vocation de permettre uniquement l'étude du régime obligatoire de la Sécurité Sociale alors

que les indicateurs non préfiltrés ont plutôt pour vocation de permettre l'étude des prestations liées à la CMU-C, les compléments Alsace-Moselle (régime local), les prestations de prévention, etc. Ainsi, leur distinction réside dans les valeurs prises selon le type de remboursement (variable `PRS_REM_TYP`) : les indicateurs préfiltrés sont nuls quand l'acte n'est pas remboursé par le régime obligatoire.

En ce qui concerne la différence²² entre le dénombrement de l'acte et sa quantité, les deux variables représentent en substance exactement la même chose et permettent de décompter le nombre d'actes par lignes agrégées. Cependant, bien qu'elles aient pour la plupart du temps la même valeur, elles diffèrent dans le cas de certains actes :

- Les transports où la quantité renseigne sur le nombre de factures alors que le dénombrement compte le nombre de courses effectuées.
- Les indemnités kilométriques où la quantité compte le nombre de kilomètres facturés.
- Les indemnités journalières où la quantité compte le nombre de jours indemnisés.
- Les frais de séjour où la quantité compte le nombre de jours hospitalisés.

Il faudra donc parfois utiliser la variable de quantité ou de dénombrement selon l'acte étudié. Par ailleurs, un certain nombre de régimes d'Assurance maladie ne renseignent pas le dénombrement des actes mais seulement la quantité, ce qui a pour conséquence la présence de nombreuses valeurs manquantes pour la première variable. Il faudra penser à les gérer dans la suite de l'étude.

2.1.3.4. Des limites de l'Open DAMIR : quid de la fréquence ?

Comme énoncé à plusieurs reprises auparavant, l'Open DAMIR est une base de données agrégées et cette agrégation, qui permet d'assurer l'anonymat à la fois des patients qui la constituent et des différents professionnels de santé, sera source d'un certain nombre de problématiques dans ce mémoire. La principale étant la détermination de la fréquence et de sa loi en présence de données agrégées.

En effet, bien que l'Open DAMIR fournisse les quantités par acte, elle ne renseigne pas le nombre de bénéficiaires par ligne. Autrement dit, nous savons qu'une ligne est associée à un groupe d'individus mais nous ne savons pas exactement le nombre d'individus dans ce groupe. Or, le nombre de bénéficiaires (exposés) est une donnée nécessaire dans le calcul de la fréquence par l'approche fréquence x coût moyen (approche priorisée de tarification de ces travaux) puisque la fréquence est déterminée comme étant le quotient entre une quantité d'actes et un nombre de bénéficiaires exposés (la formule de calcul sera présentée plus en détail dans le Chapitre 3).

Une solution potentielle qui sera utilisée serait alors de coupler l'Open DAMIR avec une base de données de l'INSEE portant sur la démographie française afin de pouvoir estimer un nombre de bénéficiaires exposés par ligne. Cette solution a été mentionnée par Pascale QUENNELLE et Marc RAYMOND de l'Institut des Actuaire [12]. La base de données retenue est celle intitulée « Estimation de population par département, sexe et âge quinquennal - Années 1975 à 2020 »²³. La méthodologie utilisée par l'INSEE pour estimer le nombre d'habitants par région, par sexe et par âge (avec un pas de 5 ans) est explicitée dans la section « Documentation » du site d'extraction de la base²⁴.

En ce qui concerne un tel choix, Arnold MEKONTSO dans son mémoire *L'open DAMIR : apport à la maîtrise des dépenses de santé* (2018, page 78 et 79) a étudié la représentativité des régions de résidence des bénéficiaires de prestations d'assurance maladie donnée par la base de l'Open DAMIR (sur la base d'occurrence de prestations) comparée à la répartition régionale de la population française. Il en a déduit

²² D'après la réponse apportée par Evelyne TOUSTOU, une des gestionnaires de la base Open DAMIR sur le site data.gouv.fr, à une question datant du 25 janvier 2019.

²³ Source : INSEE - Estimations de population. Données actualisées au 14 janvier 2020.

²⁴ <https://www.insee.fr/fr/statistiques/1893198#documentation>

une cohérence (plus précisément, une dépendance positive mesurée par un τ de Kendall) signifiant une évolution similaire (et dans le même sens) des deux notions. Par ailleurs, étant donné que plus de 93 % de la population française est couverte par la Sécurité Sociale, nous avons jugé une telle estimation convenable.

De plus, une autre limite provient de la corrélation existante entre le coût et la fréquence en santé (chose que nous n'avons pas forcément en tarification auto). Pour rappel, l'un des objectifs opérationnels est de réaliser une tarification via notamment une approche GLM fréquence x coût moyen. Mais une telle approche suppose la non-corrélation entre la fréquence et le coût. Cette non-corrélation en santé est uniquement visible niveau de garantie par niveau de garantie puisqu'un niveau de garantie élevé peut amener à une fréquence de consommation plus élevée (et donc à un risque d'anti-sélection). Une fréquence plus élevée signifie plus de remboursement de la part de la mutuelle. Cela n'est pas observé de manière aussi prononcée en automobile parce que les sinistres ne sont en quelque sorte pas « choisis » (une personne, sauf cas de suicide, ne choisit pas d'avoir un accident de voiture ; il ne choisit pas non plus quand une grêle tombe sur son véhicule) alors qu'en santé, un assuré ira plus facilement acheter des lunettes s'il est bien remboursé ou renoncera plus facilement à une audioprothèse s'il est mal remboursé (il est par ailleurs fréquent de penser que les femmes ont tendance à être plus prudentes que les hommes en matière de santé et vont donc plus souvent chez le médecin, elles présentent ainsi des fréquences plus élevées mais des coûts unitaires moins importants que ces derniers). Autrement dit, cette démarche n'est valable que sous hypothèse de se placer dans un niveau de garantie particulier. Or, l'Open DAMIR étant une base agrégée, une ligne de prestations présente un mélange de personnes possédant des garanties différentes puisqu'étant affiliés à des assureurs différents.

Pour rendre cependant cette démarche viable, nous nous placerons alors sous la garantie A ou B de VirtuaMut' en déduisant pour chaque ligne de l'Open DAMIR un montant RC et en supposant que l'ensemble des bénéficiaires de la ligne ont souscrit à la garantie A ou B de VirtuaMut'. Cela permet ainsi de placer les travaux sous le prisme d'une garantie particulière et de permettre l'approche choisie.

2.1.4. Traitements préliminaires des données de l'Open DAMIR

Tous les traitements de la base Open DAMIR ont été réalisés sous SAS (ou R) et dans un premier temps, fichier par fichier à cause de la lourdeur de ces derniers et des restrictions logistiques.

2.1.4.1. Étape 1 : Suppression des variables non pertinentes

La première étape de traitement consiste à supprimer toutes les variables de l'Open DAMIR dont il était déjà sûr de ne pas avoir à recourir. Cette suppression s'est faite sur la base de deux considérations :

- La pertinence de la variable dans une optique de tarification (un des objectifs du mémoire) ;
- Le périmètre de l'étude (pour rappel, seul le régime général de la Sécurité Sociale et sa branche maladie – frais de soins sont d'intérêt) ;

Ainsi, toutes les variables concernant les prescripteurs ont été écartées car jugées non pertinentes à l'étude. Certaines variables caractérisant l'exécutant (par exemple, sa région de localisation) ont été gardées par précaution. Elles seront supprimées à une étape ultérieure²⁵. Seuls les indicateurs quantitatifs préfiltrés ont été gardés puisqu'il s'agit ici d'étudier des données concernant le régime général de la Sécurité Sociale uniquement.

Il sera possible de trouver en Annexe 8 les variables supprimées et gardées suite à cette étape.

Annexe 8

²⁵ Sur SAS, chaque étape génère de nouveaux fichiers. Supprimer les variables de l'exécutant à une étape ultérieure permet de garder la trace de ces variables pour visée de sauvegarde.

Il reste à l'issue de cette étape 25 variables, soit un peu moins de la moitié des variables initialement présentes, dont certaines qui ont été gardées en tant que filtres pour les étapes suivantes.

2.1.4.2. Étape 2 : Utilisation des variables de filtrage puis suppressions

Les variables gardées en tant que filtres sont utilisées puis supprimées. Celles concernées sont :

- **ASU_NAT** (nature de prestation) dont seules les lignes présentant la modalité « 10 » pour la branche maladie ne sont gardées ;
- **CPT_ENV_TYP** (type d'enveloppe) dont seules les lignes présentant les modalités « 1 », « 2 », « 3 », « 9 » et « 98 »²⁶ pour en partie filtrer sur le régime général ne sont gardées ;
- **BEN_CMU_TOP** (bénéficiaire CMU-C) dont seules les lignes correspondant à des non bénéficiaires de la CMU-C n'ont été gardées ;
- **TOP_PS5_TRG** (top Périmètre hors CMU C et prestations pour information) permet de supprimer quelques prestations non pertinentes comme celles liées à la CMU-C.

Il reste désormais 21 variables en jeu.

2.1.4.3. Étape 3 : Suppression des variables non pertinentes vis-à-vis des jeux de données de la mutuelle

Afin de mettre en harmonie les jeux de données issus de l'Open DAMIR et ceux issus de VirtuaMut', toutes les variables de l'Open DAMIR qui ne sont pas présentes dans les bases de données de l'organisme assurantiel sont écartées, à condition qu'elles ne servent pas à construire une variable pertinente. Par exemple, **SOI_ANN** et **SOI_MOI** sont fusionnées afin d'apparaître sous le format « yyyyymm » en une variable unique **DATE_SOIN**.

2.1.4.4. Étape 4 : Quelques vérifications traditionnelles

Il ne s'agit plus ici d'exclusivement chercher à supprimer des variables mais à s'intéresser aux données et à effectuer des vérifications traditionnelles (similaires que pour les données de la mutuelle) :

- La quantité des compléments, majorations et suppléments d'acte doivent être à 0. Ils sont repérés grâce à la variable **PRS_REM_TYP** (type de remboursement) où seules les lignes présentant les valeurs « 0 » pour l'acte de référence ou « 99 » pour valeur inconnue sont gardées. **PRS_REM_TYP** est ensuite supprimée. Les valeurs inconnues étant nombreuses pour cette variable, nous avons dû associer à chaque code acte (**PRS_NAT**), une indicatrice prenant la valeur de « 1 » si l'acte est un complément, supplément ou majoration d'acte et « 0 » si c'est un acte de référence. Cette indicatrice a ensuite permis de mettre à 0 les quantités d'actes visées.
- Les lignes avec des montants ou quantités négatifs sont supprimées puisque cela est sûrement lié à des opérations de régularisation de la Sécurité Sociale.
- Les lignes dont le taux de remboursement de la Sécurité Sociale est supérieur à 100 % sont supprimées puisque cela est, *a priori*, impossible.
- La variable **FLT_REM_MNT** est renommée **MT_RO** (pour montant RO). Il a été observé que moins de 1 % des lignes présentaient un écart de plus de 1 € entre le montant RO renseigné et celui recalculé à partir du taux de remboursement **PRS_REM_TAU** et de la base de remboursement **PRS_REM_BSE**. Excepté les cas où cela est dû à une spécificité de renseignement des compléments d'acte, les montants ont été corrigés pour retenir la valeur calculée. La variable du taux de remboursement **PRS_REM_TAU** est ensuite supprimée car elle

²⁶ « 1 », « 2 », « 3 », « 9 » et « 98 » représentent respectivement les soins de villes, l'hospitalisation et les consommations intermédiaires, les prestations légales de l'assurance maladie, les valeurs inconnues et les sans objet.

peut être déduite du reste. À savoir que la variable **PRS_REM_BSE** est bien une somme agrégée de bases de remboursement.

- Les variables sont aussi renommées pour se conformer aux traitements de VirtuaMut’.

Variable renommée (variable de base)	Libellé
DATE_RGLT (<i>FLX_ANN_MOI</i>)	Année et mois de règlement
CODE_ACTE (<i>PRS_NAT</i>)	Nature de prestation
DATE_SOIN	Fusion de SOI_ANN et SOI_MOI
SEXE (<i>BEN_SEX_COD</i>)	Sexe des bénéficiaires
CLASSE_AGE (<i>AGE_BEN_SNDS</i>)	Tranche d’âge des bénéficiaires au moment des soins
QLT (<i>BEN_QLT_COD</i>)	Qualité des bénéficiaires
REGION (<i>BEN_RES_REG</i>)	Région de résidence des bénéficiaires
DENOMBREMENT (<i>FLT_ACT_NBR</i>)	Dénombrement de la prestation préfiltré
QTE_ACTE (<i>FLT_ACT_QTE</i>)	Quantité de la prestation préfiltrée
DEPENSE (<i>FLT_PAI_MNT</i>)	Montant de la dépense de la prestation préfiltrée
MT_RO (<i>FLT_REM_MNT</i>)	Montant versé/remboursé préfiltré
PRS_REM_BSE	Base de remboursement

Tableau 8 : Liste des variables restantes à l’issue de l’étape 4

2.1.4.5. Étape 5 : Traitement des données manquantes et aberrantes

Cette étape consiste à supprimer les lignes présentant des données manquantes ou aberrantes. Les données manquantes dans l’Open DAMIR ne sont pas forcément des variables NA (vide), cela peut aussi être des modalités prises par des variables dont la signification est « INCONNU » ou « VALEUR INCONNUE ». De ce fait, il ne suffira pas sous SAS de repérer le nombre de données manquantes indiquées par « NA » mais de regarder aussi toutes les modalités et leur signification en détail.

Les traitements suivants sont alors effectués :

- Si les adhérents de la ligne présentent une qualité de bénéficiaires égale à « 2 » (conjoint et assimilé) ou « 4 » (autre ayant-droit) avec un âge supérieur à 19 ans, la qualité est remplacée par « 1 » (assuré) ; si la qualité est égale à « 4 » (autre ayant-droit) avec un âge inférieur à 19 ans, la qualité est remplacée par « 3 » (enfant)²⁷ ; si la qualité est égale à « 9 » (valeur inconnue), la ligne est supprimée ;
- Pour les lignes présentant un âge, un genre, un code acte ou une région inconnue, elles sont aussi supprimées car considérées comme données manquantes ;
- Pour les valeurs de dénombrement manquantes, est prise la valeur de la quantité d’actes.

Toujours dans son mémoire *L’open DAMIR : apport à la maîtrise des dépenses de santé* (Chapitre 6), Arnold MEKONTSO indique que la méthode que nous avons décidé d’utiliser est communément appelée « analyse des cas complets ». Il propose cependant des alternatives de traitements de données manquantes par des méthodes plus complexes (méthode *Last Observation Carried Forward*, d’imputation par la moyenne ou par la médiane, méthode *local regression*, méthode kNN et méthode *MissForest*) et effectue une comparaison de ces méthodes sur un échantillon de l’Open DAMIR présentant uniquement des lignes complètes (le principe est de simuler des données manquantes sur des données complètes afin de connaître au préalable les valeurs manquantes et de pouvoir tester l’efficacité des méthodes). Il retient finalement la méthode *MissForest*. Il a été décidé de ne pas s’aventurer dans ces méthodes car dans le cadre des travaux, le problème des données manquantes est négligeable.

²⁷ Les personnes d’âge inférieur à 19 ans et souscripteurs ont donc été négligées. Par ailleurs, la qualité d’une personne dans l’Open DAMIR concerne la Sécurité Sociale et non un organisme assureur. Il sera cependant supposé qu’un ayant droit sous la Sécurité Sociale le sera aussi pour sa complémentaire santé.

En ce qui concerne les données aberrantes, il n'a pas été relevé de données atypiques ou extrêmes en ce qui concerne les variables d'intérêt pour une optique de tarification.

2.1.4.6. Étape 6 : Ajouter le nombre de bénéficiaires par ligne

Conformément à ce qui a été exposé en partie 2.1.3.4., le nombre de bénéficiaires pour chaque ligne agrégée de l'Open DAMIR a été déduit de la base de l'INSEE portant sur la démographie française intitulée « Estimation de population par département, sexe et âge quinquennal - Années 1975 à 2020 »²⁸. À partir de cette base de données externe (qui, soi-dit en passant, est elle aussi une *open data*), un tableau de répartition de la population française a été créé (une pour 2018 et une pour 2019 pour plus de précisions dans l'affectation) et a été rendu cohérent avec le format de données de l'Open DAMIR.

La méthode de fusion à l'Open DAMIR consiste à associer, selon les caractéristiques des adhérents de la ligne (classe d'âge, sexe et région de localisation), son effectif de population française par ligne. La variable ainsi créée est nommée **EXPO_TOTALE** conformément aux bases de VirtuaMut'.

Cela sera explicité un peu plus tard dans cet écrit mais il faut bien retenir pour le calcul de la fréquence (dans un cadre de tarification) les bénéficiaires exposés aux risques (qu'ils aient subi ou non un soin). Il est donc ici supposé une hypothèse d'exposition égale à 1 pour chaque personne.

2.1.4.7. Étape 7 : Affecter la classification des actes et un montant RC

Il s'agit ici d'affecter à chaque ligne agrégée de l'Open DAMIR, un montant de remboursement de la mutuelle pour la garantie A et B. Deux variables sont ainsi créées : **MT_RC_A** et **MT_RC_B**.

Pour cela, il a tout d'abord été nécessaire d'affecter à chaque ligne la classification retenue telle que décrite en section suivante. C'est la même classification qui a été appliquée aux données de VirtuaMut'. Les variables de sous-familles d'actes (**SS_FAM**), de familles d'actes (**FAM**) et de grands postes de soin (**POSTE**) ont donc été une nouvelle fois créées. Nous restons ainsi sur un périmètre identique.

Cela a permis ensuite d'affecter un montant de remboursement de la complémentaire à chaque ligne sur la base des prestations proposées pour chaque garantie (cf. les grilles de garanties présentes en Annexe 5) et sur la base des montants de BR renseignée sur chaque ligne. La formule appliquée est comme suit :

$$MT_RC_k = \min(DEPENSE - MT_RO, BR * TX_MUT_k + FORFAIT_k * QTE_ACTE, \\ LIMIT_k * QTE_ACTE - MT_RO)$$

Où :

- $k \in \{A, B\}$ (garantie A ou garantie B) ;
- **DEPENSE** est le montant agrégé de frais réel du soin ou de l'équipement d'une ligne ;
- **MT_RO** est le montant agrégé de remboursement de la Sécurité Sociale d'une ligne ;
- **BR** est la base de remboursement agrégée d'une ligne ;
- **QTE_ACTE** est la quantité d'actes totale d'une ligne. Pour certains types d'actes, cette variable est à être remplacée par la variable **DENOMBREMENT** ;
- **TX_MUT_k** est le taux de remboursement de la mutuelle d'après les grilles de garantie k ;
- **FORFAIT_k** est le forfait offert par la mutuelle (car pour rappel, les prestations sont exprimées à la fois en % de la BR mais aussi en montant forfaitaire) ;
- **LIMIT_k** est la limite de remboursement de la mutuelle pour l'acte concerné.

Le premier terme du minimum indique que le montant RC ne peut pas dépasser la part restante après

²⁸ Source : INSEE - Estimations de population. Données actualisées au 14 janvier 2020.

prise en charge de la Sécurité Sociale. Le deuxième terme est égal au remboursement de la mutuelle selon ses grilles de garanties, sans tenir compte d'une quelconque limite de remboursement. Le dernier terme indique que la mutuelle ne peut pas rembourser plus que la limite qu'elle impose (s'il y en a une).

Trois simplifications ont dû être faites :

- 1) Le montant RC pour les actes de pharmacie est égal à la part restante après prise en charge de la Sécurité Sociale (comme cela est souvent le cas dans la réalité). Cette « simplification » s'explique par la difficulté de prise en compte des différents cas de remboursement liés à la pharmacie (35 %, 70 % ou 85 %) et se justifie par le fait que les médicaments sont pour la plupart du temps très bien remboursés avec un reste à charge nul pour les assurés.
- 2) Par la difficulté de classification des actes hospitaliers dans l'Open DAMIR qui a été rencontrée, les taux de remboursement pour les frais de séjour ont été appliqués à tous les actes hospitaliers dont la classification présentait des parts de doute.
- 3) Les verres ont été reclassés en « très complexes », « complexes », « simples » sur la base de leur code acte.

Par ailleurs, de la même manière que pour les données de VirtuaMut', les quantités nulles, correspondant à des suppléments, majorations et compléments d'actes, ont été fusionnées à leur acte de référence.

2.1.4.8. Étape 8 : Fusion des 24 fichiers

Tous les traitements auparavant exposés ont été effectués sur des fichiers mensuels afin de limiter le temps de traitement des algorithmes. Une agrégation (en lignes) a été réalisée fichier par fichier et ensuite, une fusion des 24 fichiers mensuels a été effectuée.

Pour les mêmes raisons et de la même manière que décrites en partie 2.1.2.13., une dernière agrégation a ensuite été effectuée pour associer la bonne exposition totale à chaque ligne.

Section 2 - La segmentation retenue

2.2.1. Segmentation des prestations présentes dans l'Open DAMIR

La variable PRS_NAT de la base Open DAMIR est associée aux codes actes utilisés par l'Assurance Maladie Obligatoire afin d'identifier les différentes prestations.

PRS_NAT	Libellé Nature de Prestation	Libellé Nature de Prestation en B
0	SANS OBJET	
1096	TELECONSULTATION MEDECIN TRAITANT AVEC EHPAD	TTE
1097	TELE EXPERTISE DOSSIER TRAITANT	TDT
1098	CONSULTATION CCMU 3	U03
1099	CONSULTATION CCMU 4 ET 5	U45
1100	PROTOCOLE MURAIN - BILAN VISUEL	RNM
1101	AVIS PONCTUEL DE CONSULTANT PUPH	APU
1102	AVIS PONCTUEL DE CONSULTANT PSYCHIATRE	APY
1103	AVIS PONCTUEL DE CONSULTANT DU MEDECIN	APC

Tableau 9 : Un extrait du lexique de l'Open DAMIR (onglet PRS_NAT)

Cette variable se présente sous la forme d'une valeur numérique quantitative qu'il faudrait plutôt considérer comme étant une valeur chiffrée qualitative (i.e. voir la variable comme du texte). A chaque valeur de la variable est associée dans le lexique de l'Open DAMIR un libellé renseignant la nature de la prestation et son code acte associé. Les codes actes respectent les nomenclatures normalisées et généralisées, utilisées par la Caisse Nationale de l'Assurance Maladie (CNAMTS) et telles que présentées en partie 1.1.4. de ce mémoire. Un document [9] intitulé *Nomenclature des codifications* publié par le Système National des Données de Santé (SNDS) recense par ailleurs l'ensemble de ces codes. Dans la suite de l'écrit, par abus de langage, les valeurs prises par PRS_NAT seront confondues aux codes actes (par association bijective au lexique de l'Open DAMIR). La variable a par ailleurs été renommée comme tels.

Dans le cadre d'une tarification, il a fallu classer l'ensemble des 1 080 codes actes présents dans l'Open DAMIR dans des familles plus générales d'actes et conséquemment, associer chaque code acte à une garantie (ou risque) particulière. En effet, à chaque sous-famille d'acte ou garantie est associé un tarif propre (ce sont en fait les lignes qu'il est possible de visualiser dans une grille de garanties).

Plusieurs segmentations et donc classifications ont été réalisées :

- Les sous-familles d'actes correspondent à des garanties dans un contrat de santé basique. C'est la segmentation la plus fine utilisée et c'est elle qui sera priorisée dans la mesure du possible ;
- Les familles d'acte regroupent des sous-familles d'actes et sont souvent plus connues par les assurés. Elles seront utilisées dans certains cas où, par exemple, les informations contenues dans les bases de données de la mutuelle VirtuaMut' ne sont pas suffisamment précises pour pouvoir réaliser une tarification sur une sous-famille d'acte ;
- Les grands postes de santé permettent de faire des analyses à une échelle plus macroscopique.

Les particularités des grilles tarifaires liées à la réforme 100 % santé ne sont ici pas abordées.

L'objectif de cette segmentation fut d'être suffisamment générique pour pouvoir être adaptée à toutes les mutuelles puisque, bien que ce présent mémoire soit appliqué à la mutuelle VirtuaMut', il se voulait aussi d'être suffisamment général pour être exploitable par d'autres organismes. Aussi, cette segmentation s'est voulue être assez détaillée et précise.

Grands postes	Familles d'actes	Sous-familles d'actes
Soins courants	Honoraires médicaux	Consultations/visites Actes médicaux (techniques) Autres honoraires médicaux

	Honoraires paramédicaux	Auxiliaires médicaux (kinésithérapeutes, infirmiers, ...)
	Analyse et examen de laboratoire	Analyses médicales et examens laboratoires
	Imagerie médicale	Actes d'imagerie, de radiologie et ostéodensitométrie
	Transport, ambulance	Transport
	Pharmacie	Petit appareillage Grand appareillage Pharmacie Vaccins anti-grippes
Hospitalisation	Hospitalisation	Frais de séjour Honoraires et actes chirurgicaux Forfait journalier Chambre particulière Lit accompagnant Autres - hospitalisation
Dentaire	Dentaire	Prothèse Soins dentaires Parodontologie Implantologie Orthodontie Autres - dentaire
Optique	Optique	Monture Verres Chirurgie œil Lentilles Autres - Optique
Aides auditives	Audio	Audioprothèse Pile, accessoire Autres - audio
Autres	Cure thermique	Cure thermique
	Médecine douce	Ostéopathie, diététique, chiropractie, ...
	Prévention	Prévention
	Prestations supplémentaires	Maternité Autres (aides, inclassable, ...)

Tableau 10 : Segmentation principale retenue des actes

Il est possible de trouver en Annexe 10 la classification de l'ensemble des codes actes de la base Open DAMIR selon la segmentation ci-dessus.

Annexe 10

En ce qui concerne la démarche de cette classification, elle s'est effectuée selon plusieurs considérations :

- Le Q&A sur la page du site data.gouv.fr où la base Open DAMIR est publiée et où des internautes effectuent parfois (rarement) des demandes très ciblées de classification à l'organisme en charge de la base de données et de sa publication ;

- Selon comment sont classifiés les actes chez la mutuelle VirtuaMut' afin de faciliter dès le début le rapprochement des deux bases de données (par exemple, l'homéopathie est classifiée dans la médecine douce alors qu'il est généralement courant de l'associer à de la pharmacie) ;

- Selon la base Open DAMIR elle-même en suivant la démarche suivante :

- Filtrer la variable **ASU_NAT** (nature d'assurance) sur la valeur « 10 » afin d'isoler les prestations uniquement liées à la branche Maladie de la Sécurité Sociale (la seule d'intérêt). Tous les autres codes actes des autres branches ont alors été rangés dans « Autres (aides, inclassables, ...) ».
- À savoir que deux simplifications ont été réalisées :
 - ✚ **Simplification 1** : quand pour un code acte en particulier, la branche de la Sécurité Sociale fluctuait entre Maladie (**ASU_NAT** = 10) ou une autre branche selon les lignes, le choix d'appartenance à une autre branche a été retenu si plus de 80 % des lignes concernées présentaient une valeur de variable **ASU_NAT** différente de 10. Un seuil à 50 % a aussi été pris et le départage s'est alors fait au cas par cas : classer ou ne pas classer un code acte comme affilié à la branche maladie selon son libellé.
 - ✚ **Simplification 2** : n'ont été considérés pour la démarche décrite dans le précédent paragraphe que les mois de janvier 2018 et 2019 (choix arbitraire).

- Selon le fichier des *Données nationales : dépenses d'assurance maladie (hors prestations hospitalières) du régime général* (en considérant arbitrairement, les mois de janvier, février, mars 2019) par l'extraction des variables *l_serie*, *prs_nat* et *l_prs_nat*. Ils correspondent respectivement à la série de la statistique mensuelle à laquelle la prestation est rattachée, à sa codification chiffrée qui est en correspondance directe avec la variable **PRS_NAT** de l'Open DAMIR (même signification des données) et à son libellé ressemblant à ce qui est visible dans le lexique de l'Open DAMIR sans toutefois être identique. C'est la variable *l_serie* qui est d'intérêt car elle propose une classification du code acte. Pour rappel, la base de données des dépenses nationales étant une extraction plus ciblée de l'Open DAMIR comme exposée dans la partie 1.2.2., il s'agit ici d'une aide considérable à la classification bien que cela ne couvre pas l'ensemble des 1 080 codes actes à classer mais un peu plus d'un tiers. Cette méthode a été suggérée par l'organisme de gestion de l'Open DAMIR lors de notre requête d'aide à la classification. C'est la seule aide qui a pu être envisagée et il semblerait qu'il n'existe aucun fichier ayant déjà réalisé au préalable la classification envisagée dans ce mémoire de leur côté.

- Selon déduction après avoir réalisées toutes les considérations précédentes et en faisant des recherches ciblées sur Internet afin de comprendre la nature d'une prestation.

Par ailleurs, ne sont considérées utiles pour l'étude que les prestations concernant un assuré ordinaire d'une mutuelle. Ainsi, toutes les prestations remboursées par l'Assurance Maladie Obligatoire aux praticiens, établissements de santé tels que hôpitaux, transporteurs, ... sont classifiées dans « Autres (aides, inclassables, ...) » car elles n'entrent pas dans le périmètre des travaux.

Ces travaux de classification furent cependant complexes car ils demandent une expérience confirmée dans le domaine de la santé et une certaine expertise qu'un mémorialiste n'avait pas forcément. Ils furent aussi surtout très chronophages puisqu'il n'y avait *a priori* aucun moyen d'automatiser cette tâche. Par ailleurs, cette étape du mémoire demande de faire preuve de vigilance car elle peut amener à des impacts plus ou moins conséquents pour la suite des travaux. En effet, la tarification se basera sur une classification finale : si de grands montants en jeu sont mal classés, le tarif sera surestimé sur une catégorie d'actes en particulier et sous-estimé sur une autre. Bien heureusement, au regard de la méthode suivie et de sa rigueur, la plupart des actes dont des doutes demeurent sont des actes peu pratiqués, peu connus et peu impactants (en termes de montants). Ils auraient pu être classifiés en « Autres » mais par souci de détails, une catégorie leur a quand même été assignée pour une plupart.

Malgré la conscience de devoir réaliser une classification optimale au vu des enjeux, il fut cependant trop difficile de réconcilier trois points de vue subjectifs : celle de la mutuelle VirtuaMut', celle de la

base Open DAMIR qui adopte le point de vue de la Sécurité Sociale et notre point de vue en tant que mémorialiste.

Enfin, une autre méthode aurait pu être intéressante à croiser aux précédentes pour simplifier la démarche de classification actuelle : regarder par code acte les montants de frais réels en jeu et classer dans « Autres (aides, inclassables, ...) » les montants négligeables, autrement dit, les actes négligeables. Cela n'a cependant pas pu être réalisé pour des raisons logistiques.

2.2.2. Segmentation des prestations de la mutuelle VirtuaMut'

VirtuaMut' présente dans ses bases de données des prestations une codification propre à la mutuelle : ce sont des codes actes internes à l'organisme qui ne correspondent pas tout à fait à une nomenclature connue et donc, à la nomenclature utilisée par l'Open DAMIR. La codification utilisée correspond à celui de l'outil de gestion de la mutuelle. Par exemple, il existe des codes actes utilisés par la mutuelle mais inexistant dans l'Open DAMIR tel que « VERSUP » pour les suppléments pour les verres.

De ce fait, un travail de classification des différentes prestations de la base interne a aussi été réalisé pour la mutuelle afin de garder une cohérence dans les travaux, analogue au travail effectué pour la base Open DAMIR mais simplifié. La démarche suivie dans ce cas-ci est la suivante :

- Les codes actes dont le libellé et le code correspondent à un libellé et code déjà présents dans l'Open DAMIR ont été classifiés dans la même sous-famille/famille d'actes que leur correspondant dans l'Open DAMIR.
- Pour le reste, la sous-famille/famille d'acte a été déduite via le libellé de l'acte (souvent assez clair et suffisamment indicateur).

Quelques précisions cependant sont à mentionner :

- La mutuelle considère les petits appareillages (ex : codes MAC, MAD, ARO, AAD, ...) comme faisant partie de la sous-famille d'acte Pharmacie. Or, une distinction est faite dans les grilles de garanties. Ainsi, il a été considéré pour ces actes-là la sous-famille « Petit appareillage » comme pour l'Open DAMIR et un choix de diverger de l'outil de gestion a été effectué. La famille d'acte reste cependant « Pharmacie » (au lieu de « Matériels médicaux »).
- La mutuelle classe les actes techniques médicaux dans les actes d'imagerie (radiologie) parce qu'ils présentent des taux de remboursement identiques dans les grilles de garantie (ils sont par ailleurs présentés sur la même ligne dans les grilles). Une différenciation a cependant été réalisée lors de notre reclassification pour des raisons de cohérence et de précision.

Une remarque pourrait cependant être faite : l'Open DAMIR ne présente que les prestations remboursables par le régime obligatoire. Or, un intérêt d'un régime complémentaire est de pallier les parts non prises en charge par l'AMO. Bien que l'enrichissement de base soit par la suite visé dans notre étude, ce dernier n'est pas valable pour les actes non remboursables par la Sécurité Sociale (puisque'il n'y a pas de lignes dans l'Open DAMIR qui s'y réfèrent). Ainsi, pour ces actes-là, un tarif sur les autres régions, différent de celui de la région D, ne pourra pas être réalisé.

2.2.3. Segmentation retenue pour la tarification

Le principe de cette partie est de déterminer sur combien et sur quels segments il serait souhaitable de réaliser des tarifs (et de manière plus exacte, de réaliser un calcul de prime pure via la méthode GLM, comme explicité dans le chapitre suivant). Plus concrètement, est-il préférable d'élaborer un tarif pour la famille d'acte « Hospitalisation » ou vaudrait-il mieux en élaborer pour chacune des sous-familles de l'hospitalisation (i.e. les forfaits journaliers, le lit accompagnant, les frais de séjour, ...).

Idéalement, il faudrait réaliser un tarif pour chacune des 36 sous-familles d'actes identifiées en partie 2.2.1. (tableau 10) mais réalistiquement parlant, cela serait non seulement trop long mais aussi trop complexe. En effet, il n'y a pas nécessairement suffisamment de données pour chaque sous-famille d'actes dans la base de la mutuelle (ou même dans l'Open DAMIR du fait de la classification retenue) et les tarifs résultants seraient alors bien trop peu fiables. Il vaut mieux à ce moment-là préférer d'avoir un biais en tarifant plus globalement.

Mais comment mesurer s'il y a « assez » de données ? *A priori*, un tarif est fiable s'il y a suffisamment de sinistres (ou de soins dans le contexte de la santé) pour pouvoir en tirer des conclusions. Moins il y a de données et plus les cas d'exceptions prennent du poids et les conclusions deviennent alors biaisées et peu robustes. Cette question de la suffisance de la donnée et du nombre minimal d'observations à retenir n'a pas de réponse claire et simple : certains chercheurs préconisent de compter environ 10 observations par modalité de variables explicatives [13], d'autres utilisent des formules statistiques afin de calculer la solution mais cela est généralement sous condition de savoir au préalable le modèle et les lois envisagées. Il est par ailleurs parfois accepté qu'il faut 100 têtes en santé (contrat collectif), 500 têtes en arrêt de travail et 10 000 têtes en décès pour avoir une idée suffisamment correcte du risque actuariel pour l'élaboration de tarif [14].

La première *rule of thumb*²⁹ nous indique qu'il faudrait 80 (8x10) observations pour l'élaboration des modèles de régression sur les données de VirtuaMut' et 210 (8x10+13x10) observations pour l'élaboration des modèles de régression sur les données de l'Open DAMIR.

La seconde *rule of thumb* nous indique un minimum de 100 observations (si la notion de « têtes » est interprétée comme étant des observations différentes puisque nous souhaitons non pas tarifier des contrats santé collectifs mais individuels).

Par l'agrégation des lignes dans les jeux de données étudiés, la notion d'observations sera transposée à la notion de quantité d'actes. Ainsi, nous retiendrons la règle suivante : « il faut au moins 210 quantités de soins dans une sous-famille pour pouvoir la retenir comme un segment de tarification ». Cela reste cependant subjectif et des conclusions sur la suffisance de données pourraient en revanche peut-être être tirées une fois que nous serons parvenus à des résultats.

Afin donc de pouvoir déterminer les segments de tarification, nous avons dénombré le nombre d'actes par sous-famille d'actes :

SS_FAM	Open DAMIR		Garantie A	Garantie B
	QTE_ACTE	DENOMBREMENT	QTE_ACTE	QTE_ACTE
Actes d'imagerie, de radiologie et ostéodensitométrie	270 143 907	270 131 079	2 242	17 572
Actes médicaux (techniques)	183 377 431	182 806 405	2 094	16 924
Analyse médicale et examens laboratoires	408 891 232	1 614 160 769	4 812	35 512
Audioprothèse	5 600 955	5 962 790	100	1 543
Autres - dentaire	2 475 715	2 475 713	60	635
Autres - hospitalisation	27 093 101	26 980 072	49	447
Autres - optique	295 262	284 663	0	0
Autres (aides, inclassables ...)	66 284 006	67 156 794	15	396
Autres honoraires médicaux	38	38	0	0
Auxiliaires médicaux (kiné, infirmiers, ...)	3 572 252 328	3 412 593 311	26 756	137 731
Consultations/visites	716 301 912	743 228 264	8 657	58 960
Cure thermale	4 962 098	4 635 369	26	259
Forfait journalier	19 715 371	19 096 033	5 249	41 040
Frais de séjour	69 312 285	68 503 462	691	4 121
Grand appareillage	714 024	743 951	0	0
Honoraires et actes chirurgicaux	44 660 452	44 660 315	311	2 196
Hospitalisation	9 541 030	9 493 526	51	466
Implantologie	1 461 002	1 461 002	6	215
Lentilles	493 929	494 691	138	845
Monture	26 453 985	26 452 370	324	2 459

²⁹ Une règle d'usage, souvent déterminée empiriquement, qui n'a pas de justification théorique poussée mais qui est communément admise et utilisée.

Orthodontie	8 428 369	8 428 014	6	8
Ostéopathie, diététique, chiropractie, ...	5 389 958	5 388 784	366	3 837
Parodontologie	1 549 252	1 549 246	14	138
Petit appareillage	199 334 661	223 002 112	3 834	28 460
Pharmacie	3 249 985 023	3 264 674 252	45 376	346 877
Prothèse	19 202 504	19 202 492	314	1 724
Soins dentaires	111 415 022	111 415 233	1 570	9 674
Transport	1 793 466 485	894 836 072	341	2 956
Vaccin anti-grippe	21 475 554	21 474 354	36	374
Verres	54 497 860	51 556 529	688	5 306
Total général	10 897 264 359	11 105 337 313	104 132	734 104

Tableau 11 : Quantités d'actes par sous-familles

Dans la suite, les segments de tarifications retenus et des explications sur les choix effectués sont présentés.

Grands postes	Familles d'actes	Sous-familles d'actes	Choix de segmentation
Soins courants	Honoraires médicaux	Consultations/visites	1 segment de tarification
		Actes médicaux (techniques)	1 segment de tarification
		Autres honoraires médicaux	Non considéré
	Honoraires paramédicaux	Auxiliaires médicaux (kinésithérapeutes, infirmiers, ...)	1 segment de tarification
	Analyse et examen de laboratoire	Analyse médicale et examen laboratoire	1 segment de tarification
	Imagerie médicale	Actes d'imagerie, de radiologie et ostéodensitométrie	1 segment de tarification
Transport, ambulances	Transport	Non considéré	
Pharmacie	Pharmacie	Petit appareillage	1 segment de tarification
		Grand appareillage	
		Pharmacie	
		Vaccins anti-gripes	
Hospitalisation	Hospitalisation	Honoraires et actes chirurgicaux	1 segment de tarification
		Forfait journalier	1 segment de tarification
		Frais de séjour	1 segment de tarification
		Chambre particulière	
		Lit accompagnant	Non considéré
Autres - hospitalisation			
Dentaire	Dentaire	Prothèse	1 segment de tarification
		Soins dentaires	1 segment de tarification
		Parodontologie	Non considéré
		Implantologie	Non considéré
		Orthodontie	Non considéré
		Autres - dentaire	Non considéré
Optique	Optique	Monture	1 segment de tarification
		Verres	1 segment de tarification
		Chirurgie œil	Non considéré
		Lentilles	Non considéré
		Autres - optique	Non considéré
Aides auditives	Audio	Audioprothèse	1 segment de tarification
		Pile, accessoire	Non considéré
		Autres - audio	Non considéré
Autres	Cure thermique	Cure thermique	Non considéré
	Médecine douce	Ostéopathie, diététique, chiropractie, ...	Non considéré
	Prévention	Prévention	Non considéré
	Prestations supplémentaires	Maternité	Non considéré
		Autres (aides, inclassables ...)	Non considéré

Tableau 12 : Segmentation pour la tarification

Nous retenons donc 13 segments de tarification en tout. Les segments non retenus sont justifiés de la manière suivante :

- Pour ce qui est des « Autres dentaires », « Autres optiques », « Autre audio », « Autre hospitalisation », « Prévention », « Maternité » et « Autres (aides, inclassables ...) » : il n'était pas suffisamment clair quelle prestation et donc, quelle ligne de la grille de garanties de VirtuaMut' il fallait appliquer pour déduire un montant de remboursement de la mutuelle dans l'Open DAMIR (ni même si la garantie A ou B couvre effectivement ces actes-là et quelles lignes des données de la mutuelle s'y réfèrent). Il y a donc trop d'incertitudes. Par ailleurs, pour le cas des « Autres (aides, inclassables ...) », ce sont des prestations qui ne doivent pas être considérées par la mutuelle car elles ne la concernent pas. Pour la maternité, il semblerait que la majorité des prestations qui la concernent, avec les retraitements effectués, soient éparpillées et intégrées dans d'autres sous-familles (que ce soit pour l'Open DAMIR ou pour VirtuaMut').
- Pour le cas des actes de transport, comme mentionné en partie 2.1.2.10., les possibilités sont nombreuses tout comme les doutes sur le périmètre que cette sous-famille d'actes couvre ;
- Pour le cas des « Pile, accessoire », « Orthodontie », « Implantologie », « Autres honoraires médicaux », « Parodontologie », « Cure thermale » : il a été considéré qu'il y avait trop peu de quantités d'actes (y compris quand il y a suffisamment de quantités pour la garantie B mais trop peu pour la A afin de simplifier les travaux). Le cas de « l'audioprothèse » reste une exception comme cette sous-famille d'acte reste primordiale et trop courante pour être ignorée. Il est donc choisi dans un premier temps de la considérer comme étant un segment à tarifer.
- La chirurgie de l'œil n'est pas remboursée par la Sécurité Sociale, elle n'est donc pas présente dans l'Open DAMIR. Par ailleurs, pour les lentilles, seuls les cas de lentilles remboursées par la Sécurité Sociale sont présents dans l'Open DAMIR alors que les cas de lentilles non remboursées par la Sécurité Sociale sont aussi présents dans les données de la mutuelle. Cette dissonance de périmètre conduit à ne pas retenir ce segment.
- Les actes d'ostéopathie, de diététique, de chiropractie, etc. ont des conditions de remboursement de la mutuelle trop fluctuantes (par exemple, différentes d'une année à l'autre) et la classification effectuée présente des incertitudes. Pour simplifier l'étude, il a été considéré que ce segment ne serait pas retenu.
- Les verres n'ont pas été divisés en « simples », « complexes », « très complexes » dans un premier temps. Une étude spécifique sera faite à ce sujet.
- Pour les « Petits appareillages », « Grands appareillages », la pharmacie et les vaccins, c'est la famille d'acte « Pharmacie » qui est retenue comme segment de tarification à cause de l'incertitude de classification liée à des considérations différentes entre VirtuaMut' et Open DAMIR (malgré notre effort d'harmonisation, VirtuaMut' classait à l'origine les petits appareillages comme étant de la pharmacie simple) et à cause de la quantité d'actes parfois insuffisante.
- Pour les « Frais de séjour », la « Chambre particulière » et le « Lit accompagnant » : nous n'avons pas réussi à classer les actes hospitaliers dans ces sous-familles de manière aussi précise et avons, pour la plupart du temps, classé ces actes dans « Hospitalisation », c'est la raison pour laquelle ils n'apparaissent pas dans le tableau 12 (ils sont en fait compris dans « Hospitalisation »). Un seul segment de tarification est donc retenu et est fusionné à « Frais de séjour » comme c'est le remboursement de ce dernier qui a été retenu (cf. partie 2.1.4.7.).

Pour les segments non considérés, une approche différente du GLM sera considérée (dans la mesure du possible, la méthode statistique « directe » qui sera présentée en Chapitre 3) afin de ne pas les exclure complètement du tarif calculé.

Section 3 – Une étude descriptive comparative des deux bases de données

Il s'agit dans cette section de faire une étude descriptive des portefeuilles de VirtuaMut' et de l'Open DAMIR par le biais de deux axes : les adhérents et les dépenses. La dépense (ou frais réels) a été retenue comme variable décrite au lieu des montants de remboursement de la mutuelle car elle est plus facilement interprétable (elle ne dépend pas des garanties).

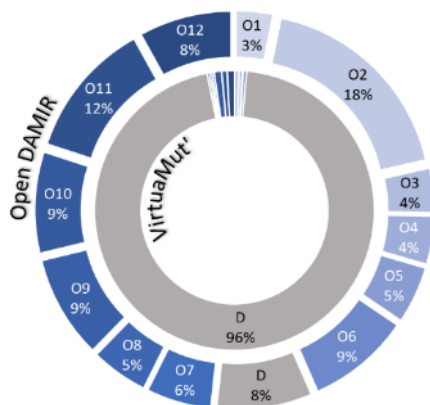
2.3.1. Les adhérents

Pour rappel, les adhérents à la garantie A (ou B) du point de vue de l'Open DAMIR sont supposés être l'ensemble de la population française. Il n'y a pas de condition d'adhésion pour ces deux garanties. La population française (France Métropolitaine, Corse et DOM-TOM) s'élevait à 66 883 761 personnes en 2018 et à 66 977 703 personnes en 2019, soit une évolution de +0,14 %.

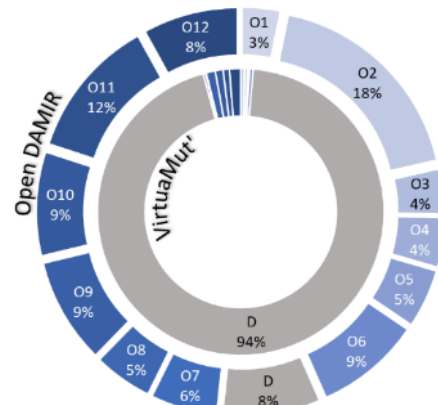
Dans cette étude statistique sur les adhérents, le dénombrement est effectué selon la durée d'exposition, ce qui explique les valeurs décimales pour les individus de VirtuaMut'.

2.3.1.1. Par région

Comparaison de la répartition des adhérents à la garantie A de l'Open DAMIR et de VirtuaMut' en 2018



Comparaison de la répartition des adhérents à la garantie B de l'Open DAMIR et de VirtuaMut' en 2018



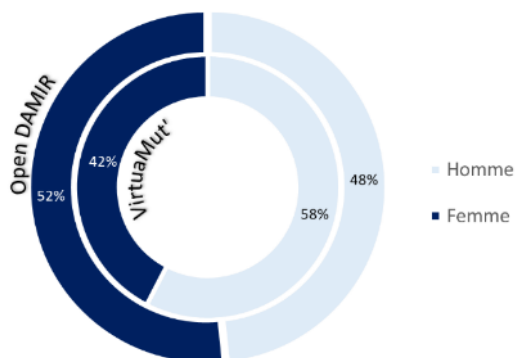
Graphique 3 : Répartition des adhérents en 2018 par région selon la garantie

Les adhérents de VirtuaMut' (96 % pour la garantie A et 94 % pour la garantie B) sont localisés en grande majorité dans la région d'établissement de la mutuelle, i.e. la région D. Le portefeuille d'assurés est donc fortement concentré sur une seule région (d'où l'intérêt de vouloir faire des tarifs sur d'autres régions). Dans l'Open DAMIR en revanche, seulement 8 % des adhérents (pour les deux garanties) seraient répartis sur la région D et dans l'ensemble, il n'est pas observé de concentration forte de manière relative (la région O2 étant la plus concentrée avec ses 18 %).

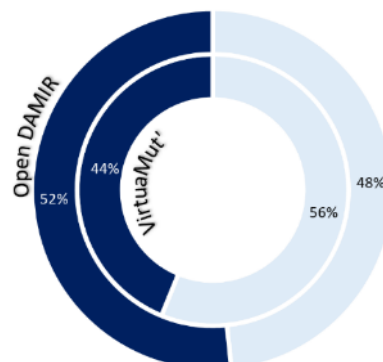
Il n'y a pas d'évolution importante entre 2018 et 2019 en termes de répartition de la population par région mise à part une légère concentration supplémentaire (+1 %) sur la région D pour la garantie B. Les graphiques associés ne seront donc pas présentés.

2.3.1.2. Par sexe

Comparaison de la répartition des adhérents à la garantie A de l'Open DAMIR et de VirtuaMut' en 2018



Comparaison de la répartition des adhérents à la garantie B de l'Open DAMIR et de VirtuaMut' en 2018



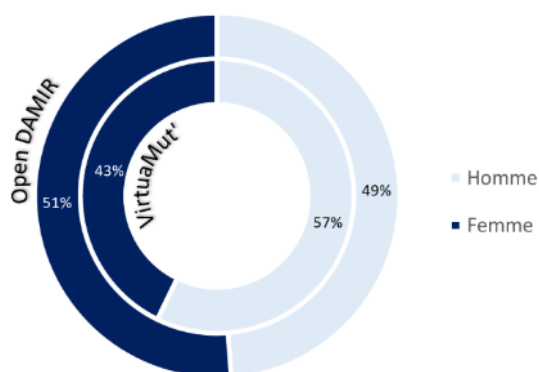
Graphique 4 : Répartition des adhérents en 2018 par sexe selon la garantie

Il s'agit ici du sexe biologique des individus. La parité homme/femme est assez équitable dans la population globale française (48 % d'hommes contre 52 % femmes). Dans celle du portefeuille de VirtuaMut', la répartition est proche pour les deux garanties mais il y a plus d'hommes que de femmes comparée à l'Open DAMIR (écart de près de 10 %).

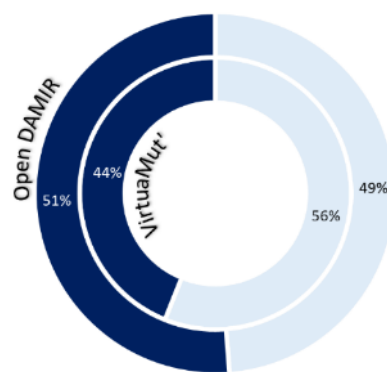
L'évolution des graphiques entre 2018 et 2019 est peu significative (moins de 1 % d'écart).

Si la région D est isolée et si seuls les adhérents vivant dans cette région sont pris en compte, il peut être observé que la répartition reste plutôt stable :

Comparaison de la répartition des adhérents à la garantie A de l'Open DAMIR et de VirtuaMut' en 2018 (sur la région D)



Comparaison de la répartition des adhérents à la garantie B de l'Open DAMIR et de VirtuaMut' en 2018 (sur la région D)



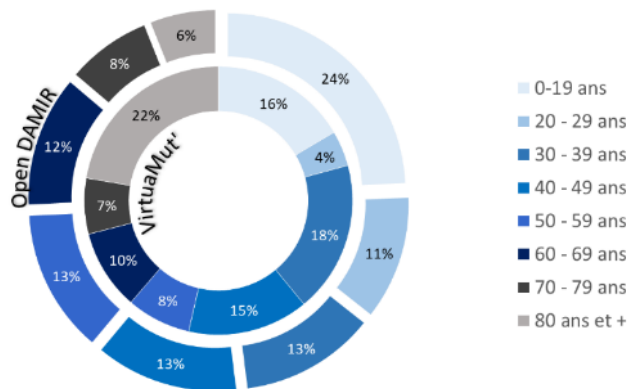
Graphique 5 : Répartition des adhérents en 2018 par sexe selon la garantie sur la région D

La parité homme/femme de la population française sur la région D est similaire à celle de la population française sur toutes les régions confondues (écart de maximum 1 %). Le portefeuille de VirtuaMut' sur la garantie A ou B étant de base concentré sur la région D, la répartition est aussi stable. Ainsi, les conclusions précédentes sur la comparaison entre Open DAMIR et VirtuaMut' restent valables.

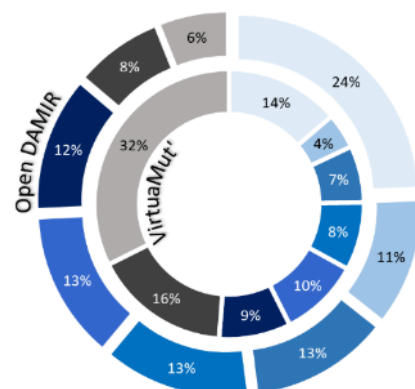
Il n'y a pas de changement de répartition significatif entre 2018 et 2019 (moins de 1 %).

2.3.1.3. Par classe d'âge

Comparaison de la répartition des adhérents à la garantie A de l'Open DAMIR et de VirtuaMut' en 2018



Comparaison de la répartition des adhérents à la garantie B de l'Open DAMIR et de VirtuaMut' en 2018



Graphique 6 : Répartition des adhérents en 2018 par classe d'âge selon la garantie

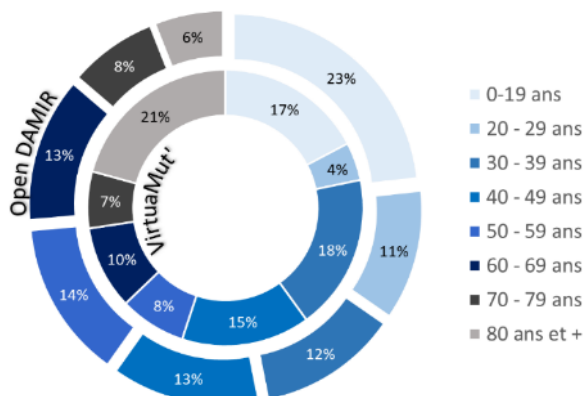
Pour la garantie A, le portefeuille de VirtuaMut' est légèrement plus âgé (les tonalités de couleurs foncées et grisées occupent un peu plus de 1/3 du cadran contre 1/4 pour l'Open DAMIR) mais certaines proportions de classes d'âge se ressemblent : les âges de 40 à 49 ans, de 60 à 69 ans et de 70 à 79 ans (le critère d'écart acceptable retenu pour émettre un tel avis est de 2 %). Cela signifie que pour ces trois classes d'âge, bien que le portefeuille de VirtuaMut' soit concentré dans la région D, la répartition des adhérents est représentative de la population française (dans son entièreté).

Pour la garantie B, le portefeuille de VirtuaMut' est globalement encore plus âgé que celui de la garantie A (environ 57 % des adhérents sont âgés de plus de 60 ans) et il y a plus de disparité dans la répartition en classes d'âge comparée à l'Open DAMIR.

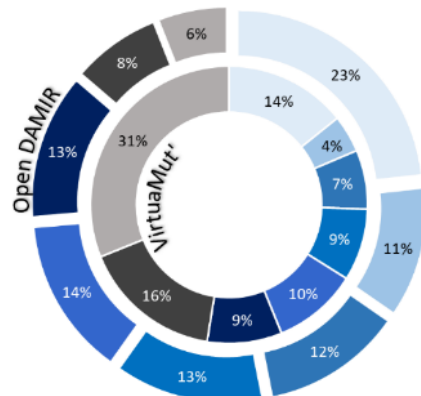
Il n'y a pas de changement significatif entre 2018 et 2019 (moins de 1 % d'écart par classe d'âge).

Si la région D est isolée et si seuls les adhérents vivant dans cette région sont pris en compte, il peut être observé que la répartition reste plutôt stable :

Comparaison de la répartition des adhérents à la garantie A de l'Open DAMIR et de VirtuaMut' en 2018 (sur la région D)



Comparaison de la répartition des adhérents à la garantie B de l'Open DAMIR et de VirtuaMut' en 2018 (sur la région D)



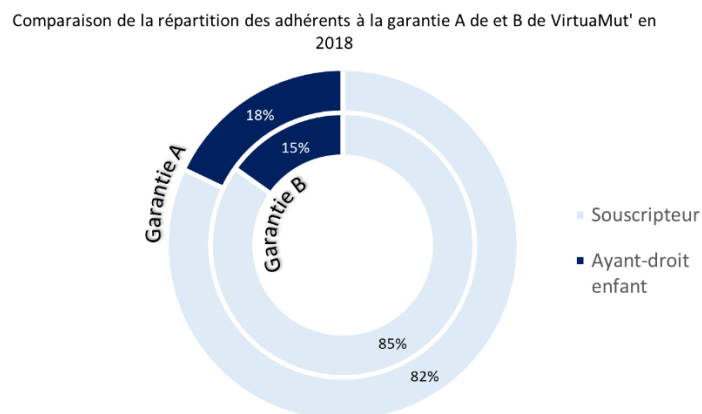
Graphique 7 : Répartition des adhérents en 2018 par classe d'âge selon la garantie sur la région D

La répartition en matière de classes d'âge de la population française sur la région D est similaire à celle de la population française sur toutes les régions confondues (écart de maximum 1 % par classe d'âge). Le portefeuille de VirtuaMut' sur la garantie A ou B étant initialement concentré sur la région D, la répartition est aussi stable. Ainsi, les conclusions précédentes restent valables.

Il n'y a pas de changement de répartition significatif entre 2018 et 2019 (moins de 1 %).

In fine, l'individu moyen du portefeuille de VirtuaMut' pour la garantie A est un homme de 51 ans vivant en région D ; pour la garantie B, c'est un homme de 59 ans vivant en région D.

2.3.1.4. Par qualité de bénéficiaire



Graphique 8 : Répartition des adhérents en 2018 par qualité de bénéficiaire selon la garantie

La répartition de la qualité des adhérents (entre souscripteur/conjoint et ayant droit enfant) est proche pour la garantie A et B. Pour le cas de l'Open DAMIR, la variable de qualité a été créée selon l'âge des bénéficiaires de la ligne (cf. partie 2.1.4.5. les individus de moins de 19 ans sont mis en ayants droit enfants et les individus âgés de plus de 19 ans en souscripteur/conjoint) afin de ne pas double-compter le nombre de bénéficiaires exposés par ligne. Il y aurait par exemple double-comptage s'il était associé le même nombre d'assurés exposés pour les lignes ayant des caractéristiques de région, de classe d'âge et de sexe communes mais avec une qualité de bénéficiaire différente (ils ne peuvent pas être à la fois souscripteurs et ayants droit). Cela fut nécessaire car il n'y avait pas d'informations sur la qualité des individus de la population française et bien que nous avons pu trouver des données de marché sur la répartition entre ayants droit et souscripteur pour le cas du régime local, rien n'a pu être trouvé pour le régime général de la Sécurité Sociale...

Par cette construction, les variables **CLASSE_AGE** et **QLT** (qualité) seront probablement colinéaires (au moins pour le cas des données de l'Open DAMIR comme la construction de la dernière dépend de la première). Cela sera reconfirmé dans des études ultérieures comme cela aura son importance dans l'une des méthodes de tarification envisagées (modèles linéaires généralisés, cf. Chapitre 3).

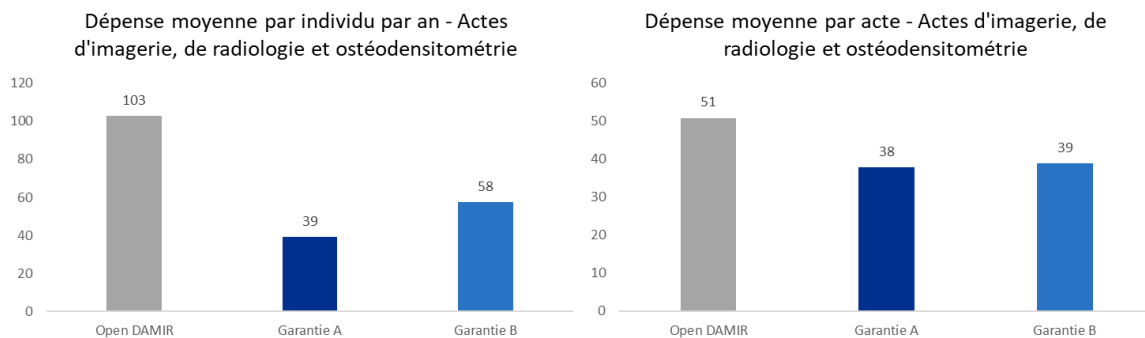
Cela est d'autant plus renforcé par le fait que la répartition entre souscripteurs/conjoints et ayants droit enfants pour les garanties A et B sont très proches de la répartition entre classe d'âge 0 et le reste (15 % d'ayants droit enfants pour la garantie B contre 14 % d'adhérents âgés de moins de 19 ans dans le graphique 6). Ces variables risquent donc aussi d'être colinéaires dans le cas des données de VirtuaMut'.

Il n'y a pas de changement de répartition significatif entre 2018 et 2019 (moins de 1 %).

2.3.2. Les dépenses

Les montants de dépenses affichés dans cette sous-section seront normalisés par le nombre de bénéficiaires exposés (ou par la quantité d'actes) afin de visualiser une dépense moyenne par adhérent (ou par acte) par sous-famille d'actes et rendre les montants comparables. Pour des questions d'interprétation et de simplifications, seuls les segments de tarification retenus en sous-partie 2.2.3. ne seront analysés. Il est aussi à préciser que les montants dépendent des retraitements faits sur les bases et qu'il faudra garder un certain regard critique sur la véracité des statistiques.

2.3.2.1. Actes d'imagerie, de radiologie et ostéodensitométrie



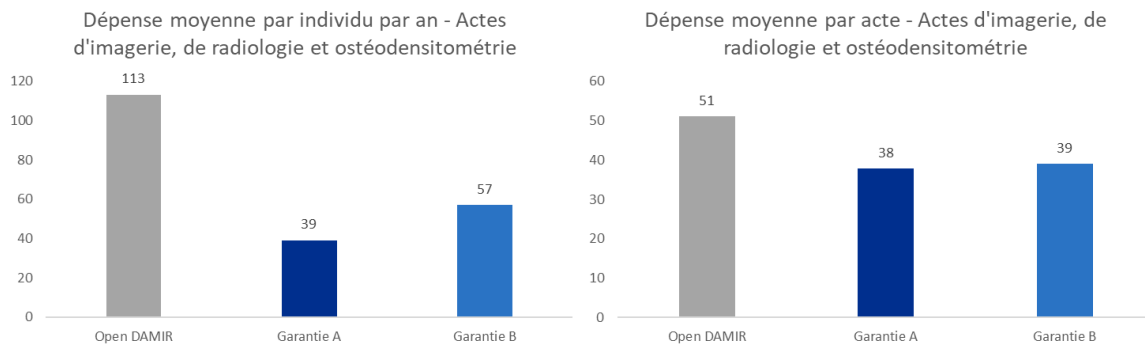
Graphique 9 : Dépense moyenne pour les actes d'imagerie, de radiologie et ostéodensitométrie

En moyenne, sur une année, un Français dépense³⁰ 103 € pour des actes d'imagerie, de radiologie et d'ostéodensitométrie (scanner, IRM, ...) et le prix moyen d'un de ces actes s'élève à 51 € d'après le graphique 9. À titre indicatif [15], le prix d'une radiographie varie de 27,50 € à 66,42 €, celui d'une IRM peut dépasser les 200 € et l'échographie dépend de la période de grossesses (environ 50 € à 80 € selon les trimestres). Ces montants semblent donc plutôt cohérents, d'autant plus que les individus concernés par de tels actes ont aussi tendance à en faire plusieurs par an.

En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 39 € par an pour de tels actes pour un prix unitaire de 38 € l'acte. Cela est inférieur à ce qui est visible sur l'Open DAMIR et pourrait s'expliquer par un portefeuille beaucoup plus restreint (seulement une petite fraction de la population française est prise en compte – moins de 0,001 %) et donc, présente moins de diversité d'actes d'imagerie, de radiologie et d'ostéodensitométrie. Il n'y aurait par exemple dans le portefeuille d'origine de VirtuaMut' pas ou peu d'adhérents atteints d'une maladie particulière (par exemple, l'ostéoporose qui génère de nombreuses fractures d'os) et certaines prestations coûteuses ne seraient pas présentes, du moins, comparés à la prise en compte de la population française dans son entièreté. L'échantillon de la population française que représentent les adhérents de VirtuaMut' ne serait alors pas représentatif de la population française globale. Cela était observable avec l'étude descriptive sur les adhérents : les portefeuilles de la mutuelle sont plus âgés. Nous nommerons ce premier argument par « portefeuille restreint » (qui fera aussi écho au fait que l'offre de soin est plus restreinte pour les adhérents de la mutuelle qui habitent dans une région précise qui leur demande parfois de changer de lieu pour atteindre les soins adéquats) afin d'élaguer les prochaines analyses. Une autre idée serait « l'effet région » et elle sera vérifiée par le graphique 10 ci-après.

³⁰ Ceci est un abus de langage. L'individu ne dépense pas le montant en tant que tel car celui-ci ne tient pas compte des remboursements qu'il recevra de la part de mutuelle et / ou de la Sécurité Sociale. Il s'agit ici de frais réel donc plutôt de « coût ».

Par rapport à la garantie A, les adhérents à la garantie B présentent un coût moyen de l'acte très proche mais une dépense annuelle plus élevée. Cela pourrait s'expliquer par un meilleur remboursement pour ces actes de la garantie B qui amènerait les adhérents à consommer plus (thématique de l'aléa moral), par un effet « âge » (le portefeuille de la garantie B est plus âgé) ou encore, par l'effet « portefeuille restreint » (il y a 10 fois plus d'adhérents à la garantie B qu'à la garantie A) qui agirait sur une échelle moins importante que précédemment.

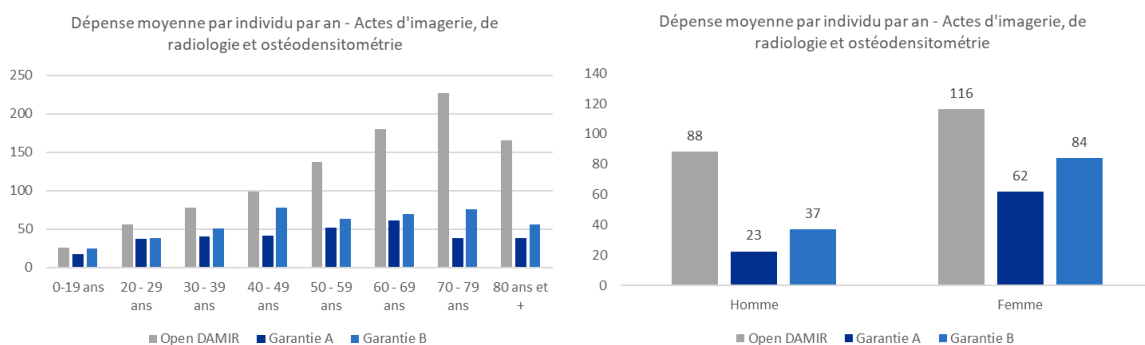


Graphique 10 : Dépense moyenne pour les actes d'imagerie, de radiologie et ostéodensitométrie sur la région D uniquement

Une explication possible de la différence entre les données issues de VirtuaMut' et de l'Open DAMIR pouvait être « l'effet région » : la région D pourrait hypothétiquement présenter une spécificité qui ferait que ses habitants ont tendance à consommer plus ou moins d'actes d'imagerie, de radiologie et d'ostéodensitométrie ou encore, une spécificité pourrait engendrer la consommation d'un acte en particulier au détriment d'autres. Les prix pourraient aussi être différents.

Le graphique 10 nous enseigne que cela serait plausible (augmentation d'environ 10 % de la dépense annuelle par individu). Cependant, et ce commentaire vaudra pour la suite de l'écrit, il reste assez difficile de juger les « effets région » : une grande part de subjectivité subsiste (car au final, peut-être que 10 € d'écart n'est pas signe d'un effet région conséquent) et nous ne savons pas à l'avance à quoi nous attendre. Il est aussi difficile de trouver des données de marché à ce sujet comme cela dépend grandement de la segmentation effectuée (autrement dit, il y a peut-être des études sur la différence de prix d'une consultation à Paris et à Toulouse mais cela risque d'être ciblé sur une consultation particulière comme celle du médecin généraliste). Les « effets région » seront étudiés plus en détail à l'étape de réalisation de l'extension des tarifs (cf. Section 3, Chapitre 3).

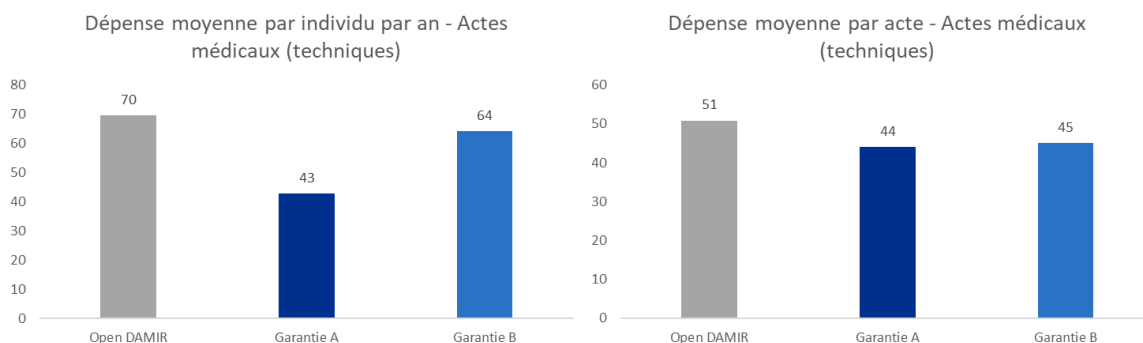
Nous nous contenterons pour l'instant de retenir un « effet région » que si cette dernière semble significative.



Graphique 11 : Dépense moyenne pour les actes d'imagerie, de radiologie et ostéodensitométrie selon les âges et le sexe

Le graphique 11 indique que la dépense annuelle moyenne pour ces actes augmente avec l'âge (jusqu'à 70 ans) et que les femmes présentent des dépenses annuelles plus élevées (les hommes, sauf cas spécifiques, ne sont par exemple pas concernés par les échographies pour cause de grossesse).

2.3.2.2. Actes médicaux (techniques)

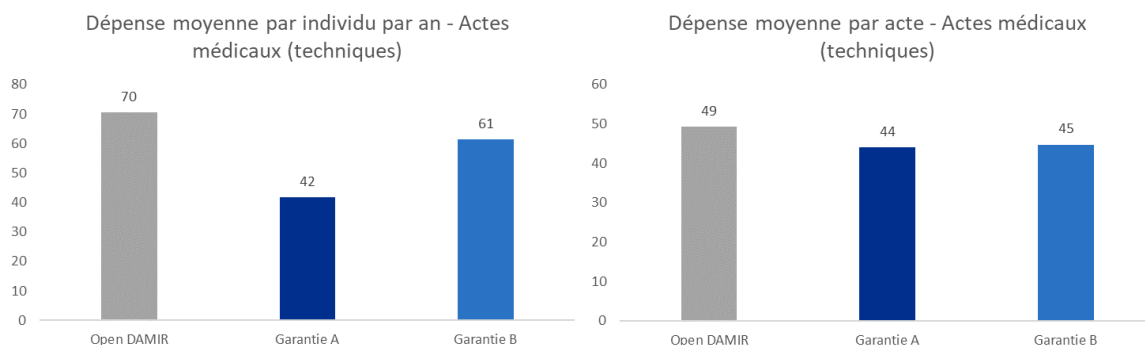


Graphique 12 : Dépense moyenne pour les actes techniques médicaux

En moyenne, sur une année, un Français dépense 70 € pour des actes techniques médicaux (y compris actes de spécialiste comme par exemple l'ablation de grain de beauté) et le prix moyen d'un de ces actes s'élève à 51 €. À titre indicatif [16], le prix d'un fond d'œil est d'environ 30 € sans dépassement d'honoraires et celui d'une ablation de grain de beauté à souvent plus de 130 €. Les actes sont cependant très divers dans cette sous-famille d'actes.

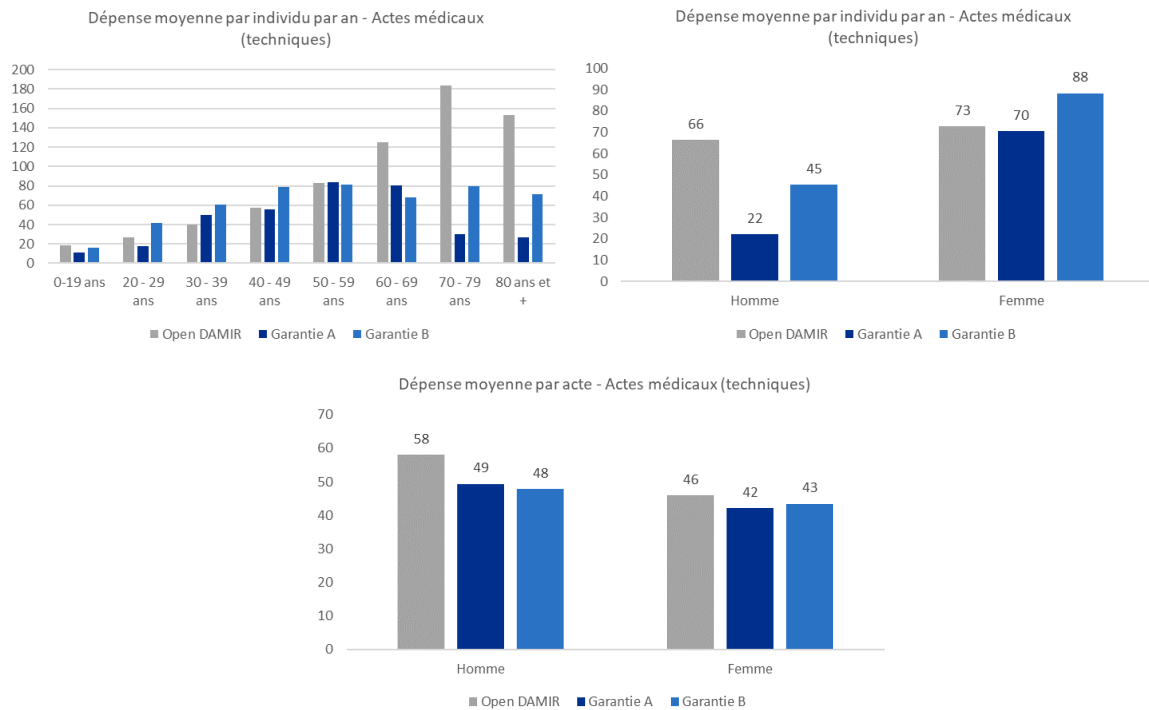
En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 43 € par an pour un prix unitaire de 44 € l'acte. Cela est inférieur à ce qui est visible sur l'Open DAMIR et pourrait s'expliquer par l'argument du « portefeuille restreint ».

Par rapport à la garantie A, les adhérents à la garantie B présentent un coût moyen de l'acte très proche mais une dépense annuelle plus élevée. Cela pourrait s'expliquer par un meilleur remboursement pour ces actes de la garantie B qui amènerait la population à consommer plus (thématique de l'aléa moral), par l'effet « âge » ou encore, par l'effet « portefeuille restreint ».



Graphique 13 : Dépense moyenne pour les actes techniques médicaux sur la région D uniquement

Le graphique 13 indique qu'il n'y a pas, *a priori*, d'« effet région ». La région D est assez représentative de ce qu'il se passe sur les autres régions.



Graphique 14 : Dépense moyenne pour les actes techniques médicaux selon les âges et le sexe

Le graphique 14 indique que la dépense annuelle moyenne pour ces actes augmente avec l'âge de manière générale (jusqu'à 70 ans pour l'Open DAMIR et jusqu'à 60 ans pour VirtuaMut') et que les femmes consomment plus annuellement (cela est plus marqué chez VirtuaMut' que chez l'Open DAMIR). En revanche, en termes de coût unitaire d'actes, les hommes l'emportent sur les femmes, ce qui confirme les dires de la fin de la sous-partie 2.1.3.4.

2.3.2.3. Récapitulatif des différentes analyses pour les autres sous-familles d'actes

Par souci de longueur, il sera mentionné ici les points notables qui ressortent du même exercice fait sur les autres sous-familles d'actes sur la base du schéma des analyses précédentes. Le détail (tel que présenté pour le cas des actes d'imagerie, de radiologie et ostéodensitométrie ou des actes techniques médicaux) sera quant à lui consultable en Annexe 11.

Annexe 11
Autres dépenses

Suite à l'analyse descriptive par segment tarifaire, nous en retenons les faits principaux suivants :

- Les femmes ont tendance à « coûter plus cher » que les hommes en matière par exemple, de prothèses dentaires, d'auxiliaires médicaux et de consultations et visites.
- Pour le cas des analyses médicales et des examens laboratoires, il est observé que les individus ayant souscrit à la garantie B sont assez représentatifs des habitudes de consommation de la population française dans son entièreté. C'est aussi le cas pour les soins dentaires et les montures mais cette fois-ci, les individus ayant souscrit à la garantie A sont aussi représentatifs au même titre que ceux ayant souscrit à la garantie B. En revanche, pour les consultations et visites par exemple, cette représentativité n'est pas reconduite (au contraire).
- Un potentiel effet « région » serait à noter pour les segments des analyses médicales et des examens laboratoires, des auxiliaires médicaux, des consultations et visites, des prothèses dentaires et des soins dentaires.
- L'analyse des verres et montures, des soins dentaires, des prothèses dentaires, des auxiliaires médicaux et des audioprothèses présente des similarités dans le sens où les dépenses sont fortement corrélées aux âges et retraduisent de manière assez fidèle les besoins liés à l'évolution

de la santé face à la vieillesse. Ainsi, par exemple, pour les prothèses dentaires, il existe un sommet de dépenses aux alentours des âges de 60 à 69 ans qui semble cohérent avec la perte de la dentition au fur et à mesure du temps. Pour les auxiliaires médicaux, le pic est très accentué et concerne les personnes âgées de plus de 70 ans.

- Pour les verres spécifiquement, une sorte de « palier » est observable entre les deux classes d'âge de 30-39 ans et 40-49 ans. Cela pourrait s'expliquer par diverses raisons : le changement de la nature des verres (passage aux progressifs par exemple), le changement des caractéristiques de verres (vision aggravée donc verres plus épais et plus coûteux à affiner, prise d'options de plus en plus nombreuses, ...), la défaillance de l'état des yeux au fil du temps, etc.
- Pour les montures spécifiquement, la forme en cloche des dépenses avec le sommet pour les individus d'âge moyen (entre 40 et 59 ans) pourrait être due à la déclinaison des yeux avec l'âge et donc, le début des équipements optiques pour une partie de la population française.
- Il semblerait que la sous-famille « audioprothèses » intègre aussi les accessoires auditifs et éventuellement les piles. Par ailleurs, une forte dépense par individu par an est observée pour la garantie B (par comparaison relative avec l'Open DAMIR et la garantie A). Cela pourrait découler d'un effet « âge » comme ce portefeuille est plus âgé.
- Pour les forfaits journaliers, un montant de dépense par acte bien trop élevée est observé (188 € au lieu des 20 € attendus par jour) pour l'Open DAMIR. À l'heure actuelle, nous ne trouvons pas d'explication à ce montant, nous décidons donc d'écarter ce segment des segments à tarifier par GLM.
- Enfin, en ce qui concerne l'hospitalisation et la pharmacie, l'étude descriptive a permis de mettre en évidence une différence de classification entre les deux bases de données pour les grands appareillages : ils sont classés en pharmacie selon notre classification harmonisée mais la mutuelle les a classés en hospitalisation sans indiquer de libellés différents, ce qui a rendu la détection plus ardue. Comme il n'y a pas moyen d'effectuer une différenciation dans les données de VirtuaMut' entre les grands appareillages et l'hospitalisation, les grands appareillages de l'Open DAMIR seront transférés pour la suite des travaux dans le segment « Hospitalisation ». De plus, comme la grille de garanties de VirtuaMut' n'indique pas de prestations claires pour les grands appareillages (il est mention de remboursement étudié au cas par cas), faute de solution, il sera pris les prestations de l'audioprothèse (comme les grands appareillages comprennent parfois les appareils auditifs).
- Dernièrement, l'étude descriptive nous a aussi permis de nous rappeler d'un point de détails que nous avons oublié jusqu'alors : la Sécurité Sociale ne rembourse pas systématiquement toutes les lentilles (seulement celles « acceptées ») alors que la mutuelle rembourse un éventail plus large de lentilles (celles acceptées et refusées par la Sécurité Sociale). Par cette divergence, la sous-famille des lentilles sera exclue d'une tarification par GLM.

L'étude descriptive s'est donc révélée primordiale dans la prise de conscience de certains sujets.

Pour faire le point

Il a été mentionné dans ce chapitre un très grand nombre de retraitements, qui ne sont pas nécessairement tous pertinents sur d'autres jeux de données. Cela permet néanmoins d'avoir une idée des éventualités. Cela permet aussi de voir à quel point la qualité des données issues des systèmes d'information d'un organisme d'assurance peut être cruciale en termes de temps et de coût. La directive Solvabilité II, en mettant un point d'honneur à cette nouvelle notion, notamment avec le rapport de la fonction actuarielle ou la politique écrite de la qualité des données, a été bien pensée. Il a aussi été fait mention d'éléments liés à l'harmonisation et à la mise en cohérence entre deux bases de données. De manière non exhaustive, la liste suivante récapitule les principaux points abordés :

- Le travail en année de règlement (et non en année de soin) imposé par une limite d'accès aux données récentes de l'Open DAMIR.
- L'importance de la différenciation entre quantité d'acte, dénombrement et exposition (totale). Notamment pour la dernière, sa détermination et l'importance de cohérence de périmètre entre la valeur prise par la variable et l'objectif de tarification poursuivi.
- Le traitement des adhérents qui présentent des changements de garantie ou des périodes vides de couverture d'assurance.
- L'estimation du nombre de bénéficiaires exposés via l'ajout de données d'une base démographique de l'INSEE.
- Les vérifications traditionnelles à faire sur les différentes variables (valeurs aberrantes ou manquantes par exemple).
- Le traitement des quantités dans le cas où celles renseignées sont des coefficients globaux de nomenclature.
- Le traitement des compléments, suppléments, majorations d'actes.
- Une sélection des variables pertinentes étapes par étape.
- ...

Par ailleurs, une classification des codes actes de l'Open DAMIR et de la mutuelle a été réalisée et bien qu'éreintante et chronophage, elle était nécessaire pour la détermination des segments tarifaires. Il ne faut pas oublier que segmenter de manière excessive signifie d'une part, un temps considérable par la suite pour les travaux à réaliser (plus il y a de segments, plus il y a de tarifs à réaliser et donc de manœuvres à effectuer) mais d'autre part, cette « sur-précision » pourrait conduire à des segments ayant trop peu d'informations et conduisant donc à des résultats biaisés et / ou peu fiables. La classification étant en partie subjective, elle sera source de biais.

L'importance de la démarche de statistique descriptive a aussi été ressentie puisqu'elle a permis de remettre en cause la segmentation retenue (les grands appareillages sont transférés en hospitalisation). De manière générale, la population de VirtuaMut' est beaucoup plus concentrée géographiquement dans la région D (environ 93 % des adhérents y habitent), est plus masculine et est plus âgée.

En somme, ce que nous souhaitons démontrer avant tout dans ce chapitre, c'est qu'avant de se plonger dans la production massive de tarifs et de résultats, avant toute phase opérationnelle, il est primordial de passer un temps suffisant sur les matières premières qui sont ici les données : à les explorer, à les comprendre, à les retraiter, toujours en gardant en mémoire l'objectif fixé.

Chapitre 3 – « Le prix d'une chose, c'est l'idée qu'on y attache »³¹

Après l'exploration des données vient la phase d'exploitation et de production. Maintenant que la base de données de VirtuaMut' peut être virtuellement enrichie par les individus constituant la base de l'Open DAMIR, il convient de s'atteler aux objectifs opérationnels de ce mémoire : déterminer un tarif par région pour la garantie A et la garantie B de la mutuelle.

Pour rappel, à l'issue du Chapitre 2, sont à disposition trois jeux de données :

- Un qui fusionne les 24 jeux initiaux de l'Open DAMIR,
- Un pour la garantie A (entrée de gamme),
- Un pour la garantie B (milieu de gamme).

Les techniques de tarification sont cependant nombreuses et l'un des premiers carrefours de choix rencontrés est celui entre l'approche statistique classique (à savoir, les GLM³² notamment) ou l'approche d'apprentissage statistique (i.e. la *data science* avec les arbres CART, les réseaux de neurones, les *Random Forest* ...). Il a été retenu l'approche qui privilégie la modélisation (car c'est avant tout un travail de modélisation qui est visé) aux dépens de l'approche *data-driven* (qui laisse parler les données), soit, celle de la statistique classique.

Par ailleurs, d'après une conférence de Xavier CONORT (2018) [17], les GLM semblent être appréciés des actuaires, entre autres, pour leur transparence et la maîtrise de la relation entre risques et facteurs de risque. Ce qui est voulu est un modèle qui serait simple à expliquer à des administrateurs dont les connaissances actuarielles peuvent parfois être limitées ; là où la *data science* a tendance à laisser un ressenti de boîte noire (bien qu'elle permette souvent d'accélérer significativement le chemin vers l'obtention des résultats).

Il s'agira donc dans ce chapitre de présenter dans un premier temps les différentes méthodes statistiques classiques de tarification dont il sera fait recours avec les hypothèses liées. Dans un second temps, il s'agira de modéliser le tarif sur la région D pour la mutuelle et enfin, d'étendre ce tarif à un périmètre national. Des études complémentaires seront réalisées à la toute fin pour assouvir certains points de curiosité rencontrés lors de cette phase opérationnelle.

³¹ Delphine DE GIRARDIN, *Les maximes et pensées* (1855)

³² Techniquement, il faudrait utiliser l'abréviation MLG (modèles linéaires généralisés) mais pour des raisons d'habitudes et du fait que l'abréviation GLM, qui vient de l'anglais Generalized Linear Model, soit tout aussi souvent usitée, cette dernière sera retenue.

Section 1 – Rappels théoriques

3.1.1. Rappel sur la tarification

3.1.1.1. Les primes d'assurance

Comme mentionné en introduction de Chapitre 1, la France possède un système de protection sociale qui permet une certaine sérénité d'esprit financière en cas de maladie. Bien que le reste à charge soit remarquablement bas par rapport à certains pays tels que les États-Unis et que la Sécurité Sociale couvre une bonne partie des dépenses notamment liées aux actes lourds ou anxiogènes (telle qu'une hospitalisation), aux yeux d'un ménage français, sa couverture pourrait paraître insuffisante en ce qui concerne, par exemple, les prothèses auditives ou l'optique. Afin de pallier cela, un individu va alors chercher à se couvrir davantage par le biais des régimes complémentaires et ira alors souscrire à un contrat d'assurance santé chez un organisme assureur qui, dans notre cas d'études, est la mutuelle VirtuaMut'. Ainsi, en échange d'une protection supplémentaire qui se traduira éventuellement par la suite, en cas de nécessité de soins avérée, par une somme pécuniaire, l'adhérent devra verser une cotisation à l'organisme assureur (selon les contrats, annuellement, trimestriellement, mensuellement, ...). En langage assurantiel, cette cotisation porte aussi le nom de prime³³.

Contrairement à l'achat d'un bien meuble, cette prime, qui représente en fait le prix d'assurance, est fixée avant même qu'un sinistre survienne et donc, avant même de savoir si le coût du sinistre sera plus cher. Il s'agit ici du principe de cycle de production inversé en assurance. Les assureurs distinguent par ailleurs plusieurs types de primes.

La **prime pure**, appelée aussi **prime technique**, correspond au coût du risque. C'est le montant que VirtuaMut' facture à son adhérent afin, *a priori*, de payer son sinistre et seulement son sinistre. C'est le montant du sinistre moyen auquel devra faire face l'assureur. C'est donc une espérance des pertes. Il existe différents risques en assurance santé (ce sont par exemple toutes les sous-familles d'actes qui ont été présentées en Chapitre 2) et à chaque risque doit être associé une prime pure. La tarification consistera justement à déterminer ces primes pures.

La **prime pure globale** est la somme des primes pures de ces différents risques sous hypothèse d'indépendance de ces derniers. En réalité, les risques ne sont pas totalement indépendants (une consultation chez le médecin conduit souvent à des achats de médicament en pharmacie par exemple) mais afin de pouvoir effectuer notre étude, une telle hypothèse théorique sera prise.

La **prime commerciale** est quant à elle ce que l'assuré paiera réellement. C'est la prime pure qui est majorée par une marge de sécurité que se permet l'organisme assureur pour faire face à la volatilité des risques (chargement de sécurité) et par différents frais et taxes. Les frais sont par exemple des frais de gestion de sinistres ou encore des frais liés à la commission des apporteurs de contrat (courtiers, agents généraux). Les taxes sont reversées au gouvernement.

Ainsi :

$$\text{Prime commerciale} = \text{Prime pure} + \text{Chargement de sécurité} + \text{Frais} + \text{Taxes}$$

Dans le cadre des travaux, seule la prime pure sera d'intérêt et par abus de langage, elle sera souvent désignée par le terme de « tarif ».

³³ La différence entre les termes « prime » et « cotisation » est floue et est sujet à débat. Dans cet écrit, elles seront considérées comme similaires.

3.1.1.2. La prime pure

Afin de calculer la prime pure d'un contrat d'assurance santé, plusieurs méthodes des statistiques classiques existent. Une démarche similaire à celle adoptée dans le mémoire de Fatemeh ABDOLLAHI (*Tarifcation d'une complémentaire santé à destination des séniors, modulaire par poste de garanties et l'impact sur la solvabilité*, 2017) sera retenue.

Il sera priorisé pour les travaux la méthode traditionnelle de « fréquence x coût moyen ». Ainsi, pour chaque risque :

$$(1) \text{ Prime pure} = \text{fréquence} \times \text{coût moyen}$$

Où

$$\text{Fréquence} = \frac{\text{Nombre total de sinistres}}{\text{Exposition totale}} = \frac{\text{Quantité d'actes totale}^{34}}{\text{Nombre total de bénéficiaires exposés}}$$

Et

$$\text{Coût moyen} = \frac{\text{Coût total des actes de soin}}{\text{Quantité d'actes totale}}$$

Autrement dit :

$$(2) \text{ Prime pure} = \frac{\text{Coût total des actes de soin}}{\text{Nombre total de bénéficiaires exposés}}$$

La prime pure telle que définie dans l'égalité précédente sera par ailleurs dans la suite dénommée « consommation » (terme fréquemment utilisé en santé).

Déterminer la prime pure en multipliant une fréquence par un coût n'est pas une équation arbitrairement choisie et il existe un fondement mathématique qui justifie une telle formule : c'est le « modèle collectif » en assurance.

Pour un risque donné k , le coût total des soins³⁵ du risque k noté X_k est :

$$X_k = \sum_{i=1}^N S_i$$

Où :

- $N \in \mathbb{N}$ est la variable aléatoire représentant le nombre total de soins sur la période considérée ;
- $S_i \in \mathbb{R}^+$ est la variable aléatoire représentant le montant du $i^{\text{ème}}$ soin avec $i \in \{1, \dots, N\}$;
- Les S_1, S_2, \dots, S_N sont i.i.d. à une variable S et indépendants de N ;
- Avec la convention que $X_k = 0$ si $N = 0$.

À titre de précision, dans notre cas de données agrégées, la formule est toujours valable mais il faut imaginer que la variable N soit en fait la somme de variables n_p où p désigne le numéro d'un agrégat d'individus. De même, il faudrait définir une variable S_{n_p} pour le montant de soin total du $p^{\text{ème}}$ agrégat d'individus. De ce fait, S_{n_p} est en fait une somme de variables S_i (p fixé, i flottant). Cela complexifie néanmoins les équations et leurs notations : la somme précédente présente un incrément de 1 ; dans notre cas, les incréments seraient les n_p et les valeurs des S_i ne sont pas connus, seulement celles des S_{n_p} .

³⁴ La seconde égalité utilise des termes plus adaptés à nos études mais cela revient à la même chose.

³⁵ Par abus de langage, le terme soins désignera ici à la fois les actes de soin mais aussi les équipements de soin.

Ainsi :

$$X_k = \sum_{n_p \in \{n_1, \dots, n_{p_{total}}\}} S_{n_p}$$

Les S_{n_p} restent i.i.d. et indépendants de N (où $N = n_1 + \dots + n_{p_{total}}$).

Pour une compréhension facilitée, nous continuerons cependant de présenter la théorie d'une manière usuelle et usitée. Il sera cependant toujours possible de l'adapter en imaginant non pas des variables liées à un individu mais des variables liées à un groupe d'individus (et donc, décomposer chaque variable en une somme de sous-variables).

En outre, la prime pure est l'espérance mathématique du coût annuel des soins déclarés d'un risque k à l'assureur (c'est ce que ce dernier s'attend à payer en moyenne pour ce risque sur l'année). Comme les S_i et N sont indépendants :

$$\text{Prime pure} = \mathbb{E}[X_k] = \mathbb{E}[S] \times \mathbb{E}[N]$$

Annexe 12

Où :

- $\mathbb{E}[S]$ est l'estimation du coût moyen ;
- $\mathbb{E}[N]$ est l'estimation de la fréquence (au sens du nombre de soins moyens).

Ceci explique donc la forme multiplicative de la prime pure (cf. équation (1)).

3.1.1.3. Méthode 1 de tarification : le modèle linéaire généralisé (GLM)

Annexe 13
GLM

Une première méthode serait de modéliser les deux termes d'intérêt par un modèle linéaire généralisé (GLM). Cette méthode est détaillée en tant que « fiche rappel » dans l'annexe 13 de ce mémoire. Deux possibilités existent en fait :

- Appliquer la méthode du GLM pour modéliser le coût moyen d'une part et la fréquence d'autre part (2 modèles) ;
- Appliquer la méthode du GLM sur la consommation (1 modèle).

L'utilisation de deux modèles permet de capter l'influence parfois contraire sur les variables à expliquer (coût et fréquence) des variables explicatives pourtant partagées (classe d'âge, région et sexe si inclus). Le pilotage du tarif en est par ailleurs meilleur. La première possibilité semble donc être celle de préférence. Cependant, l'utilisation de deux modélisations signifie aussi une exposition double au risque de modèle et notamment, à un mauvais choix de paramétrage (par exemple, choisir des lois non adaptées aux données). En revanche, il ne faut pas oublier l'hypothèse d'indépendance de la fréquence et du coût moyen sous-tendant le GLM.

Cette indépendance n'est pas vérifiée pour tous les segments de tarification retenus et il faut donc la confirmer via un test de corrélation. Si le test est en faveur d'une indépendance, il sera choisi la modélisation séparée du coût et de la fréquence (méthode nommée « **GLM fréquence x coût moyen** »). Dans le cas contraire, ce sera la seconde modélisation dite « **GLM consommation** » qui sera utilisée.

Il existe différents tests de corrélation. Soient X la variable aléatoire quantitative désignant la fréquence et Y la variable aléatoire quantitative associée au coût moyen. Dire que X et Y sont corrélées revient dans le contexte de tarification à dire que l'attribution d'une modalité à une des variables explicatives (de l'une des deux variables à expliquer) engendrerait des effets sur l'autre variable, i.e. l'évolution en valeur de l'une affecterait l'évolution en valeur de l'autre.

Parmi les nombreuses mesures de corrélation entre deux variables, la plus basique et accessible reste celle du **coefficient de corrélation linéaire** (bien que non utilisée dans le cadre de ces travaux). Ce coefficient prend une valeur dans l'intervalle $[-1 ; +1]$. Plus le coefficient se rapproche des valeurs extrêmes de l'intervalle et plus la corrélation linéaire entre les deux variables est forte. Le terme « linéaire » impose la forme de dépendance des deux variables : en effet, un coefficient égal à $+1$ (respectivement -1) signifie que l'une des deux variables est une fonction affine croissante (respectivement décroissante) de l'autre. Ce sont des cas de corrélation parfaite. Les valeurs intermédiaires illustrent un degré de dépendance intermédiaire. Dans le cas où le coefficient est nul, cela ne signifie pas pour autant l'indépendance linéaire : être nul est seulement une condition nécessaire et non suffisante de l'indépendance entre deux variables. Ce coefficient est néanmoins sensible aux valeurs aberrantes (selon les types de travaux, cela est un inconvénient ou non) et n'est pas robuste.

Annexe 14
Coefficient de
corrélation

Ainsi, il a été considéré les coefficients de **corrélation de Kendall, Spearman et Pearson** (ce dernier étant la version empirique du coefficient de corrélation linéaire). En effet, le triple test permet de diminuer le risque de faire un tarif via deux modèles alors que les hypothèses ne sont pas validées.

Annexe 14
Pearson, Kendall,
Spearman

De manière similaire au coefficient de corrélation, les valeurs intermédiaires prises par ces trois coefficients sont sujettes à l'interprétation : « à partir de quelle valeur intermédiaire pouvons-nous considérer qu'il y a une dépendance significativement suffisante pour pouvoir exclure la modélisation en deux modèles ? ». Il n'y a malheureusement pas de réponse stricte à cette interrogation. Fatemeh ABDOLLAHI (*Tarifification d'une complémentaire santé à destination des seniors, modulaire par poste de garanties et l'impact sur la solvabilité*, 2017, p.57) suggère de prendre une limite à 0,20. Un seuil un peu moins strict à 0,25 a été retenu comme il était souhaité de l'atteindre pour tous les tests. Ainsi, pour pouvoir considérer avec plus de sérénité une modélisation séparée de la fréquence et du coût par GLM, il conviendrait de respecter la condition suivante : pour les trois tests de corrélation (liés aux trois mesures de dépendance utilisées), le coefficient de corrélation obtenu est inférieur à 0,25. Les tests sont effectués sous *R* par la fonction *cor.test*.

3.1.1.4. Méthode 2 de tarification : la statistique directe

L'équation (2) permet aussi de bifurquer vers une deuxième méthode de détermination de la prime pure. Elle sera appelée la méthode « **statistique directe** ». Elle consiste en le calcul de la consommation en appliquant directement, sans modélisation quelconque, l'équation (2).

Cette méthode sera appliquée dans deux cas :

- Cas 1 : le segment de santé considéré n'est pas un segment de tarification retenu en sous-partie 2.2.3. (c'est le cas par exemple de l'implantologie) ;
- Cas 2 : la première méthode de tarification envisagée, à savoir le GLM, n'aboutit pas (c'est le cas par exemple où parmi les lois de probabilité testées, aucune ne semble adaptée aux données et les résultats qui en résultent ne sont pas satisfaisants).

Cas 1 : prime non concernée par l'extension de tarif

Le calcul de la prime pure est effectué sur l'ensemble des données liées au segment de santé non retenu en tant que segment de tarification (sous-entendu, par méthode GLM) sur la base des données de VirtuaMut', pour la garantie A ou B. Elle constitue en quelque sorte une « surprime » pour cette garantie (comme ce qu'il est parfois fait pour le cas des sinistres de type catastrophe naturelle lors d'une tarification auto) et s'ajoute directement à la prime pure totale. Les calculs se font par classe d'âge.

À préciser que dans ce cas-ci, aucune extension de tarif ne sera effectuée. En effet, ces cas concernent parfois des prestations qui ne sont pas remboursées par la Sécurité Sociale et qui ne sont donc pas présentes dans la base de l'Open DAMIR (*a contrario* de la base de VirtuaMut') ; ou encore, des

prestations dont la déduction du montant qu’aurait remboursé la mutuelle avait été trop complexe (grilles tarifaires ou classification peu claires). Autrement dit, la prime pure associée à chacun de ces segments restera identique d’une région à l’autre.

Cas 2 : prime concernée par l’extension de tarif

Sur les jeux de données de VirtuaMut’, les calculs se font par classe d’âge pour la simple raison que c’est la seule variable explicative disponible. En effet, d’une part, la « *Gender Directive* », entrée en vigueur depuis fin 2012, exclut la segmentation des tarifs selon le genre d’une personne. Cela se traduit dans les travaux par l’exclusion de la variable **SEXE** en tant que variable discriminante. D’autre part, nous nous concentrons uniquement sur les données concernant les individus de la région D pour élaborer un tarif exclusif à cette région particulière.

Sur les données de l’Open DAMIR, les calculs se font selon deux axes : la classe d’âge et la région (car cette fois-ci, les données n’étant pas concentrées sur une zone géographique particulière, il y a présence de deux variables discriminantes).

Cette méthode présente cependant une limite : elle ne permet ni le pilotage du tarif ni sa malléabilité (dans le sens où pour modifier les primes pures issues de cette méthode, il faut changer de données).

3.1.2. Les méthodes de tarification retenue par segment de santé

Voici ci-dessous un tableau récapitulatif, pour chaque segment de frais de soin, la méthodologie retenue. Des exemples d’application et de démarche par méthode seront présentés dans la section suivante.

Familles d’actes	Sous-familles d’actes	Méthode de tarification pour la garantie A	Méthode de tarification pour la garantie B	Loi(s) utilisée(s)	Extension du tarif (Oui = X)
Honoraires médicaux	Consultations/visites	GLM consommation	GLM consommation	Tweedie	X
	Actes médicaux (techniques)	GLM fréquence x coût moyen	GLM consommation	Binomiale négative x Gamma // Tweedie	X
	Autres honoraires médicaux	Non considéré*	Non considéré*		
Honoraires paramédicaux	Auxiliaires médicaux (kiné, infirmiers, ...)	Statistique directe (cas 2)	Statistique directe (cas 2)		X
Analyse et examen de laboratoire	Analyse médicale et examens laboratoire	GLM consommation	GLM consommation	Tweedie	X
Imagerie médicale	Actes d’imagerie, de radiologie et ostéodensitométrie	GLM consommation	GLM consommation	Tweedie	X
Transport, ambulances	Transport	Statistique directe (cas 1)	Statistique directe (cas 1)		
Pharmacie	Petit appareillage				
	Pharmacie	Statistique directe (cas 2)	Statistique directe (cas 2)		X
	Vaccins anti-grippe				
Hospitalisation	Honoraires et actes chirurgicaux	Non considéré*	GLM fréquence x coût moyen	Quasipoisson x Gamma	X
	Forfait journalier	Statistique directe (cas 1)	Statistique directe (cas 1)		
	Frais de séjour				
	Chambre particulière Grand appareillage	Statistique directe (cas 2)	Statistique directe (cas 2)		X
	Lit accompagnant				
Autres - hospitalisation	Statistique directe (cas 1)	Statistique directe (cas 1)			
Dentaire	Prothèse	GLM consommation	GLM consommation	Tweedie	X
	Soins dentaires	GLM fréquence x coût moyen	GLM fréquence x coût moyen	Binomiale négative x Gamma	X
	Parodontologie	Statistique directe (cas 1)	Statistique directe (cas 1)		
	Implantologie	Statistique directe (cas 1)	Statistique directe (cas 1)		
	Orthodontie	Statistique directe (cas 1)	Statistique directe (cas 1)		
	Autres - dentaire	Statistique directe (cas 1)	Statistique directe (cas 1)		
Optique	Monture	GLM consommation	GLM fréquence x coût moyen	Tweedie // Binomiale négative x Gamma	X
	Verres	GLM consommation	GLM fréquence x coût moyen	Tweedie // Binomiale négative x Gamma	X
	Chirurgie oeil	Non considéré*	Non considéré*		
	Lentilles	Statistique directe (cas 1)	Statistique directe (cas 1)		
	Autres - Optique	Non considéré*	Non considéré*		
Audio	Audioprothèse (+Pile, accessoire)	Statistique directe (cas 2)	Statistique directe (cas 2)		
	Autres - audio	Non considéré*	Non considéré*		
Cure thermique	Cure thermique	Statistique directe (cas 1)	Statistique directe (cas 1)		
Médecine douce	Ostéopathie, diététique, chiropractie, ...	Statistique directe (cas 1)	Statistique directe (cas 1)		
Prévention	Prévention	Non considéré*	Non considéré*		
Prestations supplémentaires	Maternité	Non considéré*	Non considéré*		
	Autres (aides, inclassable, ...)	Non considéré*	Non considéré*		

* Considéré.e.s comme non remboursé.e.s par la mutuelle dans les modélisations (pour cause de non-présence de codes actes ou autres).

Tableau 13 : Méthode de tarification par segment

Pour chaque segment, la méthodologie retenue est celle compatible à la fois avec les données de l’Open DAMIR et les données de VirtuaMut’ (pour chaque garantie étudiée).

Par exemple, pour la garantie A et la sous-famille des prothèses dentaires, la méthode indiquée est « GLM consommation ». Cela signifie qu'un GLM avec comme variable à expliquer la consommation, le tout restreint à la famille des prothèses dentaires, a été réalisé sur les données de la garantie A de VirtuaMut' et un second GLM sur les données de l'Open DAMIR qui sont associées à la garantie A.

Il est nécessaire de faire deux modèles à chaque fois pour pouvoir les comparer et *in fine*, aboutir au second objectif opérationnel qui est l'extension des tarifs au territoire national. Si la méthode GLM consommation a été retenue, cela signifie aussi que pour au moins un des deux jeux de données (VirtuaMut' ou Open DAMIR), il a été observé pour au moins un des trois tests de corrélation (Kendall, Spearman, Pearson), une valeur de coefficient supérieure à 0,25.

Section 2 – Tarification

3.2.1. Quelques réflexions préliminaires

Il convient encore une fois de se poser un certain nombre de questions, non plus sur les données, mais sur la méthode choisie et les différentes hypothèses à prendre.

D'une part, en ce qui concerne la séparation entre les sinistres attritionnels et les sinistres larges (voire de type catastrophe), il a été considéré qu'au regard des études sur les valeurs aberrantes et du domaine de l'assurance santé, cette séparation n'était pas nécessaire. Tous les sinistres ont été considérés comme étant attritionnels. Il n'a pas été déterminé de seuil (que ce soit arbitrairement, empiriquement ou via la théorie des valeurs extrêmes) à partir duquel un remboursement pouvait être associé à un sinistre large.

D'autre part, en ce qui concerne les montants as-if, il a été considéré les montants tels quels afin de pouvoir effectuer des comparaisons avec des tarifs déjà pratiqués, bien qu'un tarif 2021 serait *a priori* le tarif à réaliser. Il aurait été par exemple pertinent de déformer les montants en prenant en considération la dérive annuelle et naturelle des prix (inflation) ainsi que l'impact du 100 % santé. Cela sera abordé un peu plus en détail en partie 3.4.5. La priorité était avant tout de se concentrer sur la méthode globale.

Pour rappel, dans le cas d'une tarification via un GLM :

- Pour les données de VirtuaMut', la seule variable explicative retenue est la classe d'âge. La région ne constitue pas une variable explicative puisque les modèles sont réalisés avec les données sur la région D uniquement. La variable SEXE n'est pas retenue non plus à cause de la *Gender Directive*.
- Pour les données de l'Open DAMIR, les deux variables explicatives sont la classe d'âge et la région. La variable SEXE est exclue pour la même raison que précédemment.

Ainsi, les types de modèles GLM rencontrés sont les suivants :

- Pour VirtuaMut' :
 - Pour la fréquence (et la consommation) :

$$g(E(Y)) = \beta_0 + \beta_1 X_{classe\ d'âge=0} + \dots + \beta_8 X_{classe\ d'âge=80} + offset(Exposition)$$

où Y est la variable associée à la quantité d'actes (ou au coût total) et Exposition à l'exposition totale.

- Pour le coût moyen :

$$g(E(Y)) = \beta_0 + \beta_1 X_{classe\ d'âge=0} + \dots + \beta_8 X_{classe\ d'âge=80}$$

où Y est la variable associée au coût moyen.

Le but du GLM sera d'estimer les différents coefficients β_0, \dots, β_8 . Un point d'attention tout de même pour l'équation précédente qui est écrite par souci de généralité mais qui n'est pas tout à fait exacte telle quelle : un des β_1, \dots, β_8 n'apparaît pas dans le modèle car il est associé à une modalité de l'individu de référence. Pour cet individu, $X_{classe\ d'âge=classe\ de\ référence}$ vaut 1 et le coefficient associé est donc confondu avec β_0 . Cet individu de référence change selon le segment de tarification étudié. Par ailleurs, il sera pris le même individu de référence pour l'Open DAMIR et pour VirtuaMut' pour une garantie donnée et pour un segment de tarification donné pour une question de cohérence.

- Pour l'Open DAMIR :
 - Pour la fréquence (et la consommation) :

$$g(E(Y')) = \beta'_0 + \beta'_1 X_{classe\ d'âge=0} + \dots + \beta'_8 X_{classe\ d'âge=80} + \beta'_9 X_{région=D} + \beta'_{10} X_{région=01} + \dots + \beta'_{21} X_{région=012} + offset(Exposition')$$

où Y' est la variable associée à la quantité d'actes (ou au coût total) et Exposition' à l'exposition totale.

- Pour le coût moyen :

$$g(E(Y')) = \beta'_0 + \beta'_1 X_{classe\ d'age=0} + \dots + \beta'_8 X_{classe\ d'age=80} + \beta'_9 X_{region=D} \\ + \beta'_{10} X_{region=O1} + \dots + \beta'_{21} X_{region=O12}$$

où Y' est la variable associée au coût moyen.

Le but du GLM sera d'estimer les différents coefficients $\beta'_0, \dots, \beta'_{21}$. De même, l'équation précédente est écrite par souci de généralité mais n'est pas tout à fait exacte : un des β_1, \dots, β_8 et un des $\beta_9, \dots, \beta_{21}$ n'apparaissent pas dans le modèle car ils sont associés à des modalités de l'individu de référence.

Pour des raisons de commodités et de cohérence et afin de pouvoir réaliser l'extension de tarifs, la fonction lien log a été considérée pour tous les GLM. Cette dernière était compatible avec les lois testées et choisies. Cette fonction lien est par ailleurs souvent appréciée pour de tels travaux car elle assure des valeurs prédites positives et elle permet de construire des modèles multiplicatifs (qui modélisent donc un effet proportionnel de chaque facteur de risque sur la variable à expliquer) [17].

Enfin, il convient de préciser qu'à ce stade, les différents jeux de données ont été subdivisés par segments de tarification. Dans la suite, tous les traitements effectués concerneront donc les données liées à une garantie en particulière et à un segment de tarification spécifique.

3.2.2. Un exemple de cas pour la fréquence

L'idée générale est d'effectuer une tarification par une approche fréquence x coût moyen via un GLM. La partie sur la fréquence est abordée ici via un exemple illustratif. L'autre partie complémentaire sur le coût sera abordée dans la sous-section suivante. La démarche sera la même (dans son entièreté et sans sauter d'étape pour une question de rigueur) pour tous les segments de tarification dont la méthode retenue est un « GLM fréquence x coût moyen ».

Le segment retenu à titre d'exemple est celui des actes médicaux techniques de la garantie A associée à la base de données de l'Open DAMIR. Il sera possible, sous demande, de fournir les résultats pour les autres segments.

Pour ce segment en particulier, les tests de corrélation ont donné les résultats suivants :

	Valeur du coefficient
Pearson	-0,09
Kendall	-0,07
Spearman	-0,10

Tableau 14 : Résultats de tests de corrélation – Actes techniques médicaux (garantie A)

Pour l'ensemble des trois tests, une dépendance négative faible est observée, inférieure au seuil (en valeur absolue) de 0,25 imposé. Il est donc cohérent de supposer la non-corrélation entre la variable de fréquence et de coût et de procéder à un modèle GLM fréquence x coût moyen.

Remarque : la dépendance négative signifierait que plus souvent un individu consomme des actes techniques médicaux, et plus bas est le coût moyen de ses actes.

3.2.2.1. Détermination de la loi de fréquence

Au vu du nombre important de modèles à réaliser (des centaines... sans possibilité d'automatisation puisqu'il faut analyser, traiter et optimiser chaque modèle un à un), il a été décidé de ne retenir comme lois possibles d'adéquation aux données de la fréquence que celles les plus courantes, à savoir : la loi de Poisson, la loi binomiale négative et la loi de Poisson surdispersée. Cela exclut donc les lois mélanges par exemple, qui peuvent être pertinentes dans le cas où il est observé plusieurs modes dans les histogrammes empiriques générés ou les graphiques de densité de loi. Ces dernières sont cependant plus complexes et plus subtiles à calibrer.

Annexe 16

La première étape est de réaliser une rapide étude statistique pour statuer sur le caractère de surdispersion des données (i.e. quand la variance de la variable à expliquer est supérieure à son espérance) :

- En absence de surdispersion, la loi de Poisson est retenue ;
- En présence de surdispersion, il y a comparaison entre le modèle se basant sur l'hypothèse d'une loi binomiale négative ou celui se basant sur l'hypothèse d'une loi de Poisson surdispersée.

Il est aussi important de préciser que dans le cas d'un modèle de fréquence, la variable à expliquer est en réalité la quantité d'actes. Le nombre de bénéficiaires est traité en tant qu'*offset* du modèle.

Pour notre exemple, en ce qui concerne les quantités d'actes, la moyenne s'élève à $4,41 \times 10^5$ contre une variance à $9,10 \times 10^{10}$. La variance est supérieure à l'espérance, il y a donc présence d'un phénomène de surdispersion : il sera donc retenu soit la loi binomiale négative, soit la loi de Poisson surdispersée. Par ailleurs, les quantités si élevées s'expliquent par le fait qu'elles soient agrégées, elles sont donc difficilement interprétables sous cette forme.

Il est possible de déduire qu'un individu ayant adhéré à la garantie A consomme en moyenne 1,37 actes techniques médicaux par an en calculant la fréquence empirique (la somme des quantités d'actes divisée par la somme des expositions totales). Cela semble cohérent avec l'étude descriptive en sous-partie 2.3.2.2. qui présente, pour la garantie A, une dépense moyenne par individu par an de 43 € pour un acte technique médical coûtant en moyenne 44 € l'unité.

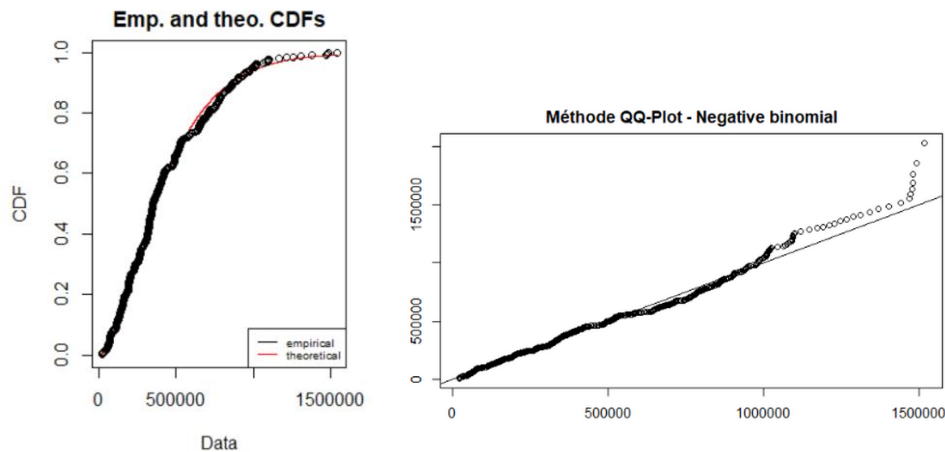
La deuxième étape est de vérifier l'adéquation des données aux lois candidates. Pour cela, au préalable, il faut supposer que les observations suivent une loi en particulière, puis estimer ses paramètres (méthodes des moments ou méthode de maximum de vraisemblance via la fonction *fit.dist* de la librairie *fitdistrplus* de R). Par la suite, plusieurs méthodes sont utilisées pour statuer sur l'adéquation des données avec la loi candidate paramétrée :

- Les méthodes graphiques pour une évaluation visuelle : histogramme des données ou graphique de la densité empirique, comparaison entre la fonction de répartition empirique et la fonction de répartition théorique, QQ-plot.
- Les méthodes statistiques pour une évaluation chiffrée : test d'adéquation du Khi-deux.

Annexe 15.5

Malheureusement, dans notre cas précis, une condition du test d'adéquation du Khi-deux n'était pas satisfaite (à savoir que le nombre d'observations multiplié par les probabilités d'appartenance doivent être toutes supérieures à 5). Les tests n'étaient donc pas concluants et c'est donc avant tout les méthodes graphiques qui ont été utilisées pour porter un jugement.

Dans notre exemple :



Graphique 15 : Tests d'adéquation de loi de fréquence – Actes techniques médicaux (garantie A)

Le graphique de gauche présente la comparaison entre la fonction de répartition empirique de la variable à expliquer et la fonction de répartition théorique de la loi candidate (ici, une loi binomiale négative). Le résultat est satisfaisant : les points empiriques (noirs) sont très proches des points théoriques (rouges). Le graphique de droite est le QQ-plot associé et il affiche aussi une superposition satisfaisante des points sur la première bissectrice (la droite d'équation $y = x$) excepté vers la fin. Cela signifie que la loi binomiale négative n'est pas parfaitement adaptée au niveau des queues lourdes de distribution. Cela reste cependant négligeable comme cela ne concerne que quelques points.

Si plusieurs lois restent candidates après les tests d'adéquation, une comparaison est effectuée sur plusieurs considérations afin de ne retenir qu'une loi unique : le critère BIC, la déviance ou la comparaison de la somme des écarts entre la fréquence calculée directement sur les données et la fréquence modélisée par modalité de variables explicatives. Ce sera la loi binomiale négative qui sera finalement retenue.

Annexe 15.3
Critère BIC

Annexe 15.2
Déviance

3.2.2.2. Multicolinéarité ?

Avant d'effectuer un GLM, il est nécessaire d'exclure la présence de variables explicatives fortement corrélées linéairement (par exemple, si deux variables explicatives donnent des informations redondantes et qu'il n'est pas possible de distinguer leur influence individuelle sur la variable à expliquer). L'idée est que les modèles de régression présentent des facteurs de risque qui doivent être indépendants deux à deux, au risque sinon d'aboutir à des résultats non fiables.

Il s'agit aussi dans cette sous-partie de constater la corrélation forte entre la variable de qualité de bénéficiaire et celle des classes d'âge, telle qu'hypothétiser en sous-partie 2.3.1.4. de par la construction de la variable QLT.

Pour cela, plusieurs outils existent : le VIF (Variation Inflation Vector), le V de Cramer, le coefficient de corrélation, ... Les variables explicatives étant toutes de nature qualitative, il a été dans un premier temps considéré le VIF (ou plutôt, sa version généralisée, notée GVIF).

Annexe 15.1
VIF

Tout comme pour les coefficients de corrélation de Kendall, Spearman et Pearson, le GVIF ne présente pas de seuil objectif à partir duquel un problème de multicolinéarité est avéré. Par ailleurs, cet indicateur n'indique pas clairement quelle variable est corrélée à quelle autre (sauf éventuellement dans notre cas où les modèles présentent trop peu de variables explicatives pour qu'il y ait ambiguïté).

Il existe cependant une *rule of thumb* pour l'interprétation du VIF [18] (cf. l'annexe 15.1 pour l'équivalence de l'interprétation dans le cas du GVIF) :

- Un VIF égal à 1 indique qu'il n'y a pas de corrélation entre une variable explicative donnée et n'importe laquelle des autres variables explicatives du modèle ;
- Un VIF compris entre 1 et 5 indique qu'il existe une corrélation modérée entre une variable explicative donnée et n'importe laquelle des autres variables explicatives du modèle. Pour la plupart du temps, cela n'est pas suffisamment significatif pour que les résultats soient impactés ;
- Un VIF supérieur à 5 indique qu'il y a potentiellement une forte corrélation entre une variable explicative donnée et n'importe laquelle des autres variables explicatives du modèle. Dans ce cas-ci, les coefficients β_i estimés ainsi que les p_{value} des tests de significativité issus de la modélisation ne sont pas fiables.

En utilisant ce test, il a souvent été observé une interprétation du GVIF équivalente à un VIF supérieur à 5 pour les variables CLASSE_AGE et QLT pour les données de VirtuaMut' (et supérieur à 1 pour les données de l'Open DAMIR). Elles seraient donc *a priori* fortement corrélées entre elles.

Dans le cas de notre exemple :

	$\left(\frac{1}{GVIF \cdot 2DF}\right)^2$
CLASSE_AGE	1,8
SEXE	1,0
QLT	2,1
REGION	1,0

Tableau 15 : Test du GVIF – Actes techniques médicaux (garantie A)

Les variables CLASSE_AGE et QLT ayant un GVIF supérieur à 1, elles sont suspectes.

Cela est cependant encore plus visible sur les données de la mutuelle. Par exemple, avec la sous-famille de consultations et visites pour la garantie A :

	$\left(\frac{1}{GVIF \cdot 2DF}\right)^2$
CLASSE_AGE	20,1
SEXE	1,1
QLT	18,3

Tableau 16 : Test du GVIF – Consultations et visites (garantie A)

Par ailleurs, pour renforcer cette conclusion de corrélation effective, une analyse des corrélations entre facteurs de risque qualitatifs a été réalisée via un test de Khi-deux. Si la p_{value} à l'issue du test est inférieure à 5% (seuil choisi), il y a rejet de l'hypothèse nulle qui est l'indépendance entre deux variables.

Annexe 15.4
Test
d'indépendance
du Khi-Deux

À titre de précision, le test de Khi-deux ne permet pas de quantifier le degré de dépendance. Mais prendre l'hypothèse qu'une p_{value} très proche de 0 (elle est à 0 quand il y a corrélation parfaite comme la corrélation d'une variable avec elle-même) est signe de forte corrélation nous a paru justifié.

Dans notre exemple, le tableau ci-après permet de constater que le couple (CLASSE_AGE, QLT) affiche une p_{value} égale à 0. Il est fort possible que la p_{value} ne soit pas exactement nulle, mais tellement proche de l'être que R l'affiche comme telle.

	CLASSE_AGE	SEXE	REGION	QLT
CLASSE_AGE	0	> 0,05	> 0,05	0
SEXE	> 0,05	0	> 0,05	> 0,05
REGION	> 0,05	> 0,05	0	> 0,05
QLT	0	> 0,05	> 0,05	0

Tableau 17 : Matrice des corrélations entre variables explicatives – Actes techniques médicaux (A)

Ainsi, la variable QLT ne sera pas considérée dans les modèles.

3.2.2.3. Définition d'un individu de référence

Avant la définition de l'individu de référence, il est d'usage de faire une étude descriptive (univariée ou bivariée) dans le but de réaliser des pré-regroupements de modalités, d'étudier la possibilité de discrétiser les variables qualitatives ou de créer des classes pour les variables quantitatives. Cela permet de réduire la complexité des modèles en réduisant le nombre de coefficients à estimer. Cependant, une étude statistique sur les dépenses avait déjà été réalisée (cf. partie 2.3.2) et par ailleurs, au regard du nombre de variables explicatives utilisées, l'étude n'était pas pertinente. Par ailleurs, pour la classe d'âge, la distinction en classes est souhaitée, de même pour la région.

En ce qui concerne l'individu de référence, il a été défini en prenant les caractéristiques du groupe d'individus ayant consommé le plus fréquemment (i.e. qui présente la plus grande quantité d'actes totale sur le segment de tarification considéré). L'individu de référence a été déterminé dans un premier temps pour chaque segment de tarification et garantie sur les données de VirtuaMut'. Une fois défini, il restera le même par segment de tarification et par garantie pour les données de l'Open DAMIR ainsi que pour le modèle de coût (ou de consommation). La région D a été fixée comme modalité de référence pour la variable REGION par souci de cohérence de modèles dans la phase suivante des travaux.

Dans le cas de notre exemple, l'individu de référence est une personne de sexe féminin, âgée entre 30 et 39 (classe d'âge 30) et habitant en région D.

3.2.2.4. Training set et test set

L'étape suivante est la séparation des données en base d'apprentissage (*training set*) et base de test (*test set*) de manière aléatoire (échantillonnage). Le seuil retenu est 85 % : cela signifie que 85 % des données iront dans la base d'apprentissage et les 15 % restantes dans la base de test. Cela est effectué pour chaque segment de tarification et pour chaque garantie.

La base d'apprentissage contient les données sur lesquelles les modèles sont construits alors que la base de test permet ensuite de valider les modèles et d'évaluer leur pouvoir de prédiction en les testant sur des données dont il est déjà su par avance le « résultat à trouver ».

3.2.2.5. Modèle GLM et optimisation

L'étape suivante est la réalisation du GLM via R.

Dans le cas où il y avait plusieurs lois candidates d'adéquation aux données, un GLM a été fait avec chacune des lois candidates afin de pouvoir effectuer les comparaisons précitées. Au final, un seul modèle est retenu. Le choix de loi est réalisé en premier lieu sur les données de VirtuaMut'. Par souci de cohérence avec le second objectif d'extension, il a été décidé de garder ce même choix de loi dans le cas de l'Open DAMIR (et toujours par segment de tarification et par garantie). Cette décision a parfois

conduit à un changement de méthode par inadéquation de la loi avec les données de l'Open DAMIR : passage de la méthode GLM fréquence x coût moyen à GLM consommation voire statistique directe. En effet, la méthode GLM consommation est en fait un choix possible à tout temps (ne pas satisfaire aux trois tests de corrélation signifie que la méthode GLM fréquence x coût moyen ne peut pas être prise mais satisfaire aux trois tests ne signifie pas que la méthode GLM consommation soit à exclure).

Avec notre exemple, cela donne :

```
Call:
glm.nb(formula = QTE_ACTE ~ CLASSE_AGE + REGION + offset(log(EXPO_TOTALE)),
       data = data_agreg_AM_training, link = log, init.theta = 12.06386427)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.47645 -0.75508 -0.00306  0.60520  2.07087

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.041912  0.068804   0.609   0.5424
CLASSE_AGE0 -0.621225  0.061684 -10.071 < 2e-16 ***
CLASSE_AGE20 -0.355532  0.061159  -5.813 6.13e-09 ***
CLASSE_AGE40  0.281999  0.063204   4.462 8.13e-06 ***
CLASSE_AGE50  0.511372  0.061661   8.293 < 2e-16 ***
CLASSE_AGE60  0.813758  0.062399  13.041 < 2e-16 ***
CLASSE_AGE70  1.134327  0.060933  18.616 < 2e-16 ***
CLASSE_AGE80  1.050878  0.062341  16.857 < 2e-16 ***
REGIONO1     -0.587025  0.077777  -7.548 4.44e-14 ***
REGIONO2     -0.085202  0.078580  -1.084  0.2782
REGIONO3     -0.149307  0.079325  -1.882  0.0598 .
REGIONO4     -0.124892  0.077013  -1.622  0.1049
REGIONO5     -0.110147  0.078472  -1.404  0.1604
REGIONO6     -0.131968  0.079257  -1.665  0.0959 .
REGIONO7     -0.120291  0.075740  -1.588  0.1122
REGIONO8     -0.114726  0.077079  -1.488  0.1366
REGIONO9     -0.056830  0.080239  -0.708  0.4788
REGIONO10    0.002499  0.075740   0.033  0.9737
REGIONO11   -0.014837  0.076417  -0.194  0.8461
REGIONO12    0.148698  0.077769   1.912  0.0559 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(12.0639) family taken to be 1)
Null deviance: 1865.45 on 352 degrees of freedom
Residual deviance: 357.88 on 333 degrees of freedom
```

Les modalités des variables explicatives ne sont pas toutes significatives (comme visible par des p value ou $\Pr(>|z|)$ supérieures à 0,05, seuil retenu dans les travaux). Par ailleurs, pour rappel, l'interprétation des coefficients estimés n'a avant tout du sens que par rapport à l'individu de référence. Par exemple, ici, un individu de classe d'âge 70 consomme en moyenne $e^{1,134327} \approx 3,11$ fois plus souvent que l'individu de référence (de classe d'âge 30).

En ce qui concerne la qualité d'ajustement des modèles, il est d'ores et déjà possible de l'évaluer via la déviance nulle et la déviance résiduelle. Leurs valeurs indiquent que :

- L'ajout de 19 modalités de variables explicatives fait diminuer la déviance de 1 507,57 (le modèle estimé s'ajuste mieux aux données que le modèle « nul ») ;
- Le quantile d'une loi du Khi-deux d'ordre 0,95 avec 333 degrés de liberté est de 376,56. Il est supérieur à la déviance résiduelle (i.e. déviance réduite du modèle). De ce fait, la qualité d'ajustement du modèle est satisfaisante.

Pour l'Open DAMIR où la variable explicative de REGION intervient, il a été souvent observé une significativité plus accrue de la variable CLASSE_AGE comparée à celle de REGION (cela est visible par le fait que les coefficients associés aux modalités de classes d'âge prennent des valeurs plus élevées que les coefficients associés aux modalités géographiques). Cela veut aussi dire que la première est plus discriminante, conformément à nos attentes. Ainsi, la variable REGION agit plutôt comme un paramètre d'affinement de tarif comparée à celle sur l'âge : la variabilité du tarif du fait de la région serait limitée.

Annexe 15.7
Test significativité

Annexe 15.2
Déviance

Ensuite, il convient d'optimiser le modèle retenu. Étant donné que les variables explicatives sont uniquement de nature qualitative et que ces dernières sont peu nombreuses, l'optimisation consiste uniquement en des regroupements de modalités (supprimer des variables revient dans ce cas-ci à supprimer toutes les modalités d'une variable explicative et dans le cas des modèles sur les données de VirtuaMut', à se retrouver sans variables explicatives).

L'objectif est d'aboutir à un modèle où toutes les modalités sont significatives. Pour cela, le regroupement a été réalisé en privilégiant deux considérations :

- Regrouper une modalité non significative avec une modalité « proche » en termes de sens (par exemple, il fait plus « sens » de regrouper les individus de la classe d'âge 30 avec ceux de la classe d'âge d'au-dessus, que regrouper ceux-ci avec la classe d'âge 80) ;
- Regrouper les modalités ayant des p values proches (donc, une significativité proche).

Ainsi, ce sont donc des considérations plutôt qualitatives qui ont été retenues. De ce fait, l'optimisation reste une étape très subjective et deux personnes la réalisant pourraient aboutir à des modèles valides présentant cependant des coefficients différents par optimisation divergente. Il aurait été idéal de tester toutes les possibilités d'optimisation conduisant à des modèles aux modalités toutes significatives puis de les comparer ou encore, de réaliser des tests de Wald qui permettraient de se conforter ou non dans les regroupements faits, mais cela serait bien trop long pour la plus-value finale que cela apporte (à savoir que cette « plus-value » fera l'objet d'une étude à part). Cette remarque prendra son importance dans l'extension de tarif.

Dans le cas de l'exemple, les modalités suivantes ont été regroupées :

- La modalité de région O1 avec la O2 ;
- La modalité de région O5 avec la O8 ;
- Les modalités de région O9, O10 et O11 avec la modalité D.

Le modèle après optimisation est le suivant :

```
Call:
glm.nb(formula = QTE_ACTE ~ CLASSE_AGE + REGION + offset(log(EXPO_TOTALE)),
       data = data_agreg_AM_training, link = log, init.theta = 10.86303746)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.07097 -0.70163 -0.00253  0.58023  1.90916

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03994    0.05157   0.774 0.438727
CLASSE_AGE0 -0.62734    0.06485  -9.674 < 2e-16 ***
CLASSE_AGE20 -0.35430    0.06441  -5.500 3.79e-08 ***
CLASSE_AGE40  0.25600    0.06635   3.858 0.000114 ***
CLASSE_AGE50  0.49591    0.06490   7.641 2.16e-14 ***
CLASSE_AGE60  0.78491    0.06553  11.979 < 2e-16 ***
CLASSE_AGE70  1.11286    0.06407  17.370 < 2e-16 ***
CLASSE_AGE80  1.03525    0.06551  15.804 < 2e-16 ***
REGIONO1-02 -0.29323    0.05075  -5.779 7.54e-09 ***
REGIONO3    -0.13184    0.06725  -1.960 0.049944 *
REGIONO4    -0.11031    0.06424  -1.717 0.085966 .
REGIONO5-08 -0.09680    0.05037  -1.922 0.054622 .
REGIONO6    -0.11431    0.06725  -1.700 0.089161 .
REGIONO7    -0.10529    0.06245  -1.686 0.091811 .
REGIONO12   0.16482    0.06522   2.527 0.011498 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(10.863) family taken to be 1)
Null deviance: 1679.77 on 352 degrees of freedom
Residual deviance: 358.42 on 338 degrees of freedom
```

Toutes les modalités sont ici significatives. Ce modèle est retenu.

3.2.2.6. Validation

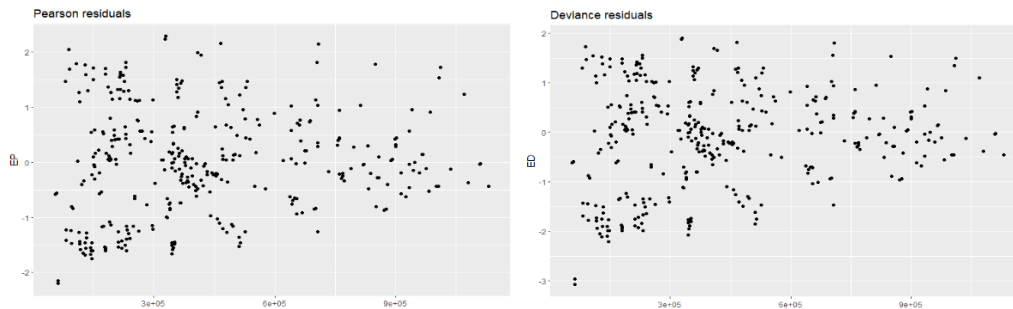
Enfin, il convient de valider le modèle retenu et optimisé. Pour cela, trois méthodes ont été identifiées :

- Une méthode statistique qui repose sur le calcul de la déviance du modèle ;
- Une méthode graphique avec les résidus de Pearson ;
- Une seconde méthode graphique avec les résidus de déviance.

Annexe 15.8
Résidus

Pour ce qui est de la première méthode, la déviance résiduelle est inférieure 381,87 (quantile d'une loi de Khi-deux d'ordre 0,95 avec 338 degrés de liberté). Le modèle est donc valide.

En ce qui concerne les méthodes graphiques qui complètent la première méthode :



Graphique 16 : Résidus de Pearson et de déviance - Fréquence – Actes techniques médicaux (A)

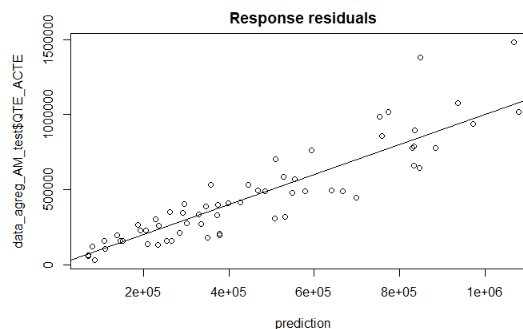
Le graphique de gauche présente des résidus de Pearson contenus dans l'intervalle $[-2 ; 2]$ pour la très grande majorité des points et centrés en 0. De plus, il n'y a pas de tendance particulière. Ce test est aussi validé. Le graphique de droite présente des résidus de déviance qui dépassent en valeur absolue 1, ce qui signifie qu'ils existent des observations participant au mauvais ajustement du modèle.

3.2.2.7. Prédiction

La dernière étape est l'utilisation de la base test en tant qu'*input* du modèle retenu afin de pouvoir évaluer son pouvoir prédictif et la justesse du modèle en cas de données « nouvelles ».

Là encore, plusieurs méthodes permettent de statuer :

- Une méthode graphique sur les données de la base test qui compare les valeurs empiriques de la variable à expliquer à ses valeurs théoriques issues du modèle : plus il y a de points alignés sur la première bissectrice et plus le pouvoir prédictif du modèle est élevé.
- Une méthode quantitative en comparant la fréquence moyenne empirique calculée à partir des données de la base test et la fréquence moyenne théorique issue du modèle. Plus les valeurs sont proches, mieux est le modèle en termes de prédiction.



Graphique 17 : Valeurs observées VS valeurs modélisées – Actes techniques médicaux (A)

Dans l'exemple considéré, les points sont relativement proches de la première bissectrice. Il est cependant assez subjectif de juger ceci d'autant que l'échelle est grande (i.e. tout écart avec la droite représente en réalité un écart important). Seulement deux ou trois points semblent être à la marge.

De plus, la fréquence empirique moyenne calculée sur la base de test s'élève à 1,417 contre 1,409 prédit. L'écart relatif absolu étant de 0,5 %, il est conclu un modèle satisfaisant en termes de prédiction.

Une dernière remarque avant d'aborder le cas du modèle de coût : il a été observé un certain antagonisme entre prédiction et adéquation. En effet, un modèle pouvait être très mauvais en termes d'ajustement aux données mais très bon en termes de prédiction... Ce constat semble être connu quand il s'agit de GLM.

3.2.3. Un exemple de cas pour le coût

Le modèle de coût est similaire au modèle de fréquence en ce qui concerne la démarche. Les remarques et analyses seront donc plus succinctes. Ce sera aussi le cas pour le modèle de consommation (la démarche de GLM et ses étapes restent relativement les mêmes à chaque fois).

L'exemple est le même que pour le cas de la fréquence puisque le modèle de coût vient en complément du modèle de fréquence dans la détermination de la prime pure.

Pour rappel, le « coût » désigne le montant moyen de remboursement de la mutuelle pour un segment de tarification et une garantie donnés. La variable à expliquer est donc le montant total remboursé par la mutuelle divisé par la quantité d'actes totale associée (MT_RC/QTE_ACTE).

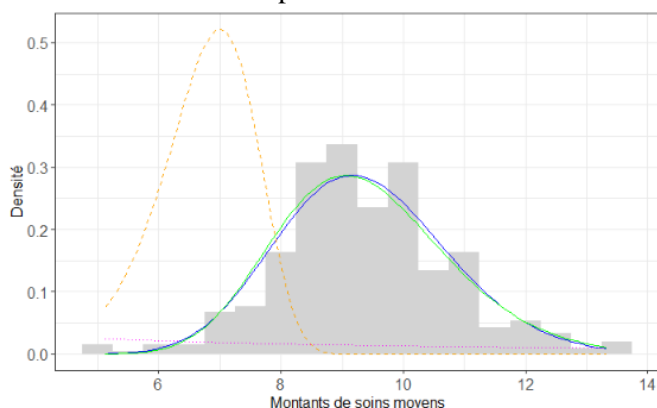
En termes de moyenne empirique interprétable, le coût moyen d'un acte technique médical serait de l'ordre de 9,66 €. C'est ici le montant que paie VirtuaMut' pour ces types d'actes (et non la dépense).

3.2.3.1. Détermination de la loi de coût

Dans le cas du modèle de coût, les lois classiques considérées sont les suivantes : loi log-normale, loi Gamma et loi Weibull.

Annexe 16

Dans le cas de l'exemple :

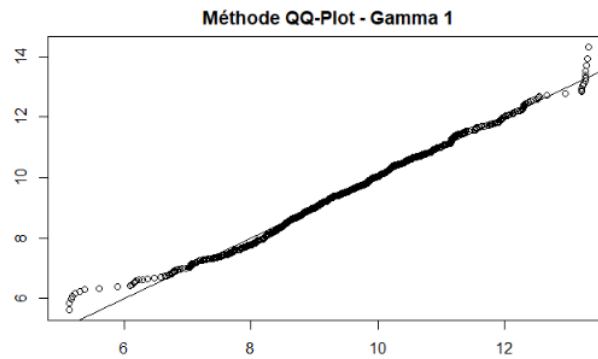


Légende :

- En gris : l'histogramme empirique ;
- En vert : la densité de la loi log-normale théorique (et dont les paramètres ont été estimés via la méthode des moments) ;
- En bleu : la densité de la loi gamma théorique (et dont les paramètres ont été estimés via la méthode de maximum de vraisemblance) ;
- En orange : la densité de la loi Weibull théorique (et dont les paramètres ont été estimés via la méthode de maximum de vraisemblance) ;

Graphique 18 : Densités théoriques et histogramme empirique – Coût – Actes techniques médicaux

Les lois Gamma ou log-normale semblent convenir. La loi Gamma a été retenue comme c'était celle qui a été retenue pour l'estimation du coût pour le segment des actes techniques médicaux sur les données de VirtuaMut' pour la garantie A. Il faudrait le vérifier avec un QQ-plot.



Graphique 19 : QQ-plot Gamma - Coût – Actes techniques médicaux (A)

L’adéquation tend à se confirmer (malgré les petits décrochages aux extrêmes).

Concernant le critère BIC :

Modèle	BIC
Lognormal	1476
Gamma	1468
Weibull	31336

Annexe 15.3
Critère BIC

Tableau 18 : BIC selon la loi candidate - Loi du coût – Actes techniques médicaux (A)

Cela permet de choisir une loi en comparaison relative avec les autres. De manière générale, dans les travaux, il a souvent été observé une différence peu notable entre la loi Gamma et la loi log-normale en termes de BIC. La loi Weibull n’a en revanche jamais été retenue dans nos modèles.

En méthode de test statistique d’adéquation de loi, dans le cas du coût, deux méthodes ont été utilisées :

- Le test de Cramer-von Mises qui peut être vu comme une version plus puissante du test suivant (car moins sensible aux *outliers*) mais il demande plus de ressources et les algorithmes tournaient parfois trop longuement comme cela implique du Bootstrap. Il y a rejet de l’hypothèse nulle (i.e. l’adéquation de la loi aux données) lorsque la statistique du test est supérieure à une valeur critique ;
- Le test de Kolmogorov-Smirnov où il y a aussi rejet de l’hypothèse nulle en cas de statistique du test supérieure à une valeur critique.

Annexe 15.6
Tests adéquation lois continues

Si une loi candidate ne passe pas l’un des deux tests (lorsque les deux sont réalisés), elle est exclue (sauf cas particuliers de sensibilité aux valeurs extrêmes qui n’ont pas été rencontrés). À ce stade, il est toujours possible que plusieurs lois candidates restent en jeu et de la même manière que précédemment (démarche identique que pour la fréquence), plusieurs GLM seront à être générés et à être comparés.

Le choix de loi est réalisé en premier lieu sur les données de VirtuaMut’. Par souci de cohérence avec le second objectif d’extension, il a été décidé de garder ce même choix de loi dans le cas de l’Open DAMIR (et toujours par segment de tarification et par garantie). Cette décision a parfois conduit à un changement de méthode par inadéquation de la loi avec les données de l’Open DAMIR : passage de la méthode GLM fréquence x coût moyen à GLM consommation voire statistique directe.

Pour l’exemple, le test de Cramer-von Mises affiche un non-rejet de l’hypothèse nulle :

	Test_stat	critical_value	Reject
95%	0.1262826	0.529221	FALSE

1 row

Tableau 19 : Résultat du test de Cramer-von Mises - Loi du coût – Actes techniques médicaux (A)

Le test de Kolmogorov-Smirnov se conclut sur une p_{value} à 0,26, supérieure à 0,05. L'hypothèse nulle n'est pas non plus rejetée.

Il semblerait donc que tous ces tests confortent l'idée de s'orienter vers une loi Gamma.

3.2.3.2. Individu de référence, training set, test set

L'individu de référence est le même que celui du modèle de fréquence et le seuil de répartition des données en base test et base d'apprentissage est resté à l'identique (à savoir, 85 %).

Pour rappel, dans le cas de notre exemple, l'individu de référence est une personne de sexe féminin, âgée entre 30 et 39 (classe d'âge 30) et habitant en région D.

3.2.3.3. Modèle GLM et optimisation

Dans le cas de l'exemple, le GLM réalisé est le suivant :

```
Call:
glm(formula = MT_RC_RN/QTE_ACTE ~ CLASSE_AGE + REGION, family = Gamma(link = "log"),
     data = data_agreg_AM_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.36120  -0.07447  -0.00401   0.06877   0.29687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.35889    0.02588  91.132 < 2e-16 ***
CLASSE_AGE0  -0.12847    0.02321  -5.536 6.28e-08 ***
CLASSE_AGE20 -0.04637    0.02301  -2.016 0.044653 *
CLASSE_AGE40  0.01932    0.02378   0.812 0.417162
CLASSE_AGE50 -0.03128    0.02320  -1.348 0.178482
CLASSE_AGE60 -0.09768    0.02347  -4.161 4.04e-05 ***
CLASSE_AGE70 -0.09074    0.02292  -3.958 9.23e-05 ***
CLASSE_AGE80 -0.08940    0.02345  -3.812 0.000164 ***
REGIONO1     -0.25161    0.02926  -8.599 3.19e-16 ***
REGIONO2     0.10755    0.02956   3.638 0.000318 ***
REGIONO3     -0.06281    0.02984  -2.105 0.036063 *
REGIONO4     -0.02699    0.02897  -0.932 0.352158
REGIONO5     -0.15139    0.02952  -5.128 4.97e-07 ***
REGIONO6     -0.10905    0.02982  -3.657 0.000296 ***
REGIONO7     -0.10151    0.02849  -3.563 0.000421 ***
REGIONO8     -0.18931    0.02900  -6.528 2.48e-10 ***
REGIONO9     -0.02994    0.03019  -0.992 0.321984
REGIONO10    -0.14118    0.02849  -4.955 1.15e-06 ***
REGIONO11     0.05770    0.02875   2.007 0.045555 *
REGIONO12    -0.04840    0.02926  -1.654 0.099040 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.01173208)
Null deviance: 8.0548  on 352  degrees of freedom
Residual deviance: 3.9577  on 333  degrees of freedom
```

Toutes les modalités ne sont pas significatives du fait de la présence de p_{values} supérieures à 0,05. Le quantile d'une loi du Khi-deux d'ordre 0,95 avec 333 degrés de liberté est de 376,56. Il est supérieur à la déviance résiduelle. De ce fait, la qualité d'ajustement du modèle est satisfaisante. Il peut être noté que la déviance résiduelle est impactée par le paramètre faible de dispersion, les tests sur les résidus seront privilégiés.

L'étape d'optimisation conduit à regrouper :

- La classe d'âge 40 avec la classe d'âge de référence :
- La région de modalité O12 avec la O1 ;
- Les régions de modalités O9, O10 avec la O4.

Après optimisation, cela donne :

```
Call:
glm(formula = MT_RC_RN/QTE_ACTE ~ CLASSE_AGE + REGION, family = Gamma(link = "log"),
     data = data_agreg_AM_training)

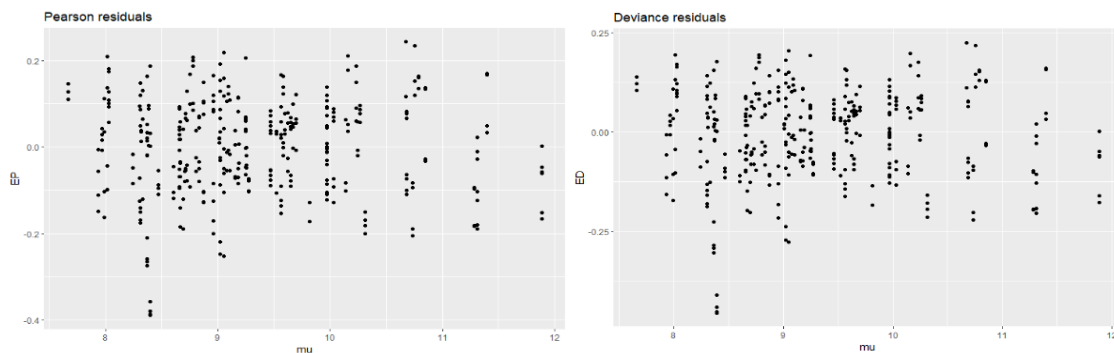
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.45632  -0.08084   0.00899   0.07953   0.22594

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.36872    0.02435  97.272 < 2e-16 ***
CLASSE_AGE0 -0.14166    0.02137  -6.628 1.35e-10 ***
CLASSE_AGE20 -0.05215    0.02117  -2.463 0.014284 *
CLASSE_AGE50 -0.04171    0.02137  -1.952 0.051743 .
CLASSE_AGE60 -0.10742    0.02168  -4.955 1.15e-06 ***
CLASSE_AGE70 -0.10007    0.02102  -4.760 2.87e-06 ***
CLASSE_AGE80 -0.09699    0.02166  -4.479 1.03e-05 ***
REGION01-012 -0.14433    0.02689  -5.367 1.49e-07 ***
REGION02      0.10652    0.03144   3.388 0.000788 ***
REGION03     -0.06298    0.03177  -1.982 0.048273 *
REGION04-09-010 -0.06899    0.02526  -2.731 0.006644 **
REGION05     -0.15186    0.03143  -4.832 2.06e-06 ***
REGION06     -0.10914    0.03175  -3.438 0.000660 ***
REGION07     -0.10218    0.03033  -3.369 0.000840 ***
REGION08     -0.19047    0.03084  -6.175 1.90e-09 ***
REGION011     0.05677    0.03060   1.855 0.064421 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.01330045)
Null deviance: 8.0548 on 352 degrees of freedom
Residual deviance: 4.7559 on 337 degrees of freedom
```

Toutes les modalités sont ici significatives. Ce modèle est retenu.

3.2.3.4. Validation

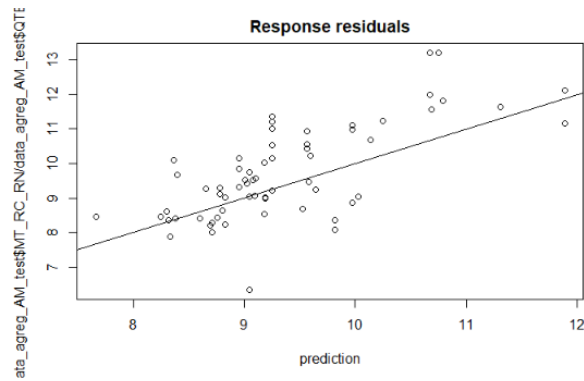


Graphique 20 : Résidus de Pearson et de déviance - Loi du coût – Actes techniques médicaux (A)

Le graphique de gauche présente des résidus de Pearson contenus dans l'intervalle $[-0,5 ; 0,3]$ et centrés en 0. De plus, il n'y a pas de tendance particulière. Le graphique de droite présente des résidus de déviance qui ne dépassent pas en valeur absolue 1. Le modèle est donc valide.

3.2.3.5. Prédiction

Le coût moyen empirique observé sur la base de test s'élève à 9,35 contre 10,02 prédit. L'écart relatif absolu étant de 7 %, il est conclu un modèle suffisamment acceptable en termes de prédiction bien qu'un seuil en deçà de 5 % aurait été plus idéal.



Graphique 21 : Valeurs observées VS modélisées - Coût – Actes techniques médicaux (A)

Les points ont l'air moins bien alignés (que pour le modèle de fréquence) sur la première bissectrice mais les échelles étant peu élevées en valeurs affichées, cela reste relativement correct.

3.2.4. Un exemple de cas pour la consommation (version GLM)

L'idée générale est d'effectuer une tarification par une approche **GLM consommation**. Cela revient à modéliser directement la prime pure. La démarche sera la même (dans son entièreté pour une question de rigueur) pour tous les segments de tarification dont la méthode retenue est un **GLM consommation**.

Le segment retenu à titre d'exemple est celui des prothèses dentaires de la garantie B associée à la base de données de l'Open DAMIR.

Pour ce segment en particulier, les tests de corrélation ont donné les résultats suivants :

	Valeur du coefficient
Pearson	0,46
Kendall	0,28
Spearman	0,44

Tableau 20 : Résultats de tests de corrélation – Prothèses dentaires (garantie B)

Pour les trois tests, le coefficient obtenu étant supérieur au seuil de 0,25, le modèle fréquence x coût moyen n'est pas retenu puisque les hypothèses d'un tel GLM ne sont éventuellement pas respectées.

La méthode ne changeant pas et par souci de redondance, les résultats pour le cas de l'exemple seront présentés succinctement.

3.2.4.1. Détermination de la loi de consommation

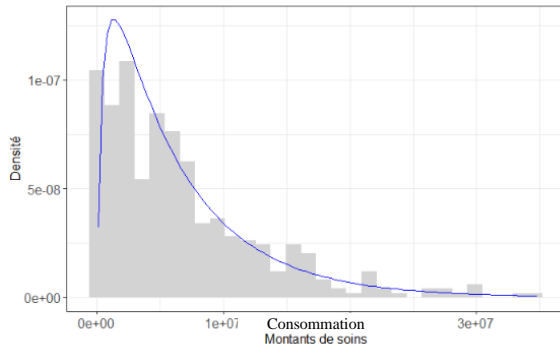
Dans le cas du modèle de consommation, la loi utilisée est la loi Tweedie³⁶ pour sa versatilité (elle englobe entre autres la loi Gamma et la loi Poisson-gamma). L'estimation du paramètre de cette loi (le paramètre γ) est réalisée via la fonction *tweedie.profile* sous R. Par ailleurs, la variable à expliquer est ici le montant de remboursement de la mutuelle (MT_RC, MT_RC_A ou MT_RC_B) et l'exposition totale est gérée en tant qu'*offset* du modèle.

Annexe 16
Tweedie

Dans le cas de l'exemple, c'est une loi Tweedie de paramètre 2,21 (loi stable ou loi Lévy tronquée) qui est estimée sur la base des données et les tests d'adéquation de loi tendent à valider ce choix de loi.

³⁶ Ceci est un abus de langage. C'est en réalité une famille de distributions.

En effet, graphiquement :



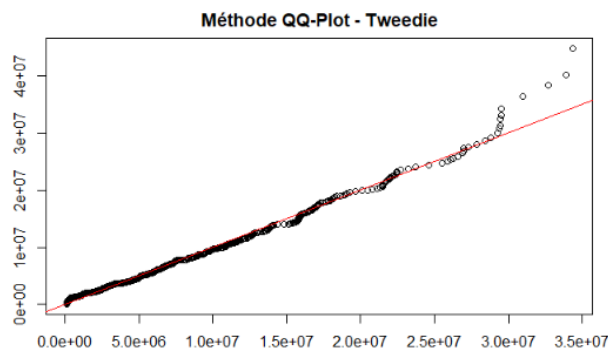
Légende :

En gris : l'histogramme empirique ;

En bleu : la densité de la loi Tweedie théorique estimée.

Graphique 22 : Histogramme empirique VS densité théorique – Prothèses dentaires (garantie B)

La forme de la courbe est assez pertinente vis-à-vis de la forme de l'histogramme.



Graphique 23 : QQ-plot Tweedie – Prothèses dentaires (garantie B)

Mis à part quelques points de queue de distribution lourde, le QQ-plot présente une adéquation de loi satisfaisante. De plus, statistiquement, via les tests de Cramer-von Mises et celui de Kolmogorov-Smirnov, cela tend à se confirmer.

Si à ce stade, la loi Tweedie n'est pas adaptée et elle ne passe pas l'un des tests précédents, sauf exception (non rencontrée dans les travaux), ce sera la méthode **statistique directe** qui sera retenue.

3.2.4.2. Individu de référence, training set, test set

L'individu de référence est défini de la même manière que précédemment présentée et le seuil de répartition de données en base test et base d'apprentissage est resté à l'identique (à savoir, 85 %).

Dans le cas de cet exemple, l'individu de référence est une personne de sexe masculin, âgée entre 40 et 49 (classe d'âge 40) et habitant en région D.

3.2.4.3. Modèle GLM et optimisation

Dans le cadre de l'exemple, le GLM réalisé est le suivant :

```
Call:
glm(formula = MT_RC_RS ~ CLASSE_AGE + REGION, family = tweedie(var.power = p_tweedie,
link.power = 0), data = data_agreg_DENT_training, offset = log(EXPO_TOTALE))
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.093330 -0.011548  0.000399  0.011052  0.067036
```

```

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.391439 0.025818 131.361 < 2e-16 ***
CLASSE_AGE0 -3.942740 0.021134 -186.561 < 2e-16 ***
CLASSE_AGE20 -1.256923 0.022674 -55.434 < 2e-16 ***
CLASSE_AGE30 -0.435207 0.024102 -18.057 < 2e-16 ***
CLASSE_AGE50 0.282841 0.024683 11.459 < 2e-16 ***
CLASSE_AGE60 0.308246 0.024789 12.435 < 2e-16 ***
CLASSE_AGE70 0.298597 0.023742 12.577 < 2e-16 ***
CLASSE_AGE80 -0.086986 0.023391 -3.719 0.000235 ***
REGION01 -0.713713 0.026320 -27.117 < 2e-16 ***
REGION02 0.173923 0.030373 5.726 2.29e-08 ***
REGION03 -0.142179 0.028203 -5.041 7.61e-07 ***
REGION04 -0.082075 0.027172 -3.021 0.002718 **
REGION05 -0.282287 0.027595 -10.230 < 2e-16 ***
REGION06 -0.251229 0.028967 -8.673 < 2e-16 ***
REGION07 -0.245125 0.027089 -9.049 < 2e-16 ***
REGION08 -0.262840 0.027486 -9.563 < 2e-16 ***
REGION09 -0.203822 0.029342 -6.946 1.98e-11 ***
REGION010 0.007943 0.028065 0.283 0.777338
REGION011 0.092955 0.029282 3.174 0.001641 **
REGION012 0.054886 0.028640 1.916 0.056166 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 0.0004285832)
Null deviance: 14.45998 on 352 degrees of freedom
Residual deviance: 0.14808 on 333 degrees of freedom

```

Toutes les modalités ne sont pas significatives du fait de la présence d'une p_{value} supérieure à 0,05. Ainsi, la phase d'optimisation conduit à regrouper la région de modalité O10 avec la région D.

Après optimisation, cela donne :

```

Call:
glm(formula = MT_RC_RS ~ CLASSE_AGE + REGION, family = tweedie(var.power = p_tweedie,
link.power = 0), data = data_agreg_DENT_training, offset = log(EXPO_TOTALE))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.093349 -0.011320  0.000401  0.010851  0.067019

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.39534 0.02169 156.545 < 2e-16 ***
CLASSE_AGE0 -3.94253 0.02110 -186.867 < 2e-16 ***
CLASSE_AGE20 -1.25668 0.02263 -55.531 < 2e-16 ***
CLASSE_AGE30 -0.43503 0.02406 -18.085 < 2e-16 ***
CLASSE_AGE50 0.28303 0.02464 11.488 < 2e-16 ***
CLASSE_AGE60 0.30846 0.02474 12.467 < 2e-16 ***
CLASSE_AGE70 0.29874 0.02370 12.604 < 2e-16 ***
CLASSE_AGE80 -0.08692 0.02335 -3.722 0.000232 ***
REGION5 -0.71777 0.02208 -32.504 < 2e-16 ***
REGION11 0.16985 0.02673 6.354 6.88e-10 ***
REGION24 -0.14624 0.02426 -6.028 4.39e-09 ***
REGION27 -0.08614 0.02306 -3.736 0.000220 ***
REGION28 -0.28636 0.02355 -12.162 < 2e-16 ***
REGION32 -0.25528 0.02516 -10.148 < 2e-16 ***
REGION52 -0.24920 0.02294 -10.862 < 2e-16 ***
REGION53 -0.26691 0.02341 -11.400 < 2e-16 ***
REGION75 -0.20789 0.02555 -8.136 8.14e-15 ***
REGION84 0.08889 0.02549 3.487 0.000553 ***
REGION93 0.05082 0.02476 2.053 0.040890 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

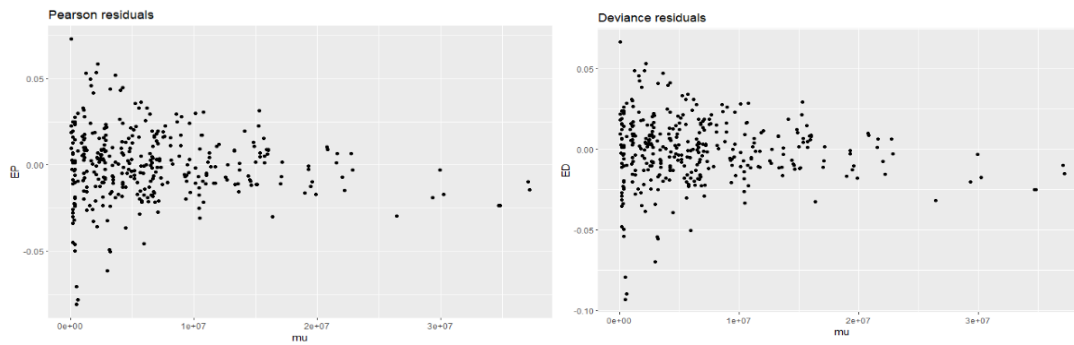
(Dispersion parameter for Tweedie family taken to be 0.0004272335)
Null deviance: 14.45998 on 352 degrees of freedom
Residual deviance: 0.14811 on 334 degrees of freedom

```

Toutes les modalités sont significatives.

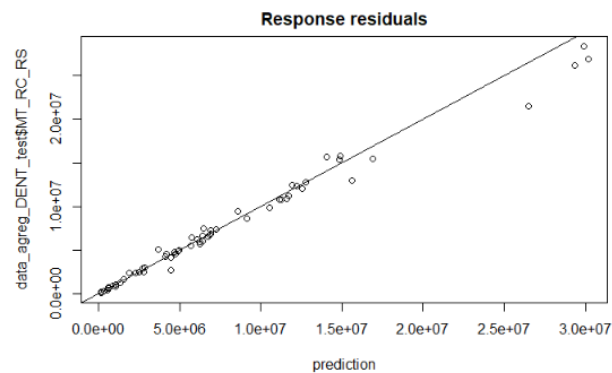
3.2.4.4. Validation

En ce qui concerne la validation, l'affichage des résidus de Pearson et de déviance présente des résidus plus ou moins centrés en 0, contenus dans un intervalle de $[-0,05 ; 0,05]$ pour la plupart des points et ne présentant pas de tendance particulière. Le modèle est donc validé.



Graphique 24 : Résidus de Pearson et de déviance - Consommation – Prothèses dentaires (garantie B)

3.2.4.5. Prédiction



Graphique 25 : Valeurs observées VS valeurs modélisées – Prothèses dentaires (garantie B)

La confrontation graphique entre points prédits et points observés est satisfaisante.

De plus, la consommation empirique moyenne observée sur la base de test s’élève à 22,21 contre 22,99 de prédite. L’écart relatif absolu étant de 3,5 %, il est conclu un modèle prédictif satisfaisant.

3.2.5. Un exemple de cas pour la consommation (version statistique)

Comme mentionné en sous-partie 3.1.1.4., cette méthode n’est utilisée que dans deux cas :

- Cas 1 : le segment de santé considéré n’est pas un segment de tarification retenu en sous-partie 2.2.3. (c’est le cas par exemple de l’implantologie) ;
- Cas 2 : la première méthode de tarification envisagée, à savoir le GLM, n’aboutit pas (c’est le cas par exemple où parmi les lois de probabilité testées, aucune ne semble adaptée aux données et les résultats qui en résultent ne sont pas satisfaisants).

3.2.5.1. Exemple pour le cas 1

Pour illustrer le cas 1, il sera pris comme exemple celui du segment Implantologie de la garantie B. Pour rappel, les calculs se font uniquement sur les données de VirtuaMut’.

La surprime annuelle obtenue par classe d’âge est la suivante :

Implantologie - Garantie B								
Méthode statistique directe (cas 1)								
Classe âge	0	20	30	40	50	60	70	80
Région D	0,00	0,00	2,68	21,39	29,14	13,25	4,90	4,65

Tableau 21 : Résultats de primes pures annuelles – VirtuaMut’ – Implantologie (garantie B)

Ainsi, par exemple, un individu de 45 ans ayant adhéré à la garantie B devra payer, en termes de prime pure annuelle, 21,39 € au titre de la garantie implantologie.

3.2.5.2. Exemple pour le cas 2

Pour illustrer le cas 2, il sera pris comme exemple celui du segment Auxiliaires médicaux de la garantie A de la base de données de l'Open DAMIR.

Auxiliaires médicaux - Garantie A								
Méthode statistique directe (cas 2)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	11,24	10,34	17,74	22,75	28,21	35,93	62,64	152,08
Région O2	15,86	9,32	13,30	18,55	26,01	32,18	46,99	88,22
Région O3	16,74	11,27	13,26	17,84	21,53	24,21	33,64	56,92
Région O4	16,32	11,87	14,45	19,37	24,07	26,06	36,72	70,80
Région O5	17,87	10,20	12,98	17,13	20,62	23,17	33,25	63,60
Région O6	24,51	12,20	17,32	24,06	30,14	35,46	56,05	133,20
Région D	22,34	13,97	17,32	22,87	28,15	31,63	45,14	91,77
Région O7	20,43	13,01	16,09	21,76	26,07	26,89	37,82	68,86
Région O8	24,57	14,76	18,31	23,36	26,82	27,92	39,78	89,34
Région O9	21,19	15,79	18,70	24,12	28,05	30,76	43,60	95,76
Région O10	28,03	18,76	22,55	29,14	34,40	36,36	52,95	141,47
Région O11	26,63	16,56	18,39	24,58	29,82	31,90	44,35	100,66
Région O12	27,15	19,76	23,52	30,52	36,76	39,85	56,82	161,12

Tableau 22 : Résultats de primes pures annuelles – Open DAMIR – Auxiliaires médicaux (garantie A)

Ainsi, par exemple, un individu habitant en région O6, âgé de 22 ans, ayant adhéré à la garantie A, devra payer, sur la base des données de l'Open DAMIR, 12,20 € par an pour la garantie auxiliaires médicaux.

Bien entendu, un tel tableau de tarification peut être érigé parce que le nombre de variables explicatives est modeste. Cela est plus difficilement représentable dans le cas d'un nombre élevé de variables explicatives (les tableaux pour stocker tous les résultats seraient *a priori* hors-normes).

3.2.6. Résultats

Par souci de clarté et de longueur, seuls les résultats de tarifs finaux en primes pures mensuelles ne seront affichés. Les résultats par segment de santé pourront être fournis sous demande.

La prime mensuelle est déduite de la prime annuelle via une division par 12. De plus, pour rappel, les cotisations présentées ne sont pas les cotisations que les adhérents doivent réellement payer : il manque les commissions et frais divers à ajouter. Il s'agit seulement de primes pures. Par abus de langage, le mot « prix », « tarif » et autres similarités désigneront la prime pure dans la suite de l'écrit.

Enfin, les montants affichés (en euros) tiennent compte de tous les segments de santé : ceux dont le tarif a été déduit d'une méthode GLM (fréquence x coût moyen ou consommation – sur la base des modèles non optimisés) et ceux dont le tarif a été déduit de calculs statistiques (cas 1 et 2). Comme les primes pures des cas 1 sont déduites uniquement des données de VirtuaMut', elles ont été ajoutées aux primes pures issues de l'Open DAMIR telles quelles.

Ce sont les GLM non optimisés qui ont été retenus pour ne pas prendre en compte la part de subjectivité créée lors des regroupements des différentes modalités. Les résultats de tarifs avec les GLM optimisés feront l'objet d'une étude à part. Ces regroupements permettant parfois d'augmenter le pouvoir prédictif des modèles, il faudra veiller à ce que la prédiction reste correcte avec les modèles non optimisés.

Enfin, il sera considéré dans la suite qu'une marge de différence de 1 % au niveau des comparaisons signifiera une équivalence de primes pures. Un seuil de 2 % sera envisagé pour les tests et un seuil de 5 % sera défini comme critique.

3.2.6.1. Garantie A

Open DAMIR (avec GLM non optimisés)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	9,98	13,08	16,91	20,63	28,00	32,19	46,96	60,77
Région O2	14,14	17,23	22,11	28,14	37,50	41,61	56,58	66,98
Région O3	13,09	16,32	19,94	25,50	34,02	38,17	52,72	60,09
Région O4	13,65	17,19	21,13	26,71	35,43	39,39	54,10	62,56
Région O5	13,72	16,24	20,29	25,70	34,27	38,37	53,21	61,90
Région O6	14,35	16,53	21,10	27,13	36,36	41,76	58,33	71,88
Région D	15,08	18,62	23,06	29,19	38,53	43,30	58,79	67,66
Région O7	13,93	16,80	20,42	25,77	34,28	38,26	53,27	62,80
Région O8	14,09	16,63	20,25	25,17	33,12	36,66	51,15	61,27
Région O9	14,82	17,83	21,73	27,19	35,95	40,58	56,13	66,52
Région O10	15,01	18,23	22,24	27,99	36,97	41,40	57,49	71,70
Région O11	14,96	18,42	22,10	28,05	37,05	41,21	56,44	66,48
Région O12	15,74	19,73	24,01	30,00	39,16	43,21	58,49	73,85

Tableau 23 : Primes pures mensuelles modélisées sur la base des données de l'Open DAMIR (A)

La table de tarification pour la garantie A déduites des données de l'Open DAMIR permet de se rendre compte d'un effet « région » (lecture en « lignes » de la table) : le tarif n'est *a priori* pas le même sur toutes les régions mais pour autant, le prix n'évolue pas de façon drastique non plus d'une région à l'autre. Cela nous conforte donc avec nos attentes des résultats.

Le prix en région D semble aussi être plus élevé que sur les autres régions (mise à part la région O12). Nous nous attendions à ce que plus de régions présentent un prix plus élevé que la région D. Cela pourrait cependant s'expliquer par le type de garantie que constitue la garantie A (une entrée de gamme). En effet, plus le niveau de garantie est bas, et plus les plafonds de remboursement de la mutuelle sont atteints rapidement. Ainsi, en termes de coût moyen, la localisation n'aurait pas tant d'effet comme les frais réels sont suffisamment élevés pour couvrir le montant maximal de remboursement de la mutuelle, quelle que soit la région considérée.

VirtuaMut' (avec GLM non optimisés)								
Classe âge	0	20	30	40	50	60	70	80
Région D	16,09	18,38	18,97	24,34	34,17	42,39	50,90	76,18
Écart en € avec la région D de l'Open DAMIR	1,01	-0,24	-4,09	-4,85	-4,36	-0,91	-7,89	8,52
Écart en %	6,3%	-1,3%	-21,6%	-19,9%	-12,8%	-2,1%	-15,5%	11,2%

Tableau 24 : Primes pures mensuelles modélisées sur la base des données de VirtuaMut' (A)

De plus, le tarif sur la région D issu des deux bases de données diffère et cela est cohérent avec le fait que de manière globale, le portefeuille de la mutuelle n'est pas représentatif de la population française sur la région d'étude (cf. Section 3 du Chapitre 2). Les écarts peuvent aller jusqu'à -22 % et globalement, les tarifs issus de VirtuaMut' sont inférieurs à ceux issus de l'Open DAMIR (à l'exception de la classe des grands âges). Cela pourrait s'expliquer par l'effet « **portefeuille restreint** ».

Il est observé une faible variation de prime pure entre les classes d'âge 20 et 30 pour le tarif basé sur les données de VirtuaMut', ce qui signifierait une éventuelle sous-tarification de la classe d'âge 30 ou une sur-tarification de la classe d'âge 20.

Si nous comparons les consommations déduites des différents choix de modèles de tarification qui ont été effectués avec les consommations empiriques du portefeuille de VirtuaMut' (cela revient en fait à comparer avec un tarif uniquement basé sur la méthode de statistique directe) :

Tarif mensuel VirtuaMut' en région D (avec GLM non optimisés) - Garantie A								
Classe âge	0	20	30	40	50	60	70	80
Résultat	16,09	18,38	18,97	24,34	34,17	42,39	50,90	76,18
Test	15,87	16,78	19,54	24,24	34,37	43,53	50,11	76,44
Écart en montant	0,22	1,59	-0,58	0,10	-0,20	-1,14	0,79	-0,26
Écart en %	1,4%	8,7%	-3,0%	0,4%	-0,6%	-2,7%	1,6%	-0,3%

Tableau 25 : Primes pures mensuelles modélisées (« Résultat ») VS observées (« Test ») – Garantie A

Il s'avère qu'en termes de primes techniques modélisées, il y a sur-tarification pour les classes d'âge 0, 20, 70 ans (notamment celle de 20 ans) et sous-tarification pour les classes d'âge 30 et 60 ans.

Au regard des écarts avec la consommation réelle (ligne « test »), il serait possible de choisir d'ajuster les tarifs modélisés conséquemment (en les majorant ou minorant directement). Ce n'est cependant pas dans nos objectifs que de piloter les résultats. Notre attention se porte plutôt sur les causes de ces différences : les principaux coupables sont les sous-familles des Verres, Montures et Prothèses dentaires. Cette dernière, qui a plus de poids dans les écarts que les deux autres, présente des consommations modélisées plus importantes pour les classes d'âge 20 et 70 ans et moins élevées pour la classe d'âge 30 de manière significative. Cela signifie que le GLM utilisé n'est pas bien ajusté aux données malgré les tests effectués et cela pourrait être dû à une quantité de données non suffisante pour une telle régression (ce segment présentait une quantité proche du seuil limite fixé). Retenir une méthode de statistique directe (cas 2) pour ce segment semble être une solution. De même pour les montures.

Tarif mensuel corrigé VirtuaMut' en région D (avec GLM non optimisés) - Garantie A								
Classe âge	0	20	30	40	50	60	70	80
Région D corrigé	15,99	16,72	19,10	24,13	33,66	43,12	49,70	75,93
Test	15,87	16,78	19,54	24,24	34,37	43,53	50,11	76,44
Écart en montant	0,12	-0,07	-0,44	-0,11	-0,71	-0,42	-0,41	-0,51
Écart en %	0,8%	-0,4%	-2,3%	-0,5%	-2,1%	-1,0%	-0,8%	-0,7%

Tableau 26 : Primes pures mensuelles corrigées VS observées (« Test ») – Garantie A

En termes de comparaison avec les tarifs réels pratiqués de la mutuelle en 2019³⁷ et après prise en compte de la donnée du ratio P/C (moyen, entre 2018 et 2019) inférieur à 100 %, il est observé une sous-tarification globale pour la première moitié des classes d'âge (de 0 à 40) et une sur-tarification sur la seconde moitié de classes d'âge (par rapport aux tarifs réels). À titre de précision, les tarifs présentés

³⁷ Pour cause de confidentialité, l'étude chiffrée ne pourra pas être présentée.

sont « bruts », i.e. sans ajustement quelconque. Par exemple, ils ne tiennent pas compte de décision de recalibrage interne de la mutuelle pour notamment mettre en œuvre un mécanisme de mutualisation intergénérationnelle (c'est-à-dire que les personnes âgées paient moins et sont « compensées » par les jeunes qui paient plus). Ce mécanisme pourrait alors expliquer ces écarts de résultats tarifaires.

3.2.6.2. Garantie B

Open DAMIR (avec GLM non optimisés)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	10,70	15,41	20,20	28,87	33,33	38,57	49,97	63,75
Région O2	17,37	23,27	30,03	42,65	49,97	55,71	69,35	77,74
Région O3	14,31	19,42	24,16	35,64	41,67	47,23	59,29	66,19
Région O4	14,84	20,24	25,30	36,83	43,05	48,46	60,76	68,75
Région O5	14,68	18,96	24,05	35,26	41,21	46,46	58,45	66,67
Région O6	15,41	19,45	25,10	36,94	43,66	50,40	64,37	77,29
Région D	16,58	22,05	27,66	39,83	46,76	52,99	66,16	74,23
Région O7	14,86	19,48	24,13	35,35	41,23	46,38	58,47	67,70
Région O8	14,79	18,99	23,57	34,22	39,41	44,03	55,48	65,64
Région O9	16,09	21,05	26,08	37,44	43,75	49,91	63,20	72,84
Région O10	16,07	21,14	26,19	37,66	44,05	49,87	63,42	77,21
Région O11	16,83	22,54	27,52	39,72	46,59	52,63	66,28	75,08
Région O12	17,35	23,34	28,81	40,66	47,38	53,10	66,40	81,27

Tableau 27 : Primes pures mensuelles modélisées sur la base des données de l'Open DAMIR (B)

Un effet région est aussi observé via la table de tarification précédente issue de l'Open DAMIR. Cette fois-ci, les régions O2, O11 et O12 présentent des tarifs plus élevés que la région D : l'extension future de tarifs effectuée devrait aller dans ce sens.

VirtuaMut' (avec GLM non optimisés)								
Classe âge	0	20	30	40	50	60	70	80
Région D	19,79	24,35	30,69	45,65	47,71	58,89	69,76	88,06
Écart en € avec la région D de l'Open DAMIR	3,21	2,30	3,03	5,82	0,95	5,91	3,60	13,83
Écart en %	16,2%	9,5%	9,9%	12,7%	2,0%	10,0%	5,2%	15,7%

Tableau 28 : Primes pures mensuelles modélisées sur la base des données de VirtuaMut' (B)

De plus, le tarif sur la région D issu des deux bases de données diffère encore une fois mais *a contrario* du cas précédent, le tarif déduit des données de VirtuaMut' a tendance à être à la hausse (jusqu'à +16 %). Cela paraît cohérent compte tenu du fait que la majorité (plus de la moitié) des adhérents à la garantie B dans la mutuelle sont des personnes âgées de plus de 60 ans (la proportion de personnes âgées est bien plus importante que la proportion visible dans la population française générale).

En termes de comparaison³⁸ avec les tarifs réels de la mutuelle en 2019, sur la base du ratio P/C moyen entre 2018 et 2019 (hors frais, taxes et chargement) et de la connaissance des ajustements réalisés par la mutuelle, il est conclu un tarif suffisamment cohérent.

Par ailleurs, la garantie B est plus coûteuse que la garantie A, ce qui est satisfaisant compte tenu du fait que la garantie B est meilleure que la garantie A (en termes de prestations proposées par la mutuelle).

³⁸ Pour cause de confidentialité, l'étude chiffrée ne pourra pas être présentée.

Si nous comparons les consommations déduites des différents choix de modèles de tarification qui ont été effectués avec les consommations empiriques du portefeuille de VirtuaMut' :

Tarif mensuel VirtuaMut' en région D (avec GLM non optimisés) – Garantie B								
Classe âge	0	20	30	40	50	60	70	80
Original	19,79	24,35	30,69	45,65	47,71	58,89	69,76	88,06
Test	19,11	26,56	31,34	49,99	48,32	58,77	69,85	87,50
Écart en montant	0,67	-2,21	-0,65	-4,34	-0,61	0,13	-0,10	0,57
Écart en %	3,4%	-9,1%	-2,1%	-9,5%	-1,3%	0,2%	-0,1%	0,6%

Tableau 29 : Primes pures mensuelles modélisées (« Résultat ») VS observées (« Test ») – Garantie B

Il s'avère qu'en réalité, en termes de primes techniques modélisées, il y a sur-tarification pour la classe d'âge 0 et sous-tarification pour les classes d'âge 20 à 50. Au vu de la cohérence avec le tarif réel de la classe d'âge 40 et de l'attente d'un tarif croissant en santé, l'écart constaté ne sera pas traité.

Les principaux coupables sont les Verres où le GLM utilisé (celui de la fréquence) n'est pas bien ajusté aux données malgré les tests effectués. Pour la classe d'âge 20 spécifiquement, il s'avère que le problème serait lié à tous les segments de tarification. Nous choisissons de nous en arrêter à là pour les correctifs comme la prime pure obtenue après correction reste cohérente avec la prime pure réellement pratiquée pour ces classes d'âge.

Tarif mensuel corrigé VirtuaMut' en région D (avec GLM non optimisés) – Garantie B								
Classe âge	0	20	30	40	50	60	70	80
Région D corrigé	19,41	24,75	30,91	46,84	48,13	58,87	69,86	88,01
Test	19,11	26,56	31,34	49,99	48,32	58,77	69,85	87,50
Écart en montant	0,29	-1,81	-0,43	-3,15	-0,18	0,10	0,01	0,51
Écart en %	1,5%	-7,3%	-1,4%	-6,7%	-0,4%	0,2%	0,0%	0,6%

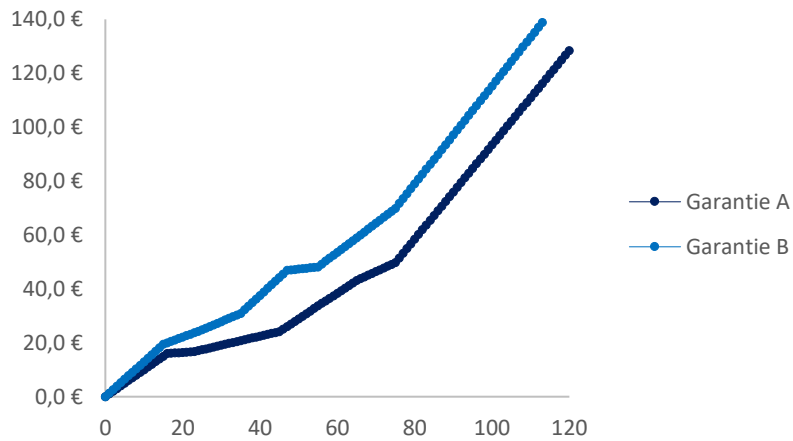
Tableau 30 : Primes pures mensuelles corrigées VS observées (« Test ») – Garantie B

Néanmoins, une solution envisageable pour la classe d'âge 20 serait de ne considérer uniquement pour cette dernière que la méthode statistique directe comme les modèles GLM semblent avoir du mal à s'ajuster sur cet intervalle d'âges.

3.2.6.3. Exemple de lissage

Que ce soit pour la garantie A ou la garantie B, les tarifs tels quels suggèrent par exemple qu'un individu de 40 ans et un individu de 49 ans (i.e. les âges des bornes extrêmes de la classe d'âge 40), pour une même garantie, paient tous deux la même prime pure. Or, ce n'est pas le cas. Il existe différentes façons d'ajuster et de lisser les tarifs afin d'en associer un par âge, l'un des plus connus étant le lissage de Whittaker-Henderson.

Nous n'aborderons pas dans nos études les choix d'ajustements et de lissages possibles, qui seront à la décision de l'instance de direction de la mutuelle, conformément à sa politique tarifaire. Nous nous contenterons cependant de proposer une visualisation des résultats via un lissage linéaire simple.



Graphique 26 : Lissage linéaire des tarifs mensuels issus des données de VirtuaMut' sur la région D

Ainsi, l'allure linéaire en pente haussière des tarifs « bruts » d'ajustements semble être cohérente avec ce qui est attendu d'un tarif d'assurance santé.

Il s'agira dans la suite d'étendre le tarif de la mutuelle à d'autres régions, grâce à ce qui est observé dans les tableaux de tarification basés sur les données de l'Open DAMIR.

Section 3 – Extension du tarif

3.3.1. Détermination des coefficients d'ajustement

La première étape pour étendre le tarif de la mutuelle sur la région D à d'autres régions (O1, ..., O12) est d'avoir une égalité entre les tarifs modélisés sur la région D via les données de la mutuelle et ceux modélisés via les données de l'Open DAMIR. En effet, c'est via les résultats de l'Open DAMIR que l'extension sera faite mais le point de départ est le tarif proposé par la mutuelle sur la région d'étude. Il est donc souhaité que les tarifs soient déjà identiques sur la région D, peu importe la source de données. Ainsi, par exemple, un individu de classe d'âge 20 devra payer en prime pure 24,75 € au titre de son adhésion à la garantie B, et cela, en utilisant les modèles de tarifs basés sur les données de la mutuelle ou ceux basés sur les données de l'Open DAMIR.

3.3.3.1. Idée 1 : La règle de 3 directement

La première idée qui nous est venue est de réaliser simplement une règle de 3 sur la table de tarification de l'Open DAMIR afin de déduire la déformation des tarifs par la donnée de région.

Ainsi, par exemple, pour déterminer un coefficient d'ajustement correspondant à la déformation du tarif due à la région O6 (par rapport à la région D), la table de tarification de l'Open DAMIR est prise et par règle proportionnelle, un coefficient multiplicateur de passage est déterminé de telle sorte que, appliqué au tarif de l'Open DAMIR sur la région D, le tarif de l'Open DAMIR en région O6 est retrouvé :

$$\text{Coefficient multiplicateur}_i = \frac{\text{Tarif région O6 pour la classe d'âge } i}{\text{Tarif région D pour la classe d'âge } i}$$

Open DAMIR (avec GLM non optimisés) – Garantie B								
Classe âge	0	20	30	40	50	60	70	80
Région D	16,69	22,14	27,60	39,40	46,45	52,80	65,88	74,08
Région O6	15,55	19,31	25,17	37,11	43,86	50,61	64,60	77,33
Coefficient multiplicateur	0,932	0,872	0,912	0,942	0,944	0,958	0,981	1,044

Tableau 31 : Coefficients de déformation dus à la région O6 (idée 1) – Garantie B

Ces coefficients sont ensuite appliqués aux primes pures en région D de la table de VirtuaMut' :

$$\begin{aligned} \text{Nouveau tarif en région O6 pour la classe d'âge } i \\ = \text{coefficient multiplicateur}_i * \text{tarif en région D pour la classe d'âge } i \end{aligned}$$

Dans VirtuaMut' (avec GLM non optimisés) – Garantie B								
Classe âge	0	20	30	40	50	60	70	80
Région D	19,41	24,75	30,91	46,84	48,13	58,87	69,86	88,01
Coefficient multiplicateur	0,932	0,872	0,912	0,942	0,944	0,958	0,981	1,044
Nouveau tarif en région O6	18,09	21,59	28,19	44,12	45,45	56,42	68,51	91,87

Tableau 32 : Nouveaux tarifs de la mutuelle en région O6 (idée 1) – Garantie B

Le tarif est ainsi étendu en région O6.

3.3.3.2. Idée 2 : Agir sur les coefficients des GLM

Cette méthode affine la méthode précédente et consiste en :

- Dans le cas d'une tarification d'un segment de santé par la méthode de statistique directe :
 - S'il s'agit du cas 1, comme précisé auparavant, il n'y a pas d'extension de tarif à réaliser. Le sous-tarif contribuant au tarif final est repris tel quel.
 - S'il s'agit du cas 2, la méthode utilisée dans la sous-partie précédente (idée 1) est reprise mais par segment de santé concerné.
- Dans le cas d'une tarification par méthode GLM, l'ajustement par la règle proportionnelle se fait via les coefficients β_0, \dots, β_p estimés issus des modèles non optimisés (avec p le nombre de paramètres des modèles). En effet, il est souhaité une égalisation des coefficients des différents modèles (ceux issus de l'Open DAMIR et ceux issus de VirtuaMut') sur la région D afin d'égaliser les tarifs. Une fois les coefficients d'ajustement déterminés par cette égalisation, ils seront supposés identiques sur les autres régions.

Pour rappel, les modèles GLM généraux utilisés pour chacune des deux garanties et pour chaque segment de tarification sont de la forme suivante :

- Pour la fréquence basée sur les données de VirtuaMut' :

$$\log(E(Y)) = \beta_0 + \beta_1 X_{classe\ d'âge=0} + \dots + \beta_8 X_{classe\ d'âge=80} + offset(Exposition)^{39}$$

où Y est la variable associée à la quantité d'actes et $Exposition$ à l'exposition totale.

- Pour la fréquence basée sur les données de l'Open DAMIR :

$$\log(E(Y')) = \beta'_0 + \beta'_1 X_{classe\ d'âge=0} + \dots + \beta'_8 X_{classe\ d'âge=80} + \beta'_9 X_{région=D} + \beta'_{10} X_{région=01} + \dots + \beta'_{21} X_{région=012} + offset(Exposition')$$

où Y' est variable associée à la quantité d'actes et $Exposition'$ à l'exposition totale.

Pour un individu d'une classe d'âge donnée (sans perte de généralité, il sera pris la classe d'âge 0) qui vit en région D, il est voulu que :

$$\log\left(\frac{E(Y)}{Exposition}\right) = \log\left(\frac{E(Y')}{Exposition'}\right)$$

Avec :

$$\begin{aligned} \log\left(\frac{E(Y)}{Exposition}\right) &= \log(E(Y)) - offset(Exposition)^{40} = \log(E(Y)) - \log(Exposition) \\ &= \beta_0 + \beta_1 X_{classe\ d'âge=0} = \beta_0 + \beta_1 \end{aligned}$$

$$\begin{aligned} \log\left(\frac{E(Y')}{Exposition'}\right) &= \log(E(Y')) - offset(Exposition') = \log(E(Y')) - \log(Exposition') \\ &= \beta'_0 + \beta'_1 X_{classe\ d'âge=0} + \beta'_9 X_{région=D} = \beta'_0 + \beta'_1 + \beta'_9 \end{aligned}$$

$$\text{Ainsi : } \beta_0 + \beta_1 = (\beta'_0 + \beta'_9) + \beta'_1$$

En réalité, β_0 peut aussi se décomposer en la somme d'un coefficient β et d'un coefficient $\beta_9 X_{région=D}$ avec $X_{région=D}$ toujours égal à 1 dans le cas des modèles sur les données de VirtuaMut'. Comme l'effet « âge » de l'individu de référence et l'effet région ne sont pas quantifiés séparément, nous supposons, les égalités suivantes :

$$\beta_0 = \alpha_0 x (\beta'_0 + \beta'_9) \quad \text{et} \quad \beta_1 = \alpha_1 x \beta'_1$$

³⁹ L'équation n'est pas tout à fait exacte telle quelle : un des β_1, \dots, β_8 n'apparaît pas dans le modèle car il est associé à une modalité de l'individu de référence. De même pour l'équation d'après.

⁴⁰ « offset » n'est qu'une notation opérationnelle qui correspond en fait ici à la fonction lien g

Les coefficients α_0 et α_1 précédemment déduits permettent d'égaliser les GLM sur la région D. Il existe donc un coefficient α_0 permettant de corriger les coefficients constants β'_0 et ceux associés aux régions (β'_9 à β'_{21}), puis des coefficients α_1 à α_8 permettant de corriger les coefficients β'_1 à β'_8 associés aux classes d'âge. Idéalement, il aurait fallu connaître la valeur de β , ce qui aurait permis d'appliquer un coefficient plus affiné pour les régions. C'est en appliquant ces coefficients sur les modèles de l'Open DAMIR qu'un tarif étendu est créé.

La démarche est identique pour le cas des modèles de coût moyen et de consommation. Par ailleurs, il est ici visible l'importance de mêmes fonctions de lien, lois d'adéquation et individus de référence lors de la phase de modélisation des modèles de tarification pour chaque segment de tarification afin de faciliter la création de ces coefficients d'ajustement et la cohérence théorique du tout.

3.3.2. Résultats finaux

3.3.4.1. Idée 1 : La règle de 3 directement

Si la démarche « idée 1 » est réalisée sur l'ensemble des autres régions (et pas que celle à titre d'exemple qui était O6), cela donne en tarif mensuel les tableaux suivants :

Open DAMIR Garantie A - Coefficients de déformation effet région								
Classe âge	0	20	30	40	50	60	70	80
Région O1	64%	68%	72%	70%	72%	75%	81%	90%
Région O2	93%	93%	97%	95%	95%	94%	95%	99%
Région O3	87%	89%	87%	87%	87%	88%	90%	89%
Région O4	91%	93%	92%	91%	91%	90%	92%	92%
Région O5	91%	87%	88%	88%	88%	88%	90%	91%
Région O6	96%	89%	92%	93%	94%	96%	99%	106%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	93%	90%	88%	88%	89%	88%	91%	93%
Région O8	94%	89%	87%	86%	86%	85%	87%	91%
Région O9	99%	96%	94%	92%	93%	94%	96%	99%
Région O10	100%	98%	96%	95%	95%	95%	98%	106%
Région O11	99%	99%	95%	95%	95%	95%	96%	98%
Région O12	104%	106%	104%	102%	101%	99%	99%	110%
Extension du tarif de VirtuaMut' - Garantie A (avec GLM non optimisés)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	10,21	11,43	13,79	16,96	24,34	32,50	40,11	68,32
Région O2	14,93	15,53	18,44	22,94	32,08	40,68	47,35	75,48
Région O3	13,97	14,79	16,52	20,96	29,43	37,84	44,67	67,62
Région O4	14,52	15,52	17,50	22,04	30,73	38,96	45,65	70,07
Région O5	14,58	14,60	16,89	21,22	29,76	37,96	44,88	69,25
Région O6	15,33	14,82	17,66	22,53	31,60	41,24	49,18	80,40
Région D	15,99	16,72	19,10	24,13	33,66	43,12	49,70	75,93
Région O7	14,87	15,09	16,83	21,20	29,81	38,14	45,28	70,74
Région O8	15,11	14,95	16,68	20,69	28,82	36,58	43,41	68,75
Région O9	15,76	15,99	17,86	22,31	31,16	40,42	47,74	75,04
Région O10	15,92	16,36	18,26	23,00	32,08	41,15	48,71	80,63
Région O11	15,90	16,62	18,11	22,94	32,02	40,85	47,66	74,73
Région O12	16,61	17,76	19,81	24,64	34,01	42,80	49,29	83,23

Tableau 33 : Nouveaux tarifs de la mutuelle après extension (idée 1) – Garantie A

Open DAMIR Garantie B - Coefficients de déformation effet région								
Classe âge	0	20	30	40	50	60	70	80
Région O1	61%	69%	74%	74%	72%	74%	77%	86%
Région O2	104%	106%	111%	107%	106%	104%	104%	104%
Région O3	87%	88%	88%	91%	90%	90%	90%	89%
Région O4	89%	92%	91%	93%	92%	91%	92%	93%
Région O5	88%	85%	87%	90%	89%	88%	89%	90%
Région O6	93%	87%	91%	94%	94%	96%	98%	104%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	89%	88%	87%	90%	89%	88%	89%	92%
Région O8	90%	86%	85%	87%	85%	84%	85%	89%
Région O9	97%	95%	94%	95%	94%	94%	96%	98%
Région O10	96%	95%	95%	95%	95%	95%	97%	104%
Région O11	102%	103%	100%	100%	100%	99%	100%	101%
Région O12	103%	106%	105%	102%	102%	100%	100%	110%

Extension du tarif de VirtuaMut' - Garantie B (avec GLM non optimisés)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	11,93	16,98	22,90	34,43	34,64	43,64	53,59	75,90
Région O2	20,21	26,32	34,36	50,19	51,15	61,21	72,56	91,86
Région O3	16,79	21,87	27,10	42,60	43,36	52,79	63,04	78,62
Région O4	17,32	22,74	28,21	43,74	44,47	53,83	64,28	81,63
Région O5	17,11	21,10	26,92	42,09	42,89	51,91	61,98	79,19
Région O6	18,09	21,59	28,19	44,12	45,45	56,42	68,51	91,87
Région D	19,41	24,75	30,91	46,84	48,13	58,87	69,86	88,01
Région O7	17,36	21,71	26,81	42,21	42,86	51,84	62,18	80,57
Région O8	17,43	21,17	26,16	40,52	40,97	49,33	59,07	78,10
Région O9	18,74	23,49	29,01	44,48	45,27	55,59	66,93	86,36
Région O10	18,62	23,55	29,23	44,58	45,87	55,98	67,65	91,78
Région O11	19,74	25,40	30,81	46,97	48,05	58,39	69,97	89,02
Région O12	20,04	26,15	32,36	48,00	48,94	59,04	70,13	96,58

Tableau 34 : Nouveaux tarifs de la mutuelle après extension (idée 1) – Garantie B

3.3.4.2. Idée 2 : Agir sur les coefficients des modèles GLM

L'idée 2 appliquée sur chaque segment de tarification conduit au résultat d'extension de tarifs suivant :

Extension du tarif de VirtuaMut' - Garantie A (avec GLM non optimisés)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	10,23	11,89	14,04	17,20	25,36	35,58	45,65	87,01
Région O2	14,03	15,38	18,26	22,75	31,38	42,28	48,95	90,58
Région O3	13,47	14,70	16,18	20,57	29,16	38,57	45,90	68,62
Région O4	13,59	15,45	17,18	21,70	30,49	39,11	46,42	72,43
Région O5	13,70	14,43	16,70	20,91	29,68	39,01	46,82	71,22
Région O6	15,49	14,79	17,40	22,09	31,97	40,68	50,07	84,40
Région D	15,99	16,72	19,10	24,13	33,66	43,12	49,70	75,93
Région O7	14,84	14,96	16,48	20,54	29,52	38,93	46,27	72,43
Région O8	14,53	15,05	16,39	20,11	28,72	36,44	43,45	66,53
Région O9	15,90	16,09	17,88	22,07	31,25	42,57	50,09	80,39
Région O10	16,96	16,93	18,99	23,87	32,49	46,64	53,51	96,52
Région O11	16,71	16,83	18,29	23,11	31,78	42,93	48,84	78,96
Région O12	18,14	18,55	21,56	26,36	35,56	49,24	53,96	105,55

Tableau 35 : Nouveaux tarifs de la mutuelle après extension (idée 2) – Garantie A

Extension du tarif de VirtuaMut' - Garantie B (avec GLM non optimisés)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	11,71	17,48	24,17	36,16	36,07	49,24	64,25	99,88
Région O2	19,07	24,98	33,34	49,17	50,81	62,37	75,53	117,02
Région O3	16,67	22,47	27,31	43,00	43,85	54,66	64,94	81,40
Région O4	17,03	23,00	28,37	43,62	44,63	54,76	65,54	86,21
Région O5	16,94	21,44	28,00	43,06	43,95	55,02	65,88	83,59
Région O6	18,25	21,51	28,34	44,53	45,13	56,94	69,62	94,57
Région D	19,41	24,75	30,91	46,84	48,13	58,87	69,86	88,01
Région O7	17,71	22,49	27,88	42,81	43,62	54,64	65,67	85,41
Région O8	17,63	21,68	26,96	41,06	41,48	50,40	60,45	75,58
Région O9	18,77	24,42	30,18	45,43	46,10	59,30	71,32	94,08
Région O10	19,15	26,21	32,44	48,84	49,73	65,19	79,11	116,27
Région O11	19,86	26,12	31,21	47,56	48,51	60,18	71,96	95,70
Région O12	20,29	27,48	34,25	50,88	51,94	66,02	79,71	129,62

Tableau 36 : Nouveaux tarifs de la mutuelle après extension (idée 2) – Garantie B

L'idée 2 a tendance à significativement exacerber la différence de tarif entre les classes d'âge 70 et 80.

3.3.4.3. Comparaisons et limites

L'idée 1 permet une extension rapide de tarif, sans rentrer dans les différentes équations mathématiques de modèles. Elle serait conseillée opérationnellement comme ses résultats semblent suffisamment cohérents et conformes aux effets régions déduits des données de l'Open DAMIR. L'idée 2 semble cependant plus naturelle théoriquement de par son étape d'égalisation de deux modèles tarifaires sur une même région avant extension de ces derniers spatialement.

Les résultats de l'idée 2 supposent par ailleurs que si un type d'individus ne consomme pas un segment de santé donné sur la région D, dans le tarif étendu, les individus du même type n'en consomment pas non plus (alors qu'il pourrait être pensé que sur d'autres régions, des individus de ce type puissent être amenés à consommer). C'est le cas de :

- La classe d'âge 0 en prothèses dentaires pour la garantie A ;
- Les classes d'âge 0 à 40 en audioprothèse pour la garantie A.

Les cas étant assez limités et la consommation pour de tels actes pour de tels individus étant à la marge, il est jugé que cela n'est pas suffisamment problématique pour remettre en cause les résultats présentés ou pour nécessiter un approfondissement spécifique dans l'immédiat.

Il reste cependant assez difficile de valider ou d'invalider de tels résultats puisqu'il n'était pas détenu au préalable des données de référence permettant la comparaison (excepté le tarif réellement pratiqué). Une solution serait d'évaluer à la fois le tarif et les effets « région » via un *benchmark* de tarifs d'autres organismes. Cependant, plusieurs limites sont à dénombrer pour un tel *benchmark* :

- La structure de frais et de portefeuille de ces tarifs ne sera probablement pas similaire à VirtuaMut', ce qui rend la comparaison peu précise ;
- Les ajustements éventuels de chaque organisme sont inconnus ;
- Les garanties ne sont pas tout à fait les mêmes et cela prend du temps d'isoler un organisme externe qui aurait une grille de garanties très proche et qui propose un devis simple et rapide ;
- Les traitements de données et les méthodes de tarification diffèrent ;
- Il a été constaté que toutes les mutuelles n'avaient pas forcément de zonier (i.e. de tarifs différents par région).

Deux mutuelles externes, dénommées « Assureur 1 » et « Assureur 2 » pour cause d’anonymat, ont été retenues. Pour chacune, les deux garanties qui paraissaient les plus proches des garanties A et B ont été sélectionnées. Un travail de *benchmarking* (via les sites Internet de ces mutuelles) a ensuite été conduit. Il a été supposé pour les deux assureurs externes la même structure de frais que celle de VirtuaMut’ afin de pouvoir effectuer les comparaisons sur la considération des primes pures (et non commerciales).

Pour précision, pour la garantie A, après comparaison des grilles de garanties, la garantie « équivalente » proposée par l’Assureur 1 est en réalité meilleure et celle proposée par l’Assureur 2 est moins bonne. Pour la garantie B, la garantie « équivalente » proposée par l’Assureur 1 est en réalité supérieure, celle proposée par l’Assureur 2 est très similaire.

Comparaison prime pure mensuelle - Garantie A								
Classe âge	0	20	30	40	50	60	70	80
Assureur 1	18,57	20,71	25,71	32,86	38,57	46,43	60,00	76,43
VirtuaMut'	15,99	16,72	19,10	24,13	33,66	43,12	49,70	75,93
Assureur 2	16,43	20,00	25,00	27,86	32,14	40,71	56,43	67,14

Tableau 37 : *Benchmark* pour la garantie A

Pour la garantie A, pour les classes d’âge 0 à 40 et 70, il est possible qu’il y ait une sous-tarification comparée à la concurrence. Le tarif tel quel serait donc *a priori* très compétitif sur la région D. Bien que cela soit à prendre avec des pincettes du fait des limites précitées d’un *benchmark*, il serait éventuellement possible d’envisager une hausse des tarifs pour les classes d’âge 20, 30, 40 et 70. Une hausse de la prime pure sur la classe d’âge 30 serait particulièrement intéressante du fait de la légère sous-tarification déjà observée en sous-partie 3.2.6.1 (résultats pour la garantie A, tableau 25).

Comparaison prime pure mensuelle - Garantie B								
Classe âge	0	20	30	40	50	60	70	80
Assureur 2	25,71	32,86	44,29	56,43	60,00	72,86	97,86	115,71
Assureur 1	27,14	30,00	36,43	52,14	53,57	67,14	85,00	107,14
VirtuaMut'	19,41	24,75	30,91	46,84	48,13	58,87	69,86	88,01

Tableau 38 : *Benchmark* pour la garantie B

Pour la garantie B, les primes pures proposées par VirtuaMut’ semblent aussi être compétitives, notamment au regard de l’Assureur 1.

Pour juger de l’effet « région », les différents coefficients de déformation de par l’effet « région » ont été comparés. Cette comparaison est plus pertinente comme elle est moins sujette aux limites d’un *benchmark*. Les tableaux présentant les différents coefficients de déformation sont exposés en Annexe 17. Plusieurs constatations en ressortent :

- Les Assureurs 1 et 2 semblent diviser la France en quatre zones et considèrent sur chaque zone, une prime pure. Leur zonier est donc moins précis que le nôtre mais contrairement à d’autres organismes, la prime pure se décline quand même selon l’espace géographique (bien que la segmentation soit grossière).
- L’Assureur 1, pour la garantie équivalente à la A, considère la même prime pure pour les zones suivantes :
 - Régions D, O4, O6, O11 ;
 - Régions O3, O5, O7, O8, O9 ;
 - Régions O10 et O12 ;
 - Région O2 ;
 - Il n’y a pas de tarifs proposés pour la région O1.

Seules les régions O2, O10 et O12 présentent un tarif plus élevé que la région D. Pour comparaison, dans nos études, ce sont les régions O10, O12 puis O11 en moyenne et O2 et O6 sur la seconde moitié des classes d'âge qui présentent des primes pures plus élevées que la région D. Cela semble plutôt cohérent bien que l'effet région soit plus impactant dans notre cas.

- L'assureur 1, pour la garantie équivalente à la B, considère la même prime pure pour les zones suivantes :
 - Régions D, O3, O5*, O7*, O8*, O9*⁴¹ ;
 - Régions O2 et O12 ;
 - Régions O10 ;
 - Régions O4, O6*, O11 ;
 - Il n'y a pas de tarifs proposés pour la région O1.

Les régions O2, O4, O6, O10, O11 et O12 présentent un tarif plus élevé que la région D. Pour comparaison, dans notre cas, ce sont les régions O2, O11, O12 puis O9 et O10 sur la seconde moitié des classes d'âge qui présentent des primes pures plus élevées que la région D. Compte tenu de l'identité réelle des régions, notre zonier semble plus cohérent avec la réalité des choses.

- L'assureur 2, pour les garanties équivalentes aux garanties A et B, considère la même segmentation de zones suivantes :
 - Régions D, O2 et O4 ;
 - Régions O3, O9, O11, O12 ;
 - Régions O5, O6, O7, O8 ;
 - Régions O1 et O10.

La région D (et les régions O2 et O4) présente cependant la prime pure la plus élevée par rapport aux autres zones. Au regard de l'identité réelle de la région D, cela ne semble pas être un bon choix de considération.

Cette étude comparative a cependant permis de conclure sur un fait : il semblerait qu'il n'y ait pas de cohérence entre les organismes en ce qui concerne le zonier... Chacun a sa propre façon de faire et ses propres considérations en ce qui concerne les primes pures sur la France entière.

⁴¹ * : à quelques coefficients près.

Section 4 – Tests et sensibilités

3.4.1. Une question de segmentation

Il a été mentionné auparavant l'importance et les difficultés de la classification des différents codes actes en postes, familles et sous-familles d'actes ainsi qu'exposé le choix retenu de méthode de tarification pour chaque sous-famille (voire famille) d'actes. Par ailleurs, la rétention des différents segments de tarification pour la méthode GLM ainsi que la subtilité du jugement sur la quantité suffisante de données par segment ont été abordées (notamment en sous-partie 2.2.3).

Cette sous-section a pour but d'étudier l'éventualité de résultats qui naîtrait d'une segmentation différente de tarif : que se passerait-il dans le cas où le segment de tarification « Verres » était non pas unique mais divisé en trois ? En effet, contrairement aux autres segments, celui-ci a pour particularité de couramment se fragmenter en « Verres simples », « Verres complexes » et « Verres très complexes » dans les grilles de garanties des mutuelles. Cette différenciation a par ailleurs dû être faite pour la déduction du montant RC sur les jeux de données de l'Open DAMIR. Une segmentation plus approfondie serait donc pertinente à étudier.

Une telle nouvelle segmentation a conduit à plusieurs constatations une fois les modèles recalibrés, les algorithmes relancés et les résultats adaptés :

- Pour les verres très complexes, il n'y avait pas suffisamment de données (dans le sens qui a été défini en sous-partie 2.2.3) pour les deux garanties, sur les données de VirtuaMut'. Cela rendait alors non viable une tarification par méthode GLM.
- Pour les verres complexes et simples, l'adéquation d'une loi Tweedie s'est soldée par un échec (notamment sur les données de l'Open DAMIR). Ou bien, lorsque cette dernière était adéquate, la prédiction à l'issue des modèles GLM n'était pas acceptable (surestimation importante sur les premières classes d'âge).

Cela illustre donc les précédents propos sur la difficulté de créer une tarification suffisamment segmentée mais non à l'excès : il y a parfois manque de données sur certains segments voire des résultats non fiables à l'issue des travaux de tarification.

Nous avons alors retenu pour continuer l'étude la méthode **statistique directe** pour chacun des trois sous-segments. Cela permet ainsi de comparer notre choix initial de méthode de tarification (à savoir le GLM pour la garantie A pour les verres) avec celle plus directe sur les statistiques des données :

Test optique / Tarif étendu								
Classe âge	0	20	30	40	50	60	70	80
Région O1	4,7%	2,1%	0,3%	0,1%	0,4%	-0,6%	0,5%	0,0%
Région O2	0,3%	0,6%	-1,0%	-0,1%	0,4%	0,4%	1,5%	0,0%
Région O3	0,0%	0,6%	-0,5%	-0,6%	-0,2%	-0,5%	0,9%	-0,1%
Région O4	0,5%	0,6%	-0,3%	-0,2%	0,2%	-0,1%	1,2%	-0,1%
Région O5	0,6%	1,1%	-0,4%	-0,5%	-0,2%	-0,4%	1,0%	-0,1%
Région O6	0,0%	1,2%	-0,5%	-0,5%	-0,2%	-0,5%	0,8%	-0,1%
Région D	0,1%	0,6%	-0,3%	0,2%	0,3%	-0,1%	1,2%	-0,1%
Région O7	0,3%	1,0%	-0,2%	-0,6%	-0,2%	-0,5%	0,9%	-0,2%
Région O8	-0,2%	1,0%	0,0%	-0,4%	-0,1%	-0,6%	0,9%	-0,2%
Région O9	0,6%	0,9%	-0,1%	-0,3%	0,2%	-0,2%	1,1%	0,0%
Région O10	1,0%	1,0%	-0,2%	-0,3%	-0,1%	-0,5%	0,7%	-0,1%
Région O11	-0,1%	0,4%	-0,3%	-0,1%	0,2%	-0,1%	1,2%	-0,1%
Région O12	1,2%	0,7%	-0,3%	-0,1%	0,3%	0,0%	1,1%	-0,1%

Tableau 39 : Écarts relatifs des tarifs finaux mensuels retenus et de la méthode **statistique directe** appliquée sur les verres simples, complexes et très complexes pour la garantie A

Pour la garantie A, les écarts de méthodes sont moindres. L'écart le plus important étant pour la classe d'âge 0 lors de l'extension de tarif à la région O1 où il est observé que la méthode GLM tend à surestimer légèrement le tarif (écart relatif de 4,7 %, soit 0,50 € par mois).

Pour la garantie B, après la correction effectuée des résultats, la méthode **statistique directe** avait déjà été considérée pour les verres donc le test n'a pas lieu d'être.

Au vu des écarts relatifs présentés et de leur traduction en montant d'euros, le choix d'une méthode GLM vaut le coup : les résultats de la modélisation restent suffisamment proches de la réalité (des données) et il y a possibilité de piloter les modèles (chose non faisable via la méthode **statistique directe**). C'est la raison pour laquelle la prime pure pour la garantie A n'avait pas été corrigée pour le segment Verres bien que celui-ci avait été désigné comme étant un des trois coupables de l'écart constaté lors de la comparaison avec les consommations réelles du portefeuille.

3.4.2. Utilisation des modèles optimisés

Il a été choisi pour les résultats finaux des tarifs et leur extension, une estimation sur la base des modèles GLM non optimisés (pour les segments concernés par cette méthode) dans le sens où les modalités n'ont pas été regroupées et certaines d'entre elles apparaissent non significatives dans les modèles GLM obtenus. Un tel choix a été justifié par le fait qu'au vu du nombre restreint de variables explicatives (une ou deux variables seulement), les modèles étaient déjà suffisamment simples et les données facilement accessibles. Par ailleurs, le regroupement des modalités s'étant fait de manière subjective (à moins de faire tourner un nombre très conséquent d'algorithmes pour chaque modèle GLM, le choix est généralement souvent subjectif), il a été préférable de ne pas tenir compte de cette subjectivité en excluant les modèles optimisés.

Il s'agit dans cette sous-section de tester l'éventualité de résultats suivante : que se passerait-il si, pour tarifier, les modèles GLM optimisés étaient pris au lieu des modèles non optimisés ?

Les écarts relatifs entre les deux possibilités sont présentés ci-après.

Tarif étendu (GLM non optimisé VS GLM optimisé)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	-1,0%	0,6%	0,1%	0,3%	0,2%	0,6%	1,9%	0,0%
Région O2	-0,7%	-0,3%	2,2%	1,6%	1,4%	2,3%	3,2%	0,2%
Région O3	-1,1%	0,0%	0,1%	-0,1%	0,2%	0,9%	2,8%	0,0%
Région O4	-1,5%	-0,5%	-0,3%	-0,4%	0,0%	0,8%	2,8%	0,0%
Région O5	-1,1%	-0,1%	0,0%	-0,1%	0,2%	0,9%	2,8%	0,0%
Région O6	-1,1%	-0,3%	-0,1%	-0,2%	0,1%	0,8%	2,5%	0,0%
Région D	-1,1%	-1,0%	-1,0%	-0,9%	0,0%	1,1%	2,7%	0,0%
Région O7	-1,3%	-0,8%	-0,6%	-0,6%	0,0%	0,7%	2,8%	0,0%
Région O8	-1,4%	-0,8%	-0,7%	-0,7%	-0,1%	0,7%	2,8%	-0,1%
Région O9	-2,6%	-2,6%	-3,6%	-2,9%	-1,8%	-0,6%	1,5%	-0,6%
Région O10	-1,3%	-1,0%	-1,3%	-1,1%	-0,3%	0,6%	2,0%	-0,1%
Région O11	-1,2%	-0,7%	-1,2%	-1,1%	-0,2%	0,8%	2,6%	0,0%
Région O12	-0,9%	-0,1%	-2,6%	-2,4%	-0,4%	1,1%	2,4%	0,2%

Tableau 40 : Écarts relatifs des résultats finaux mensuels retenus et des résultats obtenus avec utilisation de modèles GLM optimisés pour la garantie A

Pour la garantie A, aucun écart n'est supérieur à 5 % et la grande majorité des écarts est sous le seuil de 2 %. En termes de montants, cela signifie par exemple, pour l'écart le plus important associé à la région

O9 et la classe d'âge 30, que l'utilisation des modèles non optimisés au lieu des modèles optimisés ferait payer à un assuré en primes pures, 0,64 € de moins par mois.

Tarif étendu (GLM non optimisé VS GLM optimisé)								
Classe âge	0	20	30	40	50	60	70	80
Région O1	6,6%	-0,8%	-2,1%	0,4%	1,4%	0,3%	0,1%	0,2%
Région O2	6,3%	-1,0%	-4,7%	0,2%	2,0%	1,0%	0,5%	0,3%
Région O3	6,3%	-1,3%	-4,1%	-0,2%	1,4%	0,5%	0,0%	0,2%
Région O4	6,6%	-0,9%	-3,7%	0,2%	1,8%	0,8%	0,5%	0,5%
Région O5	6,6%	-0,8%	-2,9%	0,2%	1,7%	0,7%	0,3%	0,3%
Région O6	5,9%	-0,9%	-3,1%	0,2%	1,6%	0,6%	0,3%	0,3%
Région D	5,4%	-1,4%	-4,7%	-0,3%	1,2%	0,3%	0,0%	0,2%
Région O7	6,7%	-0,8%	-3,2%	0,2%	1,7%	0,7%	0,3%	0,3%
Région O8	6,4%	-1,1%	-3,5%	0,1%	1,5%	0,6%	0,2%	0,3%
Région O9	5,1%	-1,6%	-4,3%	-0,6%	0,8%	-0,1%	-0,5%	-0,3%
Région O10	5,3%	-0,8%	-3,9%	0,1%	1,4%	0,4%	0,2%	0,3%
Région O11	6,0%	-0,6%	-4,2%	0,2%	1,9%	1,0%	0,5%	0,5%
Région O12	6,4%	0,3%	-2,3%	1,1%	2,7%	1,6%	1,2%	0,8%

Tableau 41 : Écarts relatifs des résultats finaux mensuels retenus et des résultats obtenus avec utilisation de modèles GLM optimisés pour la garantie B

Pour la garantie B, des écarts relatifs un peu plus conséquents sont observés, notamment sur la classe d'âge 0 et 30 (à savoir qu'il suffit que l'écart soit important sur la région D pour que, lors de l'extension, cet écart se répercute sur les autres régions). Autrement dit, pour la classe d'âge 0, le regroupement des modalités a son impact. Cependant, en termes de montants, cela signifie par exemple, pour l'écart le plus important associé à la région O7 et la classe d'âge 0, que l'utilisation des modèles non optimisés au lieu des modèles optimisés fait payer à un assuré en termes de primes pures, 1,18 € de plus par mois.

Par ailleurs, il a été observé, pour la garantie B, une « inversion » des tarifs entre les classes d'âge 40 et 50 : c'est-à-dire que le tarif pour la classe d'âge 50 était moins élevé (de quelques centimes) que celui pour la classe d'âge 40. Or, pour un tarif de mutuelle santé, il était attendu une prime croissante selon les âges. Cette « inversion » a ensuite été rectifiée après le constat de quelques regroupements de modalités inadéquats et les résultats de primes pures via l'utilisation de modèles GLM optimisés font état d'un tarif croissant mais dont la différence de cotisations entre la classe d'âge 40 et 50 pour la garantie B reste faible (une vingtaine de centimes d'euros).

Compte tenu des écarts relatifs et de leur traduction en montant d'euros, ainsi que du temps passé pour optimiser l'ensemble des modèles GLM sur les différents segments de tarification, il nous a paru que le coût et le temps étaient trop conséquents par rapport à la plus-value sur le tarif qu'apporterait l'optimisation. Retenir les modèles non optimisés semble corrects opérationnellement. Cela est d'autant plus renforcé par le fait que retenir les modèles optimisés reviendrait à accepter la part de subjectivité qui vient avec (avec la méthode de regroupement choisie).

3.4.3. Le seed

Il s'agit dans cette sous-section d'un test inhérent au programme *R*. Tout comme *Excel*, *R* utilise un générateur de nombres aléatoires (en réalité, pseudo-aléatoires) afin de générer des séquences de nombres lorsqu'il est fait appel à des fonctions telles que *runif*, *sample*, etc. soit, toute fonction demandant la génération de nombres « au hasard ». Cet hasard n'est cependant pas si aléatoire que cela car la séquence de nombres générés est en fait prédéterminée selon le « seed » (en français, la

« graine »). Ce *seed* est à fixer via la fonction *set.seed()* et une fois défini, toute séquence de nombres aléatoires générés à partir de cette fixation sera immuable. Autrement dit, à un *seed* est associé une séquence prédéterminée de nombres qui seront utilisés comme nombres aléatoires en cas d'appel.

Cette fonctionnalité permet notamment de figer les résultats d'algorithmes présentant une part d'aléatoire. Dans notre cas, cela permet d'obtenir entre autres les mêmes montants de primes pures à chaque simulation pour un segment de tarification donné (il faut sinon s'imaginer travailler avec un résultat qui change à chaque fois). À titre de précision, il ne s'agit pas ici de remettre en cause le générateur de nombres pseudo-aléatoires de *R*, qui par défaut, utilise le générateur *Marsenne-Twister* considéré comme étant de bonne qualité. Il s'agit surtout d'étudier l'impact de ce mécanisme sur la tarification élaborée.

Plus concrètement, dans les travaux, le choix d'un *seed*, et donc d'une amorce de séquence de nombres pseudo-aléatoires, a son impact à deux étapes de l'élaboration du tarif (et seulement dans le cas de la méthode de tarification via un GLM) :

- Les tests d'ajustement de loi aux données ;
- La répartition des lignes de données entre la base d'apprentissage (sur laquelle est construit un modèle GLM) et la base de test (sur laquelle un modèle de GLM est testé, validé, et sur laquelle le pouvoir prédictif est évalué).

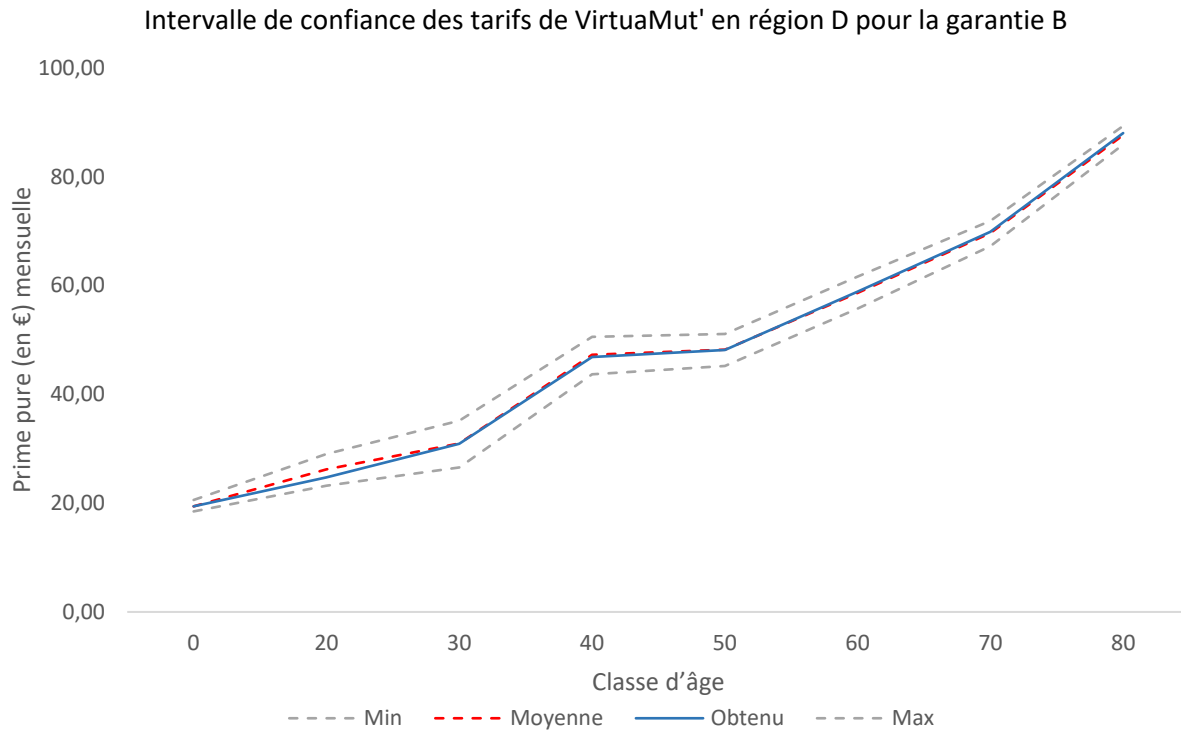
Dans le cas des tests d'ajustement de loi aux données, le Bootstrap sur 1 000 itérations intégré au test de Cramer-von Mises assure que le rejet ou non de l'hypothèse nulle (qui est l'adéquation de la loi aux données) ne soit pas (ou peu) sensible au *seed*. Combattre le caractère aléatoire par un nombre important d'itérations permet d'obtenir des conclusions fiables et non soumises au choix du *seed*.

Il reste donc le second point qui est la ventilation des lignes entre base d'apprentissage et base de test. Pour évaluer l'impact du *seed*, il a donc été décidé de simuler 1 000 fois les primes pures pour chaque segment de tarification concerné par la méthode GLM avec un *seed* variant de 1 à 1 000. Cela permet donc d'une part, d'obtenir une estimation de bornes inférieure et supérieure des résultats finaux (ces bornes mettront donc en évidence la sensibilité du tarif au *seed*) et d'autre part, de déduire un tarif moyen « neutre » au *seed*. À titre de remarque, un autre intervalle de confiance de tarifs aurait pu être fait en considérant l'intervalle de confiance pour chaque coefficient β_i estimé par les modèles GLM.

À titre d'exemple, pour la garantie B, sur la région D (tarif non étendu, sur la base des données de VirtuaMut') :

Tarifs VirtuaMut' - région D								
Classe âge	0	20	30	40	50	60	70	80
Min	18,47	23,22	26,58	43,68	45,20	55,76	67,19	85,90
Moyenne	19,41	26,21	30,99	47,25	48,20	58,64	69,64	87,63
Obtenu	19,41	24,75	30,91	46,84	48,13	58,87	69,86	88,01
Max	20,58	29,05	35,17	50,57	51,10	61,65	71,88	89,32

Tableau 42 : Sensibilité du tarif au *seed* – Garantie B



Graphique 27 : Sensibilité du tarif au *seed* – Garantie B

Lors des travaux, un *seed* égal à 1 a été pris. Ce choix arbitraire a finalement conduit à un tarif (courbe **bleue**) proche du tarif moyen (courbe **rouge**) qui tient compte d'une variation du *seed* de 1 à 1000. Le tarif finalement retenu est donc proche de celui neutre à ce paramètre. De plus, par l'intervalle de confiance ainsi créé, il est possible de conclure que le tarif reste sensible à ce paramètre. L'explication réside dans le fait que pour pouvoir associer le nombre adéquat de bénéficiaires exposés par ligne de prestations agrégées dans les différentes bases de données (cf. partie 2.1.2.13), il a fallu procéder à une agrégation importante des lignes par segment de tarification. Cela signifie donc que les modèles GLM se retrouvent à tourner sur des bases d'apprentissage avec un faible nombre de lignes et cela conduit ainsi à des différences notables lors de la division des lignes entre base d'apprentissage et base de test.

Il sera aussi intéressant de noter que le tarif réel pratiqué par la mutuelle en 2019 est inclus dans cet intervalle de confiance. Des conclusions similaires peuvent être déduites pour la garantie A.

Par ailleurs, à titre de précision, le *seed* peut aussi avoir un impact sur le regroupement des modalités en générant un premier modèle GLM non optimisé avec des significativités légèrement différentes à chaque nouvelle simulation. Les regroupements étant en partie subjectifs, il est possible qu'un changement de significativité de modalités conduise à un choix de regroupements différent et donc, à un modèle GLM optimisé divergent. Comme les résultats finaux sont basés sur des modèles non optimisés, les résultats n'en sont pas affectés.

3.4.4. Modèles basés sur la consommation uniquement

Il a été émis lors du retraitement des données de la mutuelle, notamment en sous-partie 2.1.2.6., quelques réserves sur la fiabilité de la variable QTE_ACTE. Nous avons cependant souhaité faire des modèles séparés de fréquences et de coût quand cela s'y prêtait afin de faire une démarche théorique propre et complète et afin de pouvoir donner la visibilité sur des effets éventuellement différents des variables explicatives sur les deux variables à expliquer.

Il s'agit cependant dans cette sous-section de réaliser le scénario dans lequel la seule méthode de tarification par GLM retenue serait celle sur la variable de consommation. La variable dénombrant de ce fait la quantité d'actes serait complètement exclue des modélisations et des calculs. Ainsi, l'incertitude qui y est liée ne serait plus sujet d'actualité.

À titre d'exemple, pour la garantie B, sur la région D (tarif non étendu, avec des modèles GLM non optimisés, sur la base des données de VirtuaMut'), les seuls segments de tarification dont une méthode GLM fréquence x coût moyen a été retenue sont les suivants :

- Honoraires et actes chirurgicaux ;
- Soins dentaires ;
- Monture ;

Tous les autres segments suivent déjà soit une méthode GLM consommation, soit une méthode statistique directe. Ils ne sont donc pas sensibles à la variable QTE_ACTE. Pour ces trois segments, il a donc été considéré une méthode GLM consommation.

Tarif mensuel VirtuaMut' en région D (GLM non optimisé) – Garantie B								
Classe âge	0	20	30	40	50	60	70	80
Original	19,41	24,75	30,91	46,84	48,13	58,87	69,86	88,01
Test	19,40	24,75	30,93	46,94	48,27	58,96	69,98	88,14
Écart en montant	0,01	0,00	-0,02	-0,11	-0,13	-0,10	-0,12	-0,12
Écart en %	0,0%	0,0%	-0,1%	-0,2%	-0,3%	-0,2%	-0,2%	-0,1%

Tableau 43 : Tarif retenu (Original) VS tarif issu du scénario (« Test ») – Garantie B

Il est observé que les écarts relatifs sont très faibles avec le changement de méthode (moins de 0,3 % en valeur absolue). Plusieurs conclusions peuvent donc être tirées :

- *A priori*, le temps et l'énergie que nécessite le retraitement des quantités d'actes de la base de données de la mutuelle n'apportent pas tant de plus-value ;
- Le retraitement effectué a l'air cependant correct ;
- Sauf en cas de volonté de séparer les effets des variables explicatives sur la fréquence et le coût et sauf en cas d'études de ces deux variables (par exemple, s'il y a volonté d'étudier le nombre de sinistres par an et les variations sur plusieurs années ainsi que l'impact sur la prime pure), effectuer une étude sur la consommation uniquement est suffisante.

Des conclusions similaires peuvent être déduites pour l'étude analogue sur la garantie A.

Et si les GLM étaient abandonnés et que le tarif était déterminé uniquement par la méthode statistique directe ? C'est ce qui avait déjà été présenté lors du recalibrage du tarif et de sa correction en sous-partie 3.2.6. de résultats (comparaison entre les tarifs finaux et les consommations réelles du portefeuille). Outre le fait que la calibration des modèles devient désuète et que le tarif soit fortement restreint aux données sur lequel il est déduit (autrement dit, il n'y a pas de pouvoir de prédiction et il n'y a pas d'ajustement de modèles car il n'y a pas de modèle), l'extension de ce dernier à l'échelle nationale ne pourra se faire que via l'idée 1 d'extension. Cependant, il y a gain de temps et d'énergie, il n'y a aussi que peu de notions mathématiques sous-jacentes. De ce fait, cela pourrait être une perspective quand il s'agit d'estimer rapidement un tarif, sans volonté de l'estimer ni de le modéliser. Il est cependant observé dans le tableau de résultats de test ci-dessous un écart significatif de la prime pure pour les classes d'âge de 20 et 40 ans avec le tarif final déduit de modèles. De plus, il est aussi visualisé une décroissance de prix entre les classes d'âge de 40 et 50 ans, ce qui n'est pas très satisfaisant en notre sens.

Tarif mensuel VirtuaMut' en région D (GLM non optimisé) – Garantie B								
Classe âge	0	20	30	40	50	60	70	80
Original	19,41	24,75	30,91	46,84	48,13	58,87	69,86	88,01
Test	19,11	26,56	31,34	49,99	48,32	58,77	69,85	87,50
Écart en montant	0,29	-1,81	-0,43	-3,15	-0,18	0,10	0,01	0,51
Écart en %	1,5%	-7,3%	-1,4%	-6,7%	-0,4%	0,2%	0,0%	0,6%

Tableau 44 : Primes pures mensuelles corrigées (« Original ») VS observées (« Test ») – Garantie B

En revanche, pour la garantie A, les écarts relatifs sont plus faibles. Peut-être n'est-il pas nécessaire de réaliser des modèles dans le cas de petites garanties (dans le sens des garanties d'entrée de gamme) ?

Tarif mensuel VirtuaMut' en région D (GLM non optimisé) - Garantie A								
Classe âge	0	20	30	40	50	60	70	80
Résultat	15,99	16,72	19,10	24,13	33,66	43,12	49,70	75,93
Test	15,87	16,78	19,54	24,24	34,37	43,53	50,11	76,44
Écart en montant	0,12	-0,07	-0,44	-0,11	-0,71	-0,42	-0,41	-0,51
Écart en %	0,8%	-0,4%	-2,3%	-0,5%	-2,1%	-1,0%	-0,8%	-0,7%

Tableau 45 : Primes pures mensuelles corrigées (« Original ») VS observées (« Test ») – Garantie A

3.4.5. Déformation du tarif par l'inflation et les 100% santé

Ce point a été rapidement abordé en début de chapitre mais il s'agit ici de le développer un peu plus. Le principe d'une tarification consiste, dans la majorité des cas, à se baser sur un historique de données antérieur afin d'en déduire un tarif pour une année ultérieure à cet historique. Par exemple, utiliser des données de 2018 et 2019 pour en déduire un tarif pour 2020 ou 2021.

Pour se faire, une des étapes de la tarification consiste en la transformation des différents montants (notamment le coût) en montants dits « *as-if* », c'est-à-dire, des montants qui pourraient être comparables à l'année de cotation (à savoir, 2021 par exemple). Ainsi, une inflation annuelle des coûts (liée à l'inflation des dépenses en matière de santé) dans la fourchette de [1 % ; 2,5 %] pourrait être à appliquer deux ou trois fois aux montants de l'historique de données afin d'obtenir *in fine* des coûts qui seraient « équivalents » à des coûts de 2021.

D'autres effets sont cependant possibles à prendre en compte et notamment, dans le contexte d'actualité des années 2019 à 2021 : la réforme 100 % santé dont il a été choisi de ne pas aborder. Pour la prendre en compte, il faudrait cependant faire une étude de son impact à part (une estimation possible sur la base de travaux sur d'autres mutuelles serait de l'ordre de 0 % à 5 %) et de l'intégrer ensuite dans les facteurs de transformations en montants « *as-if* ».

Dans une moindre mesure, notre historique de données couvrant 2019, l'impact 100 % santé reste marginal. De plus, étant donné que nous souhaitons avant tout étudier la démarche d'extension et nous concentrer sur la méthodologie, élaborer un tarif sans étape de montants « *as-if* » (et donc, ne pas réaliser un tarif 2021) pour pouvoir ensuite faire une comparaison avec les tarifs disponibles, réels et connus de 2019 nous a paru plus intéressants. En somme, l'ensemble des travaux réalisés restent valables, il suffirait cependant de changer les coûts.

Pour faire le point

Afin d'élaborer des tables de tarification pour VirtuaMut' pour les garanties A et B, trois méthodes de tarification ont été considérées : un GLM sur la variable de fréquence et un autre sur celle du coût, un GLM sur la variable de consommation et enfin, un calcul direct basé sur les statistiques issues des données. Le choix de la méthode repose sur différents tests effectués au préalable (notamment pour juger de la pertinence mais surtout de la compatibilité du choix vis-à-vis des hypothèses qui sous-tendent chaque modèle).

Une fois les tables générées, il a été réalisé une extension des tarifs sur le périmètre national notamment à l'aide de coefficients de passage appliqués sur les coefficients β_i estimés via les modèles GLM ; ou sur la base d'une estimation d'un effet « région » directement sur les tables tarifaires déduites des données de l'Open DAMIR. Cette dernière base se présente en soi comme étant une référence permettant l'extension voulue.

Il convient ensuite de discuter les résultats, de les critiquer et éventuellement tenter de les remettre en cause. Un *benchmark* a donc été élaboré dans ce contexte. Cependant, par ses limites, la comparaison concurrentielle reste limitée et il semblerait au final que chaque organisme ait son propre zonier, sans cohérence unanime particulière entre les différents acteurs sélectionnés.

Il s'agit aussi d'évaluer la robustesse des résultats et les impacts des différentes décisions prises ultérieurement, tant sur les modélisations en elles-mêmes que sur les traitements de données. Pour s'aiguiller, différentes questions ont été mises en avant :

- Quelles hypothèses ont été posées lors du traitement des données ? Lors de la modélisation ? Avant cela, dans la construction ou la récolte même des données ?
- Ces hypothèses ont-elles des alternatives ? S'ébranchent-elles en d'autres possibilités que nous pouvons tester ? Et est-ce que cela a un sens de les tester ?
- Quels sont les paramètres utilisés lors de la modélisation ? Peuvent-ils différer ? Dans quelles situations ?

Le but en soi est de passer en revue tous les choix qui ont été pris et de se demander pour chaque choix, si le test d'une alternative était intéressant.

Ainsi, nous avons retenu les tests suivants et leurs conclusions globales :

- Un test sur la segmentation via la sous-famille des verres où il a été vu que segmenter davantage n'était pas forcément intéressant, voire même incompatible, par la création de sous-segments ayant trop peu de données ;
- Un test sur l'optimisation des modèles GLM qui sont sujets à de la subjectivité et qui ne sont donc pas appréciables dans le cas de modèles à faible nombre de variables explicatives ;
- Un test sur les méthodes choisies de tarification où il a été vu qu'au final, utiliser uniquement des modèles basés sur la consommation (statistique directe ou GLM) était suffisant pour les objectifs poursuivis ;
- Un test sur le « *seed* » de R qui a permis de mettre en avant une sorte d'intervalle de résultats de tarification possible.

Ces tests ont parfois conclu sur des simplifications de démarche mais aussi parfois sur des sensibilités des tarifs à certains paramètres (notamment ici, au *seed*, du fait du caractère agrégé des bases de données). Cette agrégation aura donc jusqu'au bout constitué un obstacle.

Conclusion

Il n'a cessé d'être rappelé mais pour une dernière fois : l'objectif opérationnel de ce mémoire fut de proposer une démarche de tarification permettant d'outrepasser les contraintes spatiales du tarif de produits déjà préexistants mais restreints à une région géographique particulière. Pour cela, une fusion de deux sources de données a été réalisée : les données de sinistralité de la mutuelle VirtuaMut' qui permettent d'estimer un tarif de base, et les données publiques issues de l'Open DAMIR qui permettent d'étendre ce tarif à un périmètre national. Il s'agit donc d'allier données privées et données publiques en matière de santé.

Tout au long de cet écrit, chaque chapitre a déjà été résumé en grande partie par une page récapitulative en fin de chapitre. Il ne s'agira donc pas ici de se répéter indûment comme il suffirait de les coupler afin d'obtenir une conclusion générale, synthétique et chronologique des travaux. Ces travaux ont par ailleurs abouti et une tarification étendue au territoire français entier a bien été réalisée avec succès. Cette dernière a ensuite été soumise à diverses études (comparatives, de sensibilités, etc.) afin de pouvoir juger de sa pertinence. Le but principal est donc accompli : équiper la mutuelle d'une certaine autonomie, compréhension et surtout maîtrise des tarifs de quelques-uns de ses produits, lorsque ces derniers sont vendus sur les autres régions que celle de sa zone de confort. Cette conclusion poursuivra donc avant tout la quête de mettre en évidence les limites et extensions du sujet.

Le premier obstacle ironique et omniprésent des travaux reste l'agrégation des données de l'Open DAMIR assurant son anonymisation. En effet, au premier abord, cette base de données colossale aurait tendance à évoquer un certain sentiment d'optimisme comme il est souvent question de manque de données aux travaux actuariels : or, ici, il est dénombré plus de 720 millions de lignes sur 55 variables, soit plus d'une centaine de gigaoctets de données, sur 2 ans d'historique. L'intuition dirait qu'avec autant de données en matière de santé, des conclusions et des résultats pertinents devraient pouvoir en être extraits (et c'est le cas). Malheureusement, cette abondance a aussi son prix : cela rend son traitement problématique puisqu'il faut des solutions logistiques adaptées qui ne sont pas forcément déjà disponibles au sein des petits organismes. Fort heureusement, les machines virtuelles avec les avancées technologiques du début du XXI^e siècle sont relativement faciles à prendre en main et accessibles à tous à un budget correct, même sans connaissances informatiques au préalable. Une fois cette première embûche passée, la deuxième surgit, spécifique aux objectifs visés via les travaux de ce mémoire : comment déterminer une fréquence à partir de l'Open DAMIR si celle-ci n'informe pas sur le nombre de bénéficiaires par ligne ? Il a donc fallu contourner ce problème en ayant recours à une seconde base de données publique : celle de la démographie française de l'INSEE. Ceci constitue alors une première estimation et donc, un premier biais. Et alors qu'il était pensé qu'il était question de « trop de données », l'association du bon nombre de bénéficiaires par ligne et la cohérence de périmètres de données pour la démarche de tarification exigent d'agrèger les données de manière significative (en fonction des segments de tarification et des caractéristiques des individus). L'excès de données se transforme alors en carence de données, ou, pour être plus précis, en carence de lignes de données, ce qui rend les primes pures estimées sensibles au *seed* de *R* avec les méthodes de tarification retenues.

Par ailleurs, l'Open DAMIR présente aussi des limites même sur la publication de ses données qui ne permet pas une étude en années de soin les plus récentes. Mais outre cela, s'il fallait répondre à la question qui sous-tendait tous les travaux effectués, une réponse affirmative pourra être donnée quant à l'utilité réelle pour une mutuelle telle que VirtuaMut' de la base de données de l'Open DAMIR. D'une part, parce que celle-ci a malgré tout permis d'être la référence utilisée pour une extension de tarifs ; mais d'autre part, elle constitue par sa nature une base informative sur les prestations et les consommations en matière de santé de la population française. Elle peut donc, sur la base de l'étude des

consommations, être utilisée comme source de données de marché pour avoir une idée générale de la consommation par sous-famille, famille et poste de soin. Il est aussi probable qu'être familière avec cette base puisse offrir dans le futur des opportunités qui jusqu'à-là n'avaient pas été encore envisagées.

Cela nécessite cependant d'avoir tout de même établi la classification et la segmentation des actes de soin, tâche très chronophage et non unanime, puisque cela demande la cohésion entre trois regards différents avec une tentative d'appel au renfort à l'Assurance maladie dont l'aide fut limitée. Ceci constitue le deuxième obstacle majeur des travaux, qui a pu être franchi toutefois. Une fois faite, cette classification pourra être réutilisée au fil des années et nécessitera éventuellement quelques mises à jour rapides selon les améliorations et ajouts à l'Open DAMIR.

Des questions peuvent tout de même rester en suspens... Qu'en est-il au final du coût de tels travaux ? Est-ce réellement plus avantageux que l'achat de données ? Quel est l'intérêt face, par exemple, à un *benchmarking* du marché via la méthode du *webscrapping* ? N'est-il pas plus simple de tester un tarif puis de le réajuster les années suivantes ? La discussion sur les intérêts d'un tel projet ne vient qu'ici en conclusion puisqu'il a été jugé plus pertinent et compréhensible de les aborder qu'une fois les travaux connus et abordés.

En ce qui concerne les coûts, l'estimation finale se base sur les prérequis suivants :

- La mobilisation d'une personne (coût de son salaire) ;
- Pour une période d'environ 1 mois ;
- L'acquisition des logiciels tels que la machine virtuelle (estimation à environ 750 € pour une utilisation intensive sur le mois avec la configuration similaire aux travaux présentés), SAS (licence basique) et R qui sont tous deux souvent déjà des logiciels disponibles en interne.

Le temps est estimé sur la base du fait que la démarche à suivre, la classification des codes actes de soin et des exemples de retraitements et de franchissement d'obstacles ont déjà été donnés dans cet écrit. Les parties les plus chronophages et les plus difficiles sont *a priori* ainsi déjà couvertes et en cas de travaux similaires à effectuer, il ne s'agira plus d'avancer à tâtons. Autrement dit, le travail de recherche est déjà effectué, il suffira de passer directement aux travaux opérationnels.

Cela semble donc moins coûteux que l'achat de bases de données ou les tests de tarifs en passant par des ajustements au fil des années. En effet, cette dernière pratique revient à « acheter de l'expérience », c'est-à-dire que si une mutuelle pose son tarif pour une année N donnée et se rend compte que celui-ci est erroné l'année d'après (que ce soit au-dessus ou au-dessous des prix du marché), le biais qui s'en créer devient une perte pour l'organisme (si c'est trop onéreux, il y a perte de clients et si c'est trop peu onéreux, il y a perte d'argent). Mais cette perte est multipliée par autant d'assurés qui auront souscrit et ces derniers auront été engagés pour une nouvelle année au moment de la réalisation du biais d'erreur en N+1. Les pertes peuvent alors assez rapidement atteindre des montants élevés sur deux ans. Une solution pour pallier cela pourrait être l'étude des prix du marché avant de se positionner sur un tarif via la méthode du *webscrapping*, qui n'engendre pas de coût. Cependant, les limites d'un *benchmarking* ont été précédemment abordées. La comparaison de tarif pour isoler un cas pertinent reste difficile du fait :

- De la méconnaissance de la structure de frais et de portefeuille des autres organismes ;
- De la méconnaissance des ajustements effectués (notamment le niveau de mutualisation) par chaque organisme sur leurs primes pures ;
- Les garanties ne sont pas tout à fait les mêmes et cela prend du temps d'isoler un organisme externe qui aurait une grille de garanties très proche et qui propose un devis accessible.

Les autres faiblesses notables d'une telle pratique de *benchmarking* sont aussi :

- L'absence de zonier : certains organismes n'adaptent que très peu voire pas leur tarification selon les espaces géographiques. L'avantage de notre démarche réside donc en partie dans son

extension de tarifs en 13 zones géographiques. Cette extension est par ailleurs maîtrisée et comprise, son origine est connue et son calibrage possible ;

- L'attente des données concurrentielles et le retard qui s'ensuit : cela serait notamment un inconvénient fort en cas de changement de tarifs généraux comme dans le cas par exemple de la réforme 100 % santé ou de tout autre impact réglementaire. En créant de soi-même son propre modèle de tarification, VirtuaMut' serait en mesure d'effectuer les calibrations et adaptations nécessaires.

Et en sommes, il vaut mieux être maître complet du tarif proposé et de sa malléabilité, qui gouverne le résultat technique de l'organisme, que suiveur et dépendant d'autrui pour des sujets d'un tel calibre. Cela ne signifie pas pour autant que le tarif proposé dans cette étude est le bon tel quel, seulement qu'il sera mieux que d'avancer à l'aveuglette ou d'être totalement dépendant d'une source externe.

Pour finir, il est à noter que ce mémoire comporte des sujets qui sont passés sous silence et qui pourraient constituer des axes d'amélioration. Tout d'abord, il serait envisageable de non seulement enrichir les bases de données de VirtuaMut' en lignes mais aussi en colonnes en ajoutant d'autres variables qui pourraient influencer un tarif (notamment, si l'information sur la catégorie socio-professionnelle est disponible). Ensuite, la tarification réalisée est effectuée par tranche d'âges. Or, une telle pratique est fortement déconseillée car il y a des sauts dans le taux de résiliation à chaque borne supérieure et inférieure des classes d'âge. D'où l'intérêt de lisser les résultats obtenus. Un lissage linéaire simple a été mentionné mais une étude plus poussée sur les lissages tarifaires pourrait constituer un nouveau pan à étudier. De plus, en ce qui concerne la démarche elle-même, pour ce qui est des tarifications par GLM, la considération de lois d'ajustement plus complexes que les lois usuelles retenues (comme les lois mélanges) pourraient sensiblement améliorer les modèles. Pour ce qui est de l'extension du tarif, une solution serait éventuellement d'agir sur les lignes en elles-mêmes de l'Open DAMIR (en sélectionner ou en enlever certaines pour reproduire et simuler un portefeuille plausible de VirtuaMut'). Pour ce qui est de l'association des bénéficiaires exposés pour l'Open DAMIR, peut-être existe-t-il une solution plus adéquate. Et enfin, seul l'effet en « lignes » (i.e. la déformation liée aux régions) a été examiné, l'effet en « colonnes » (i.e. la déformation du tarif liée à l'âge) pourrait aussi être un autre axe d'étude complémentaire.

Pour conclure, bien que les doutes aient été présents pendant un temps, l'union entre Open Data et Assurance Santé fait bien une force non négligeable, bien que pour le moment limité, vis-à-vis du cadre d'étude et des objectifs visés. Nous restons cependant optimistes quant à son renforcement dans le futur.

Liste des abréviations

ACPR	Autorité de Contrôle Prudentiel et de Résolution
ACS	Aide à l'acquisition d'une Complémentaire Santé
AIC	Akaike Information Criterion
AME	Aide Médicale d'État
AMI	Acte Médico-Infirmier
AMO	Assurance Maladie Obligatoire
ASIP Santé	Agence des Systèmes d'Information Partagés de Santé
AWS	Amazon Web Services
BIC	Bayesian Information Criterion
BRSS (ou BR)	Base de Remboursement de la Sécurité Sociale
CARSAT	Caisse d'Assurance Retraite et de la Santé au Travail
CAS	Contrat d'Accès aux Soins
CCAM	Classification Commune des Actes Médicaux
CdAM	Catalogue des Actes médicaux
CDO	Chief Data Officer
CGSS	Caisses Générales de Sécurité Sociale
CMA	CoMorbidity Associée
CMU-C	Couverture Maladie Universelle Complémentaire
CNAM	Caisse Nationale d'Assurance Maladie
CNAMTS	Caisse Nationale d'Assurance Maladie des Travailleurs Salariés
CNIL	Commission Nationale de l'Informatique et des Libertés
CPAM	Caisses Primaires d'Assurance Maladie
CSBM	Consommation de Soins et de Biens Médicaux
CSG	Contribution Sociale Généralisée
CSS	Complémentaire Santé Solidaire
DAMIR	Dépenses d'Assurance Maladie Inter Régimes
DMP	Dossier médical personnel
DPO	Data Protection Officer
DREES	Direction de la Recherche, des Etudes de l'Evaluation et des Statistiques

DRSM	Direction Régionale du Service Médical
DSS	Direction de la Sécurité Sociale
EGB	Échantillon Général des Bénéficiaires
FNMF	Fédération Nationale de la Mutualité Française
FR	Frais Réel
GAFA	Google, Apple, Facebook et Amazon
GHM	Groupe Homogène de malade
GHS	Groupe Homogène de séjour
GLM	Generalized Linear Model
GVIF	Generalized Variation Inflation Vector
HAS	Haute Autorité de Santé
IA	Institut des Actuaire
INDS	Institut National des Données de Santé
ISFA	Institut de Science Financière et d'Assurances
LFSS	Loi de Financement de la Sécurité Sociale
LPPR	Liste des Produits et Prestations Remboursables
MLG	Modèle Linéaire Généralisé
Montant RC	Montant pris en charge par le Régime Complémentaire
Montant RO	Montant pris en charge par le Régime Obligatoire
MSA	Mutualité Sociale Agricole
NABM	Nomenclature des Actes de Biologie Médicale
NGAP	Nomenclature Générale des Actes Professionnels
OPTAM	Option de pratique tarifaire maîtrisée
PMSS	Plafond Mensuel de la Sécurité Sociale
PUMA	Protection Universelle Maladie
RAC	Reste à charge
RGPD	Règlement Général sur la Protection des Données
RSA	Revenu de solidarité active
RSI	Régime Social des Indépendants
SNDS	Système National des Données de Santé
SNIIRAM	Système National Inter Régimes d'Assurance Maladie

T2A	Tarifcation à l'Activité
TM	Ticket Modérateur
TRSS	Taux de Remboursement de la Sécurité Sociale
UGECAM	Union de Gestion des Établissements de Caisse d'Assurance Maladie
UNOCAM	Union Nationale des Organismes Complémentaires d'Assurance Maladie
URSSAF	Unions de Recouvrement des cotisations de Sécurité Sociale et d'Allocations Familiales
VIF	Variation Inflation Vector

Liste d'infographies

Couverture : (Source de l'image de base) Inconnu, https://i.pinimg.com/564x/75/19/5d/75195d7f32fcfe6ca5b87d3e2cffebe4.jpg	p.2
Schéma 1 : Prise en charge de la CSBM en 2018	p.10
Tableau 1 : Quelques caractéristiques des organismes assureurs	p.13
Schéma 2 : Récapitulatif des différents montants	p.16
Schéma 3 : Structure du SNIRAM	p.25
Tableau 2 : Aperçu de l'état de l'art sur les Open data en santé	p.26-27
Schéma 4 : Les axes d'analyse des données de l'Open DAMIR	p.28
Tableau 3 : Les dix variables initiales des jeux de données des prestations de VirtuaMut'	p.34
Tableau 4 : Dénombrement de lignes pour chacune des garanties	p.35
Tableau 5 : Les dix variables initiales du jeu de données des adhérents de VirtuaMut'	p.35-36
Tableau 6 : Table de correspondance des âges de l'Open DAMIR	p.37
Graphique 1 : Exemple de boîte à moustache pour le cas des verres de la garantie B	p.40
Tableau 7 : Les variables pertinentes de VirtuaMut' pour une garantie donnée	p.43
Graphique 2 : Dénombrement de lignes de l'Open DAMIR	p.44
Tableau 8 : Liste des variables restantes à l'issue de l'étape 4	p.49
Tableau 9 : Un extrait du lexique de l'Open DAMIR (onglet PRS_NAT)	p.52
Tableau 10 : Segmentation principale retenue des actes	p.52-53
Tableau 11 : Quantités d'actes par sous-familles	p.56-57
Tableau 12 : Segmentation pour la tarification	p.57
Graphique 3 : Répartition des adhérents en 2018 par région selon la garantie	p.59
Graphique 4 : Répartition des adhérents en 2018 par sexe selon la garantie	p.60
Graphique 5 : Répartition des adhérents en 2018 par sexe selon la garantie sur la région D	p.60
Graphique 6 : Répartition des adhérents en 2018 par classe d'âge selon la garantie	p.61
Graphique 7 : Répartition des adhérents en 2018 par classe d'âge selon la garantie sur la région D	p.61
Graphique 8 : Répartition des adhérents en 2018 par qualité de bénéficiaire selon la garantie	p.62
Graphique 9 : Dépense moyenne pour les actes d'imagerie, de radiologie et ostéodensitométrie	p.63
Graphique 10 : Dépense moyenne pour les actes d'imagerie, de radiologie et ostéodensitométrie sur la région D uniquement	p.64
Graphique 11 : Dépense moyenne pour les actes d'imagerie, de radiologie et ostéodensitométrie selon les âges et le sexe	p.64
Graphique 12 : Dépense moyenne pour les actes techniques médicaux	p.65
Graphique 13 : Dépense moyenne pour les actes techniques médicaux sur la région D uniquement	p.65
Graphique 14 : Dépense moyenne pour les actes techniques médicaux selon les âges et le sexe	p.66
Tableau 13 : Méthode de tarification par segment	p.74
Tableau 14 : Résultats de tests de corrélation – Actes techniques médicaux (garantie A)	p.77
Graphique 15 : Tests d'adéquation de loi de fréquence – Actes techniques médicaux (garantie A)	p.79
Tableau 15 : Test du GVIF – Actes techniques médicaux (garantie A)	p.80
Tableau 16 : Test du GVIF – Consultations et visites (garantie A)	p.80
Tableau 17 : Matrice des corrélations entre variables explicatives – Actes techniques médicaux (A)	p.81
Graphique 16 : Résidus de Pearson et de déviance - Fréquence – Actes techniques médicaux (A)	p.84
Graphique 17 : Valeurs observées VS valeurs modélisées – Actes techniques médicaux (A)	p.84
Graphique 18 : Densités théoriques et histogramme empirique – Coût – Actes techniques médicaux	p.85
Graphique 19 : QQ-plot Gamma - Coût – Actes techniques médicaux (A)	p.86
Tableau 18 : BIC selon la loi candidate - Loi du coût – Actes techniques médicaux (A)	p.86

Tableau 19 : Résultat du test de Cramer-von Mises - Loi du coût – Actes techniques médicaux (A)	p.86
Graphique 20 : Résidus de Pearson et de déviance - Loi du coût – Actes techniques médicaux (A)	p.88
Graphique 21 : Valeurs observées VS modélisées - Coût – Actes techniques médicaux (A)	p.89
Tableau 20 : Résultats de tests de corrélation – Prothèses dentaires (garantie B)	p.89
Graphique 22 : Histogramme empirique VS densité théorique – Prothèses dentaires (garantie B)	p.90
Graphique 23 : QQ-plot Tweedie – Prothèses dentaires (garantie B)	p.90
Graphique 24 : Résidus de Pearson et de déviance - Consommation – Prothèses dentaires (garantie B)	p.92
Graphique 25 : Valeurs observées VS valeurs modélisées – Prothèses dentaires (garantie B)	p.92
Tableau 21 : Résultats de primes pures annuelles – VirtuaMut’ – Implantologie (garantie B)	p.92
Tableau 22 : Résultats de primes pures annuelles – Open DAMIR – Auxiliaires médicaux (garantie A)	p.93
Tableau 23 : Primes pures mensuelles modélisées sur la base des données de l’Open DAMIR (A)	p.94
Tableau 24 : Primes pures mensuelles modélisées sur la base des données de VirtuaMut’ (A)	p.94
Tableau 25 : Primes pures mensuelles modélisées (« Résultat ») VS observées (« Test ») – Garantie A	p.95
Tableau 26 : Primes pures mensuelles corrigées VS observées (« Test ») – Garantie A	p.95
Tableau 27 : Primes pures mensuelles modélisées sur la base des données de l’Open DAMIR (B)	p.96
Tableau 28 : Primes pures mensuelles modélisées sur la base des données de VirtuaMut’ (B)	p.96
Tableau 29 : Primes pures mensuelles modélisées (« Résultat ») VS observées (« Test ») – Garantie B	p.97
Tableau 30 : Primes pures mensuelles corrigées VS observées (« Test ») – Garantie B	p.97
Graphique 26 : Lissage linéaire des tarifs mensuels issus des données de VirtuaMut’ sur la région D	p.98
Tableau 31 : Coefficients de déformation dus à la région O6 (idée 1) – Garantie B	p.99
Tableau 32 : Nouveaux tarifs de la mutuelle en région O6 (idée 1) – Garantie B	p.99
Tableau 33 : Nouveaux tarifs de la mutuelle après extension (idée 1) – Garantie A	p.101
Tableau 34 : Nouveaux tarifs de la mutuelle après extension (idée 1) – Garantie B	p.102
Tableau 35 : Nouveaux tarifs de la mutuelle après extension (idée 2) – Garantie A	p.102
Tableau 36 : Nouveaux tarifs de la mutuelle après extension (idée 2) – Garantie B	p.103
Tableau 37 : Benchmark pour la garantie A	p.104
Tableau 38 : Benchmark pour la garantie B	p.104
Tableau 39 : Écarts relatifs des tarifs finaux mensuels retenus et de la méthode statistique directe appliquée sur les verres simples, complexes et très complexes pour la garantie A	p.106
Tableau 40 : Écarts relatifs des résultats finaux mensuels retenus et des résultats obtenus avec utilisation de modèles GLM optimisés pour la garantie A	p.107
Tableau 41 : Écarts relatifs des résultats finaux mensuels retenus et des résultats obtenus avec utilisation de modèles GLM optimisés pour la garantie B	p.108
Tableau 42 : Sensibilité du tarif au seed – Garantie B	p.109
Graphique 27 : Sensibilité du tarif au seed – Garantie B	p.110
Tableau 43 : Tarif retenu (Original ») VS tarif issu du scénario (« Test ») – Garantie B	p.111
Tableau 44 : Primes pures mensuelles corrigées (« Original ») VS observées (« Test ») – Garantie B	p.112
Tableau 45 : Primes pures mensuelles corrigées (« Original ») VS observées (« Test ») – Garantie A	p.112
Schéma 5 : Organisation et structure de la Sécurité Sociale	p.127
Tableau 46 : Suite de l’état de l’art des Open data en santé	p.130
Tableau 47 : Grilles de garanties simplifiées	p.131
Tableau 48 : Liste des variables de l’Open DAMIR	p.135-136
Tableau 49 : Choix de traitement des variables (étape 1)	p.137-138
Tableau 50 : Correspondance des régions	p.139-140

Tableau 51 : Segmentation des actes de soin	p.141
Graphique 28 : Dépense moyenne pour les analyses médicales et examens laboratoires	p.145
Graphique 29 : Dépense moyenne pour les analyses médicales et examens laboratoires uniquement sur la région D	p.145
Graphique 30 : Dépense moyenne pour les analyses médicales et examens laboratoires selon les âges et le sexe	p.146
Graphique 31 : Dépense moyenne pour les auxiliaires médicaux	p.146
Graphique 32 : Dépense moyenne pour les auxiliaires médicaux uniquement sur la région D	p.147
Graphique 33 : Dépense moyenne pour les auxiliaires médicaux selon les âges et le sexe	p.147
Graphique 34 : Dépense moyenne pour les audioprothèses	p.147
Graphique 35 : Dépense moyenne pour les audioprothèses uniquement sur la région D	p.148
Graphique 36 : Dépense moyenne pour les audioprothèses selon les âges et le sexe	p.148
Graphique 37 : Dépense moyenne pour les consultations et visites	p.149
Graphique 38 : Dépense moyenne pour les consultations et visites uniquement sur la région D	p.149
Graphique 39 : Dépense moyenne pour les consultations et visites selon les âges et le sexe	p.150
Graphique 40 : Dépense moyenne pour les forfaits journaliers	p.150
Graphique 41 : Dépense moyenne pour des honoraires et actes chirurgicaux	p.151
Graphique 42 : Dépense moyenne pour des honoraires et actes chirurgicaux uniquement sur la région D	p.151
Graphique 43 : Dépense moyenne pour des honoraires et actes chirurgicaux selon les âges et le sexe	p.152
Graphique 44 : Dépense moyenne pour les actes hospitaliers	p.152
Graphique 45 : Dépense moyenne pour de la pharmacie	p.152
Graphique 46 : Dépense moyenne pour des montures	p.153
Graphique 47 : Dépense moyenne pour des montures uniquement sur la région D	p.153
Graphique 48 : Dépense moyenne pour les montures selon les âges et le sexe	p.154
Graphique 49 : Dépense moyenne pour des verres	p.154
Graphique 50 : Dépense moyenne pour des verres uniquement sur la région D	p.154
Graphique 51 : Dépense moyenne pour les consultations et visites selon les âges et le sexe	p.155
Graphique 52 : Dépense moyenne pour des prothèses dentaires	p.155
Graphique 53 : Dépense moyenne pour des prothèses dentaires uniquement sur la région D	p.156
Graphique 54 : Dépense moyenne pour des prothèses dentaires selon les âges et le sexe	p.156
Graphique 55 : Dépense moyenne pour des soins dentaires	p.156
Graphique 56 : Dépense moyenne pour des soins dentaires uniquement sur la région D	p.157
Graphique 57 : Dépense moyenne pour des soins dentaires selon les âges et le sexe	p.157
Tableau 52 : Paramètres de la famille exponentielle pour des lois usuelles	p.161
Tableau 53 : Espérance et variance de la famille exponentielle	p.161
Tableau 54 : Fonctions de lien	p.162
Tableau 55 : Plusieurs tables de coefficients de déformation pour la garantie A	p.176-177
Tableau 56 : Plusieurs tables de coefficients de déformation pour la garantie B	p.177-178

Bibliographie

Cours

Mylène FAVRE- BEGUET, Cours de Protection Sociale (dispensé à l'ISFA)

Anne MARION, Cours d'Assurance Santé (dispensé à l'ISFA)

Xavier MILHAUD, Cours de Tarification (dispensé à l'ISFA)

Aurélien COULOUMY, Cours de SAS (dispensé à l'ISFA)

[14] Vincent BENOIT, Cours de Prévoyance Collective (dispensé à l'ISFA)

Mémoires

Arnold MEKONTSO, *L'open DAMIR : apport à la maîtrise des dépenses de santé*, 2018

Alban GARNIER, *Apport des Open Data pour évaluer les impacts de la réforme 100% santé*, 2020

Fabien LAGADEC, *Tarification d'un contrat de complémentaire santé par un Modèle Linéaire généralisé*, 2009

Laetitia FENET, *Le risque dentaire en assurance complémentaire santé*, 2012

Jing WANG, *Tarification santé : Mesure des risques associés aux produits modulaires*, 2015

Élodie PAGET, *Amélioration de l'outil de tarification santé d'Actélior à partir des actes codés avec la Classification Commune des Actes Médicaux*, 2008

Fatemeh ABDOLLAHI, *Tarification d'une complémentaire santé à destination des séniors, modulaire par poste de garanties et l'impact sur la solvabilité*, 2017

Pour la partie 1.1.1 sur les régimes obligatoires

[1] Direction de la Sécurité sociale, *Les chiffres clés de la Sécurité sociale 2018*, édition 2019, document disponible via le lien :

<https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/DSS/2019/CHIFFRES%20CLES%202019.pdf>

[2] Le monde avec AFP, *Grand âge : l'Assemblée nationale vote le principe d'une cinquième branche de la Sécurité sociale*, juin 2020, article disponible sur le site de Le Monde :

https://www.lemonde.fr/societe/article/2020/06/16/grand-age-l-assemblee-nationale-vote-le-principe-d-une-cinquieme-branche-de-la-securite-sociale_6042972_3224.html

[3] Ministère des solidarités et de la santé, *Présentation de la sécurité sociale*, octobre 2019, disponible sur le site :

<https://solidarites-sante.gouv.fr/affaires-sociales/securite-sociale/article/presentation-de-la-securite-sociale>

Sécurité sociale, *Les branches*, disponible sur le site :

<https://www.securite-sociale.fr/la-secu-cest-quoi/organisation/les-branches>

Pour la partie 1.1.2 sur les régimes complémentaires

[4] Isabelle DANTON, *Comptes 2019 : les Tops 20 France (résultats 2018)*, janvier 2020, article disponible via le lien :

<https://www.argusdelassurance.com/classements/classements-assureurs/comptes-2019-les-tops-20-france-resultats-2018.157599>

[5] Laure VIEL, *Santé et prévoyance : le marché progresse de 2,8% en 2018*, septembre 2019, article publié sur le site de l'Argus de l'assurance :

<https://www.argusdelassurance.com/assurance-de-personnes/sante-et-prevoyance-le-marche-progresse-de-2-8-en-2018.153189>

La rédaction de Vie Publique, *Santé : panorama de l'assurance complémentaire en France*, avril 2019, article disponible via le lien :

[https://www.vie-publique.fr/en-bref/20309-sante-panorama-de-lassurance-complementaire-en-france#:~:text=les%20mutuelles%20\(au%20nombre%20de,d'assurances%20\(103\).](https://www.vie-publique.fr/en-bref/20309-sante-panorama-de-lassurance-complementaire-en-france#:~:text=les%20mutuelles%20(au%20nombre%20de,d'assurances%20(103).)

Le nombre de mutuelles diminue de près de 9% sur un an, août 2017, article publié sur le site de DevisProx :

<https://www.devisprox.com/pro/nombre-mutuelles-diminue-15175.html>

Pour la partie 1.1.3.2. sur les grilles de garanties

Comment choisir sa mutuelle en fonction des principaux postes de dépenses ?, article disponible sur le site de Mutuelle.fr :

<https://mutuelle.fr/infos/mutuelles/comment-choisir-sa-mutuelle-en-fonction-des-principaux-postes-de-depenses/>

Pour la partie 1.1.4. Nomenclature des actes et pour la Section 2 - La segmentation retenue

[9] Système national des données de santé, *NOMENCLATURE DES CODIFICATIONS*, créé en août 2001, mise à jour en mars 2017, validé par Claude GISSOT et Hélène CAILLOL, document disponible via le lien :

https://www.snds.gouv.fr/download/SNDS_Nomenclature_sous_produits.pdf

Site de l'ameli, page sur la nomenclature et le codage

<https://www.ameli.fr/medecin/exercice-liberal/remuneration/consultations-actes/nomenclatures-codage>

Antoine FRUCHARD, *Code CCAM : définition*, mise à jour en juillet 2020, article disponible sur le site de Réassurezmoi :

<https://reassurez-moi.fr/guide/mutuelle-sante/ccam>

Page web de la Mutuelle complémentaire de la ville de Paris, de l'assistance publique des administrations annexes présentant les prestations :

<https://www.mc602.com/prestations/petit-appareillage-15.html>

Petit appareillage

Antoine FRUCHARD, *Quel remboursement des analyses et examens médicaux par la mutuelle santé ?*, mise à jour en juillet 2020, article disponible sur le site de Réassurezmoi :

<https://reassurez-moi.fr/guide/mutuelle-sante/remboursement-analyses-examens>

Analyses et examens médicaux

Codifications des Prestations Pharmacies, disponible sur le site de resopharma :

https://www.resopharma.fr/modalites/codes_prestations_pha.php

Pharmacie

Comment identifier les soins non-remboursés par la Sécurité sociale ?, article disponible sur le site de la mutuelle ADREA

<https://www.adrea.fr/le-mag/je-suis-un-particulier/soins-non-remboursable-secu>

Actes non remboursés

CNAMTS, *NOMENCLATURE GÉNÉRALE DES ACTES PROFESSIONNELS des Médecins, Chirurgiens-dentistes Sages-femmes et Auxiliaires médicaux*, mise à jour en novembre 2005

<https://www.cafat.nc/documents/20479/2502774/Nomenclature+Generale+Actes+Professionnels+30+mars+2005.pdf>

Médecins et auxiliaires médicaux

Ministère de la santé et des sports, ministère du budget, des comptes publics et de la réforme de l'État, UNCAM, *Cahier des charges des règles de facturation des séjours*, version d'octobre 2010, disponible via le lien suivant :

https://solidarites-sante.gouv.fr/IMG/pdf/FIDES-Cahier_des_charges_Sejours_AMO_et_ATIH_V0-2.pdf

Séjour hospitalier

GIE SESAM-Vitale, *AVENANT 12, Système de facturation SESAM-Vitale*, mai 2017, version 02.11, document consultable au :

https://www.sesam-vitale.fr/documents/20182/54734/PDT-CDC-076+Avenant+12_EV91-A+Convention+m%C3%A9dicale+2016_Tarification+NGAP+v02+11.pdf

Des compléments et suppléments d'actes

Codification des prestations obligatoires à destination des AMO, mis à jour en janvier 2018, document visible sur :

https://www.ameli.fr/fileadmin/user_upload/documents/Annexe10-B2-AMO_Janvier_2018.pdf

Des compléments et suppléments d'actes

Pour la partie 1.2.1. Un second point sur la législation : Open Data

[6] Clorsia NKOUA-GAYAN, *Open data santé : les enjeux de l'ouverture des données dans la gestion du dossier médical*, avril 2020, article publié sur le site Influenceurs du Web :

<https://influenceursduweb.org/open-data-sante-les-enjeux-de-louverture-des-donnees-dans-la-gestion-du-dossier-medical/>

[7] *Données de santé : pourquoi l'opendata est une nécessité et comment y parvenir*, avril 2014, article publié sur le site La Gazette des communes :

<https://www.lagazettedescommunes.com/232095/donnees-de-sante-pourquoi-lopendata-est-une-necessite-et-comment-y-parvenir/>

[8] *Qu'est-ce que l'Open Big Data va changer à la santé ?*, écrit en août 2016 et mis à jour en août 2020, article disponible sur le site de Innov' Asso :

<https://www.innovasso.fr/dossier/quest-lopen-big-data-va-changer-a-sante/>

C.ANNO, *L'open data des données de santé, une vraie révolution?*, février 2016, article publié sur le site Le Petit Juriste :

<https://www.lepetitjuriste.fr/lopen-data-donnees-de-sante-vraie-revolution/>

Squire Patton Boggs, *Loi de santé et « Open Data » pour certaines données de santé*, janvier 2016, article publié sur le site de La Revue :

https://larevue.squirepattonboggs.com/loi-de-sante-et-open-data-pour-certaines-donnees-de-sante_a2791.html

Site officiel de la CNIL sur le RGPD :

<https://www.cnil.fr/fr/comprendre-le-rgpd>

Pour la partie 1.2.2. sur l'état de l'art

Sur la SNIIRAM :

Site de l'assurance maladie (ameli), page sur la SNIIRAM

<https://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/sniiram/>

[11] Dominique POLTON, Philippe RICORDEAU, *Le SNIIRAM et les bases de données de l'Assurance Maladie en 2011*, mars 2011, slides disponibles via le lien :

https://solidarites-sante.gouv.fr/IMG/pdf/CNAMTS_Le_SNIIRAM_et_les_bases_de_donnees_de_l_assurance_maladie_en_2011.pdf

Sur l'Open DAMIR :

Portail de l'Assurance Maladie, *Open Damir : base complète sur les dépenses d'assurance maladie interrégimes*, consulté en août 2020 :

<http://open-data-assurance-maladie.ameli.fr/depenses/index.php>

[12] Pascale QUENNELLE, Marc RAYMOND, *Comment les nouvelles bases de données santé en open data peuvent-elles être source d'inspiration pour l'assurance santé de demain*, slides disponibles sur le site de l'Institut des Actuaire :

https://www.institutdesactuaire.com/global/gene/link.php?doc_id=11745&fg=1

[10] Alice VITARD, *Open data : Santé publique France publie des jeux de données sur le Covid-19*, Mars 2020, article disponible sur :

<https://www.usine-digitale.fr/article/open-data-sante-publique-france-publie-des-jeux-de-donnees-sur-le-covid-19.N944391>

Pour le Chapitre 2

[13] Ben BOLKER, *Minimum number of observations for multiple linear regression*, juin 2012, réponse à une question sur le forum StackExchange, visible au :

<https://stats.stackexchange.com/questions/29612/minimum-number-of-observations-for-multiple-linear-regression>

[15] Santiane, *Imagerie médicale : quel remboursement ?*, mai 2020, article disponible sur le site de la mutuelle Santiane :

<https://www.santiane.fr/mutuelle-sante/guides/imagerie-medecale-remboursement>

[16] Centre de dermatologie Bichat-Confluence, *Les tarifs*, page consultée en décembre 2020 :

<https://dermato-lyon-confluence.com/les-tarifs/>

UNOCAM, *Tarifs moyens nationaux 2020*, document téléchargeable à l'adresse :

<https://unocam.fr/?mdocs-file=1275>

Cabinet Alcimed, *Analyse économique du secteur des appareillages optiques et auditifs*, mars 2011, résultats visibles sur le site :

[https://www.france-assos-sante.org/66-millions-dimpatients/offre-de-soins-et-tarification/lunettes-combien-ca-coute/#:~:text=Selon%20l%C3%A9tude%20men%C3%A9e%20par,143%20%E2%82%AC%20pour%20les%20verres\).](https://www.france-assos-sante.org/66-millions-dimpatients/offre-de-soins-et-tarification/lunettes-combien-ca-coute/#:~:text=Selon%20l%C3%A9tude%20men%C3%A9e%20par,143%20%E2%82%AC%20pour%20les%20verres).)

Lucie CALVET, *Troubles de la vision : sept adultes sur dix portent des lunettes*, mis à jour en décembre 2020, article disponible sur le site de la DREES :

<https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/publications/etudes-et-resultats/article/troubles-de-la-vision-sept-adultes-sur-dix-portent-des-lunettes>

Pour le Chapitre 3

[17] Xavier CONORT, *How Actuaries and Data Scientists could learn from each other*, septembre 2018, slides de conférence consultable à l'adresse :

<http://www.actuaries.jp/lib/meeting/reikai2018-02-siryo2-en.pdf>

[18] ZACH, *How to Test for Multicollinearity in SPSS*, juin 2020, article disponible sur :

<https://www.statology.org/multicollinearity-spss/>

[19] Andrius BUTEIKIS, *Practical Econometrics and Data Science*, 4.5 Multicollinearity, mis à jour en octobre 2020, extrait d'un ouvrage disponible sur :

http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-5-Multiple-collinearity.html#ref-Fox1992

[20] Arthur CHARPENTIER, *Actuariat IARD - ACT2040, Partie 4 - modèles linéaires généralisés*, 2013, slides de cours disponibles sur :

<http://freakonometrics.free.fr/slides-2040-4.pdf>

Loi binomiale, famille Tweedie

[21] I. AISSAOUI-FQAYEH, S. EL-ADLOUNI, T.B.M,J. OUARDA, A. ST-HILAIRE, *Développement de l'estimateur GLM-ML pour le modèle log-normal non stationnaire et application à des précipitations extrêmes* (p.15-16), mai 2006, rapport de recherche disponible via :

<http://espace.inrs.ca/id/eprint/583/1/R000860.pdf>

[22] KJETIL B HALVORSEN, *Is Weibull distribution a exponential family?*, mai 2017, réponse à une question sur le forum StackExchange, visible au :

<https://stats.stackexchange.com/questions/277466/is-weibull-distribution-a-exponential-family>

Annexes

Annexe 1 : Fiche complémentaire sur la Sécurité Sociale.....	129
Annexe 2 : Fiche complémentaire sur les nomenclatures des actes.....	132
Annexe 3 : Fiche complémentaire sur la législation santé.....	133
Annexe 4 : Fiche complémentaire sur l'état de l'art (Open data).....	134
Annexe 5 : Les prestations de VirtuaMut' et de la Sécurité Sociale.....	135
Annexe 6 : Compléments sur le traitement des données provenant de la mutuelle VirtuaMut'	136
Annexe 7 : Liste des variables de l'Open DAMIR.....	139
Annexe 8 : Liste des variables retenues ou supprimées (étape 1).....	141
Annexe 9 : Tableau de correspondance des régions entre VirtuaMut' et Open DAMIR.....	143
Annexe 10 : Classification des actes de l'Open DAMIR selon la segmentation retenue	145
Annexe 11 : Les autres dépenses	149
Annexe 12 : Démonstration – Pour la prime pure.....	162
Annexe 13 : Modèles linéaires généralisés (GLM).....	163
Annexe 14 : Mesures de corrélation.....	168
Annexe 15 : Autres compléments à la méthode GLM.....	170
Annexe 16 : Lois usuelles	177
Annexe 17 : Résultats du benchmarking	180

Annexe 1 : Fiche complémentaire sur la Sécurité Sociale

Cette fiche en annexe permet de compléter la sous-section 1.1.1. sur les régimes obligatoires.

Histoire

Si nous devons résumer brièvement l'histoire de la création de la Sécurité Sociale et plus généralement, du système de santé français, nous retiendrions les points suivants :

- Dès le Moyen Âge : existence d'assistance au sein de certaines corporations professionnelles.
- 1791 (soit, 2 ans après la Révolution française) : décret d'Allarde. Interdiction d'existence de syndicats, corporations, mutuelles... C'est la fin du dispositif limité d'entraide.
- 1898 : Charte de la mutualité. Cette loi fonde les principes mutualistes que nous pouvons retrouver dans le code de la mutualité de nos jours. Elle encourage les « sociétés de secours mutuel » qui étaient venues remplacer le dispositif provenant du Moyen Âge.
- 1914 : Assassinat de Jean Jaurès qui voulait initialiser à l'image de l'Allemagne un système de santé. Début de la Première Guerre Mondiale.
- 1945 : Création de la Sécurité Sociale sous le gouvernement du Général De Gaulle avec les ordonnances du 4 et 19 octobre 1945. Création innovatrice de familles d'ayants droit (conjoint, enfants, retraités) qui persistent encore aujourd'hui.

Retour
p.11

Les 3 objectifs initiaux

- **L'unicité** poursuit la création d'un réseau coordonné de caisses qui se substituerait aux différents organismes déjà existants mais disparates.
- **L'universalité** désigne la volonté de couvrir l'ensemble de la population française (et non que certaines professions).
- **L'uniformité** est la tentative de mettre de l'ordre à la disparité des caisses par souci d'équité.

L'uniformité n'a cependant pas pu être atteinte puisqu'il existe encore de nos jours des régimes spéciaux.

Retour
p.11

Organisation

D'un point de vue institutionnel, la Sécurité Sociale se compose :

- Des organismes de tutelle sous l'égide de l'État et pilotés par la Direction de la Sécurité Sociale (DSS) qui est elle-même rattachée au ministère des solidarités et de la santé et au ministère de l'action et des comptes publics ;
- Des caisses de Sécurité Sociale qui se répartissent sur tout le territoire national et qui assurent à une échelle locale le versement des prestations. Il y a plus de 400 caisses en France, appartenant chacune à un régime de la Sécurité Sociale tel que développé juste après.

Retour
p.11

Les quatre régimes principaux

- **Le régime général de la Sécurité Sociale**
- **Le régime des travailleurs non-salariés non agricoles** couvre l'ensemble des artisans, commerçants et professions libérales. Anciennement appelé RSI pour Régime Social des Indépendants, ce régime a été supprimé en janvier 2020 et est devenu la SSI (Sécurité Sociale des Indépendants) lors de son adossement au régime général de la Sécurité Sociale après 2 ans de période transitoire.
- **Le régime agricole** couvre les exploitants et salariés agricoles et certaines autres professions rattachées à l'agriculture (ex : industrie agro-alimentaire). Ce régime est géré par la caisse centrale de la Mutualité Sociale Agricole (MSA) ;
- **Les régimes spéciaux** qui sont des régimes restreints à une profession ou entreprise particulière (ex : régime de la SNCF, de la RATP, des cultes, des militaires, régime Alsace-Moselle, ...). Il en existe plus de 200. Ils présentent une dimension historique et ont refusé de se fondre au régime général lors de sa création.

Plus spécifiquement, le régime Alsace-Moselle concerne les habitants de trois départements : le Bas-Rhin (67), le Haut-Rhin (68) et la Moselle (57). Appelé aussi régime local, il présente plus de 2,5 millions de bénéficiaires. Ses prestations sont plus avantageuses que celle du régime général de la Sécurité Sociale.

Retour
p.11

Les autres branches de la Sécurité Sociale (hors branche maladie)

La branche Accidents du travail et maladies professionnelles s'occupe des risques professionnels auxquels sont confrontés les travailleurs. Elle gère ainsi le système légal d'assurance des dommages corporels liés au travail (indemnisation des victimes et fixation de la contribution des entreprises au financement du système) et met en œuvre la politique de prévention des risques professionnels pour améliorer la sécurité et la santé des personnes actives. Elle est gérée par la CNAM.

La branche Vieillesse et veuvage (retraite) distribue les pensions versées aux retraités et celles de réversion aux conjoints survivants. Elle se décompose en plusieurs régimes fonctionnant pour la plupart par répartition (notamment les régimes obligatoires du pilier 1). Tous les régimes obligatoires de cette branche incluent des mécanismes de solidarité intergénérationnelle. Elle est gérée par la Caisse Nationale d'Assurance Vieillesse des Travailleurs Salariés (CNAVTS), parfois abrégée en Caisse Nationale d'Assurance Vieillesse (CNAV).

La branche Famille (dont logement, RSA, handicap, ...) s'occupe des prestations familiales dans un but d'atténuer les inégalités entre les ménages via quatre domaines prioritaires : l'accompagnement des familles dans leur vie quotidienne, l'accueil du jeune enfant, l'accès au logement et la lutte contre la précarité ou le handicap. Elle est gérée par la Caisse Nationale d'Allocations Familiales (CNAF) et localement par les caisses d'Allocations Familiales (CAF) éparpillées dans toute la France.

Les prestations de la branche Famille sont de deux types :

- Les prestations légales (aides financières prenant la forme de compléments de revenus ou de revenus de substitution comme le RSA) ;
- L'action sociale (complément aux prestations légales prenant plusieurs formes et s'adressant à l'ensemble des familles allocataires, notamment celles ayant des difficultés financières et sociales).

La branche Cotisation et recouvrement a plusieurs missions dont la collecte auprès des entreprises, des travailleurs indépendants et des particuliers des cotisations et contributions sociales. Elle les redistribue ensuite au bénéfice des autres branches pour financer l'ensemble des prestations.

Elle est gérée par l'Agence centrale des organismes de Sécurité Sociale (ACOSS) et son réseau se compose d'une vingtaine d'Unions de Recouvrement des cotisations de Sécurité Sociale et d'Allocations Familiales (URSSAF).

Retour
p.11

Financement

Quatre sources principales financent la Sécurité Sociale en France :

- Les cotisations sociales provenant des actifs (salariés et employeurs) et calculées sur la base du salaire des travailleurs ;
- La Contribution Sociale Généralisée (CSG) ;
- Les autres impôts et les taxes (de toute nature, y compris par exemple la TVA sur l'alcool ou les produits de pharmacie) ;
- D'autres sources telles qu'un transfert de l'État afin de compenser les pertes ou même des transferts interrégimes.

Retour
p.12

Synthèse

Voici ci-dessous un schéma tiré du site Internet du Ministère des solidarités et de la santé [3] qui reprend l'organisation de la Sécurité Sociale :

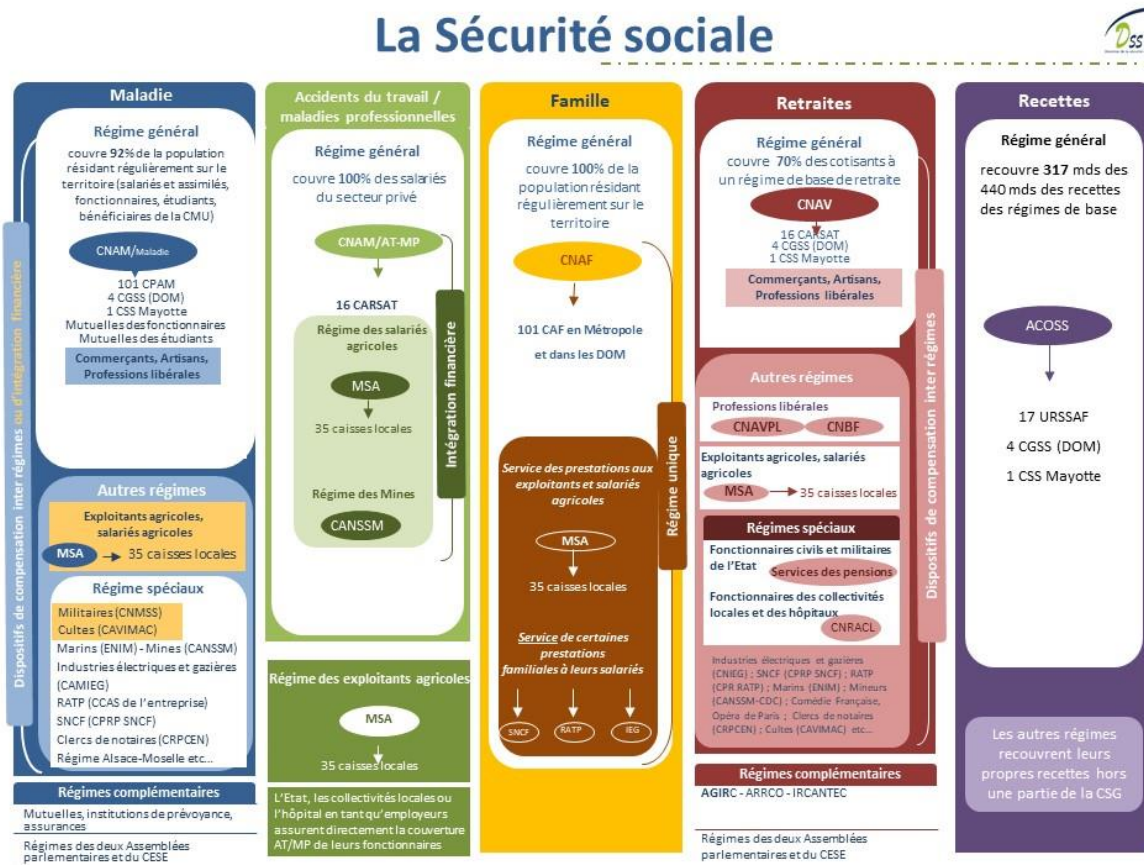


Schéma 5 : Organisation et structure de la Sécurité Sociale

Retour p.12

Annexe 2 : Fiche complémentaire sur les nomenclatures des actes

Exemple de calcul de la base de remboursement à partir de la CCAM

Code acte	Libellé	Tarif	Code de regroupement
MFQK002	Radiographie du coude selon 1 ou 2 incidences	19,95 €	ADI

Code acte	Libellé	Tarif	Commentaire
Modificateur E	Urgence	25,15 €	L'acte a été fait dans l'urgence
Modificateur U	Radiographie âge inférieur à 5 ans	49 %	Un enfant bouge beaucoup

$$BR = 19,95 \text{ €} * \underbrace{(1+49\%)}_{\text{Modificateur U}} + \underbrace{25,15 \text{ €}}_{\text{Modificateur E}} = 54,88 \text{ €}$$

Modificateur E Modificateur U

Retour
p.18

Groupe Homogène de séjour (GHS)

Le tarif du GHS comprend plusieurs choses :

- Dans le secteur public, cela comprend la rémunération des praticiens et auxiliaires médicaux, les prestations d'hébergement et les forfaits d'environnement technique telle que la location d'une salle d'opération ;
- Dans le secteur privé, cela comprend les prestations d'hébergement et les forfaits d'environnement technique seulement puisque le tarif du praticien libéral vient s'ajouter au tarif du GHS dans ce cas-ci.

Les autres nomenclatures

La Liste des Produits et Prestations Remboursables (LPPR) se présente sous la forme de liste recensant les fournitures et appareillages médicaux fixés par arrêtés ministériels (ex : pansements, prothèses auditives, ...) avec leur BR associée. Semblable à un catalogue, il n'y a pas ici de code particulier.

La base des médicaments peut se présenter de diverses manières selon si la classification se fait d'après les principes actifs directement (ex : hypertenseurs, vasodilatateurs, ...) ou d'après les dénominations communes internationales (DCI) (nom générique qui permet d'identifier au niveau mondial les principes actifs et les substances pharmaceutiques). L'identification du médicament s'effectue selon un code CIP (club inter-pharmaceutique) ou code UCD (unité commune de dispensation) utilisé pour la délivrance des médicaments dans les établissements de santé.

La Nomenclature des Actes de Biologie Médicale (NABM) qui définit les actes de biologie médicale comme son nom l'indique. Cette nomenclature est liée à celle de la NGAP et fonctionne d'une manière similaire.

Retour
p.19

Annexe 3 : Fiche complémentaire sur la législation santé

Dossier médical personnel (DMP)

Les quatre intervenants principaux sont :

- Le patient : détenteur du dossier, il y a accès via Internet. Il est libre de refuser son accès à certains professionnels de santé ;
- Les professionnels de santé : à partir d'un logiciel adapté, ils peuvent alimenter le dossier (d'analyses, de constatations, de résultats, ...)
- Le portail Internet d'accès au dossier : il permet d'enregistrer les accès côté patient et côté professionnel afin de garder un historique ;
- Les hébergeurs : ils reçoivent les données confidentielles et doivent être habilités par agrément de l'État.

Ce dossier permet de suivre les pathologies du patient dans le temps et de faciliter la coordination entre professionnels de santé. Il est coordonné par le médecin traitant.

Retour
p.19

CMU-C

La couverture maladie universelle (CMU) permet l'accès au soin et le remboursement des frais de santé à toute personne française ou étrangère sous deux conditions :

- Cette personne doit résider en France de manière stable et régulière – pendant au moins 3 mois ;
- Elle ne doit pas être couverte par un autre régime obligatoire.

Deux catégories ont été votées en juillet 1999 : la CMU de base et la CMU Complémentaire (CMU-C). La CMU de base a été instaurée dans le but de lutter contre l'exclusion aux soins. Elle peut être accordée de manière gratuite sous condition de ressources (il faut être en deçà d'un certain palier). Le remboursement est le même que celui de la Sécurité Sociale. Le 1^{er} janvier 2016, la CMU est remplacée par la PUMA (Protection Universelle Maladie).

La CMU-C complète la CMU de base (sans pour autant la remplacer) et fonctionne similairement à une mutuelle. C'est en fait une complémentaire santé pour des individus en situation précaire financièrement. Elle exige aussi d'être en deçà d'un certain palier financièrement et a été plus tard remplacée par la CSS (Complémentaire santé solidaire).

ACS

Comme son nom l'indique, l'Aide à la Complémentaire Santé (ACS) est versée à l'organisme assureur d'un individu qui en bénéficie afin de réduire le tarif du contrat santé choisi. Mise en place au 1^{er} janvier 2005, c'est donc avant tout une aide à la souscription de contrat d'assurance santé. Pour en bénéficier de manière gratuite, il faut satisfaire des conditions de ressources. Cette aide a été supprimée et a été remplacée par la CSS au 1^{er} novembre 2019.

Retour
p.21

Annexe 4 : Fiche complémentaire sur l'état de l'art (Open data)

Titre :	Les Comptes nationaux de la santé en 2013
Source :	https://www.data.gouv.fr/fr/datasets/les-comptes-nationaux-de-la-sante-en-2013/
Contenu :	Des informations générales sur la CSBM (valeur, évolution sur les différents grands postes de soin et biens médicaux, structure de financement, la prise en charge par les différents organismes, des informations sur les pays étrangers, ...)
-	Les comptes nationaux de la santé en 2010, 2011, 2013 sont disponibles mais pas pour les années récentes. Le cadre législatif a significativement changé depuis 2013 et ces comptes devraient être différents maintenant (notamment avec la réforme 100 % santé qui change le paysage de l'optique, l'audio et le dentaire).

Titre :	Annuaire santé de la Cnam
Source :	https://www.data.gouv.fr/fr/datasets/annuaire-sante-de-la-cnam/
Contenu :	Les coordonnées des professionnels de santé exerçant à titre libéral et celles des établissements de soins, les actes pratiqués, le secteur conventionnel auquel appartient un professionnel de santé, les tarifs pratiqués, s'il accepte ou non la carte vitale, les données tarifaires pour certaines prestations d'hospitalisation, ...
-	Au premier abord, cette base serait plus utile à un assuré qui recherche un médecin qu'à un assureur ; mais un assuré ordinaire aurait plutôt tendance à faire une simple recherche sur Internet plutôt que s'aventurer dans un fichier Excel.
+	En revanche, pour ses informations de localisation des praticiens, de leurs tarifs et de leur secteur conventionnel, il est possible qu'elle puisse servir d'aide ou de bases statistiques dans le cas d'une problématique d'enrichissement d'une base de données préexistante en colonne (ajout d'une variable) car le tarif d'un médecin dépend de sa zone géographique.

Tableau 46 : Suite de l'état de l'art des *Open data* en santé

De nombreuses autres bases existent, parmi elles :

- Des bases de données sur les médicaments avec
 - La base de données publique des médicaments dont la dernière mise à jour date du 16 mars 2016 qui permet au public et aux professionnels de santé d'accéder à des données de référence (administratives, scientifiques, ...) et à des documents de référence sur les médicaments commercialisés à partir de 2014 ;
 - L'Open Medic (base complète sur les dépenses de médicaments interrégimes) qui est l'homologue de l'Open DAMIR mais pour les médicaments et qui fournit des informations complémentaires au fichier Medic'AM (Médicaments remboursés par l'Assurance Maladie).
 - L'Open PHMEV (bases sur les prescriptions hospitalières de médicaments délivrées en ville) qui est constituée de bases annuelles interrégimes, portant sur les remboursements de médicaments prescrits par les établissements publics et ESPIC (établissements de santé privés d'intérêt collectif) et délivrés en officine de ville de 2014 à 2017.
- Des bases de données portant sur le domaine de la perception et du qualitatif comme le baromètre d'opinion de la DREES qui suit sur une longue période la perception de la protection sociale et les avis sur la santé, la cohésion sociale et les inégalités en France ;
- Des bases de données sur les renoncements aux soins comme celle sur l'état de santé et le renoncement aux soins des bénéficiaires du RSA ;
- Des bases de données relatives aux décès comme la Base des Causes Médicales de Décès du CépiDc, le Centre d'épidémiologie sur les causes médicales de décès (qui fait partie de l'INSERM) et qui recense toutes les causes médicales de décès à partir des certificats de décès complétés par les médecins ;

Annexe 5 : Les prestations de VirtuaMut' et de la Sécurité Sociale

Précisions :

Il est à noter qu'il n'y a pas eu d'évolution dans les grilles de garantie entre 2018 et 2019. De ce fait, les grilles suivantes valent pour les deux années d'études.

Pour les actes de consultations, de visites, les actes médicaux techniques, les actes d'imagerie, de radiologie et d'ostéodensitométrie et les honoraires et actes chirurgicaux, ont été retenus les remboursements en vigueur pour un praticien ayant adhéré à l'OPTAM par prudence (la mutuelle rembourse plus dans ce cas-ci).

Sauf mention de « (FR) » dans certaines lignes de prestations (pour « Frais réels »), les pourcentages sont exprimés en fonction de la BR.

Tableau 47 : Grilles de garanties simplifiées

		Garantie A			Garantie B		
		% SS	% Mutuelle	Limite	% SS	% Mutuelle	Limite
Soins courants	Consultations/visites	70%	30%		70%	70%	
	Actes médicaux techniques	70%	30%		70%	90%	
	Auxiliaires médicaux (kinésithérapeutes, infirmiers, ...)	60%	40%		60%	100%	
	Analyses médicales et examens laboratoire	60%	40%		60%	100%	
	Actes d'imagerie, de radiologie et ostéodensitométrie	70%	30%		70%	90%	
	Petit appareillage	60%	90%		60%	240%	
	Pharmacie	Variable (15/30/65% pour la SS, soit 85/70/35% pour la mutuelle)			Variable (15/30/65% pour la SS, soit 85/70/35% pour la mutuelle)		
Hospitalisation	Frais de séjour	80%	20%		80%	20%	
	Honoraires et actes chirurgicaux	80%	0%		80%	120%	800 €
	Forfait journalier	0%	100% (FR)		0%	100% (FR)	
Dentaire	Prothèse	70%	130%	1 100 €	70%	130%	1 300 €
	Soins dentaires	70%	30%		70%	90%	
Optique	Monture	60%	80% (FR)	120 €	60%	80% (FR)	150 €
	Verres	Variables selon type de verres			Variables selon type de verres		
	Chirurgie œil	0%	0%		0%	500 €	
	Lentilles	60%	200 €		60%	260 €	
Aides auditives	Audioprothèse (<i>par oreille</i>)	60%	325 €		60%	600 €	
	Pile, accessoire	60%	30 €		60%	80 €	

Retour
p.33

Retour
p.50

Annexe 6 : Compléments sur le traitement des données provenant de la mutuelle VirtuaMut'

6.1. Compléments sur le retraitement 2.1.2.2. Rajout des informations sur les adhérents

La variable **OPTION** de la base des adhérents a servi à identifier correctement l'individu dans le cas où ce dernier apparaissait plusieurs fois dans la base des adhérents avec des souscriptions de garanties successives différentes.

Une fois la variable **REGIME** utilisée comme filtre, elle devient alors non informative et elle est aussi supprimée de la base de données.

Dans le cas de la garantie A, 3 adhérents n'ont pas pu être identifiés car la variable **REF_PERSONNE** avait été mal renseignée et 19 lignes de prestations ont dû être supprimées (l'impact est non significatif comme ces prestations représentent moins de 0,1 % des prestations totales). 2 adhérents présentaient une erreur dans leur identifiant mais leur référence était suffisamment proche d'une référence existante pour que nous puissions les modifier directement. Dans le cas de la garantie B, 1 adhérent n'a pas été identifié et 4 lignes de prestations ont alors été supprimées ; 2 références ont pu être corrigées.

Retour
p.37

6.2. Retraitement de la variable SEXE

Conformément à son homologue dans la base Open DAMIR, la variable **SEXE** prend la valeur « 1 » pour le sexe masculin et la valeur « 2 » pour le sexe féminin.

Retour
p.37

6.3. Compléments sur le Retraitement (2.1.2.6.) de la variable QTE ACTE

Exemples de retraitements réalisés :

Par exemple, si pour une audioprothèse, la dépense engagée était proche du prix moyen estimé à 1 500€ en 2018, la quantité associée est mise à 1 quand bien même celle renseignée est à 1,9 ou 1,3. Pour les médicaments, il est tout à fait cohérent d'afficher une quantité supérieure à 1 comme il est courant d'acheter plusieurs médicaments en pharmacie en une seule visite. Autre exemple, les forfaits journaliers, quant à eux, sont souvent décomptés en jour et seule la date de début d'hospitalisation est inscrite, la quantité entière renseignée est dans ce cas-là gardée.

Retour
p.38

6.4. Agrégation des lignes par individu

Avant l'agrégation faite pour égaliser avec la granularité de l'Open DAMIR, une première agrégation a été réalisée : celle des lignes par individus sur les différents mois. Cela permet d'effectuer une première sauvegarde des travaux sur une base mensuelle. Ainsi, les différents montants (dépense, montant RO, montant RC) et les quantités d'actes sont agrégés selon toutes les autres variables en faisant abstraction des jours de règlement de sinistres et de soins délivrés. Suite à cela, une ligne indiquera désormais les différents montants de dépenses et de remboursements ainsi que les quantités d'actes dans le cadre d'un code acte en particulier, sur un mois particulier, pour un adhérent de la mutuelle. La base de données relative à une garantie donnée est donc toujours ici une base tête par tête, mais l'information n'est plus écrite de manière journalière. Pour rappel, la granularité de l'information dans l'Open DAMIR est mensuelle.

Retour
p.40

6.5. Compléments sur la création de la variable EXPO_TOTALE (2.1.2.12.)

Pour déterminer les valeurs de EXPO_2018 et EXPO_2019, il faut considérer plusieurs situations :

- Soit, l'adhérent apparaît plusieurs fois dans la base des adhérents, il souscrit à la même garantie (A ou B) à chaque fois et il n'y a pas de discontinuité apparente entre la date de fin de son premier contrat et la date de début de son second contrat (idem pour entre le deuxième et troisième contrat mais c'est un cas rare). Dans ce cas-là, DATE_DEBUT prend la valeur de la date la plus récente entre la date de début de son premier contrat et la date de début de l'année de soin i considérée (01/01/2018 ou 01/01/2019). Et DATE_FIN prend la valeur la moins récente entre la date de fin de son dernier contrat et la date de fin de l'année de soin i considérée (31/12/2018 ou 31/12/2019).
Dans le cas où le contrat n'est pas encore terminé et que la variable DATE_FIN est vide, cette variable est forcée à la date limite de l'année de soin i considérée.
- Soit, l'adhérent apparaît plusieurs fois dans la base des adhérents, il souscrit à la même garantie à chaque fois (A ou B) mais il y a discontinuité des dates, c'est-à-dire qu'il existe une période située entre le 01/01/2018 et le 31/12/2019 pendant laquelle l'adhérent n'est plus couvert par VirtuaMut'. Dans ce cas-là, les variables DATE_DEBUT et DATE_FIN prennent des valeurs forcées de telle sorte que le nombre de jours de présence de l'adhérent en 2018 soit conservé (et idem pour 2019) et qu'il n'y ait plus de dates discontinues.
- Soit, l'adhérent n'apparaît qu'une fois dans la base des adhérents avec la garantie A ou B, il faut se ramener au premier cas sans la nécessité de distinguer les contrats puisqu'un seul contrat a été souscrit ici au titre de la garantie en question.

Il faut aussi s'assurer que DATE_FIN soit plus récente que DATE_DEBUT.

Quelques exemples pour mieux comprendre ce qui est concrètement fait :

Exemple 1 : Selon la base des adhérents, l'adhérent Emma (code de référence 63194) a souscrit :

- À la garantie A du 01/01/2010 au 30/11/2018
- À la garantie A du 01/12/2018 à « » (contrat en cours)

Il n'y a pas de discontinuité de dates.

Pour déterminer la valeur prise par EXPO_2018 :

- DATE_DEBUT = 01/01/2018 car la date de début de son premier contrat (01/01/2010) est moins récente que la date de début de notre année de soin (2018)
- DATE_FIN = 31/12/2018 car la date de fin de son dernier contrat n'est pas encore déterminée (le contrat étant encore en cours après le 31/12/2018).

Par ailleurs, dans ce cas, EXPO_2019 est égale à 1.

Exemple 2 : Selon la base des adhérents, l'adhérent Ray (code de référence 81194) a souscrit :

- À la garantie A du 01/01/2010 au 30/04/2019
- À la garantie A du 01/07/2019 à « » (contrat en cours)

Il y a discontinuité de dates.

Pour déterminer la valeur prise par EXPO_2019 :

- DATE_DEBUT = 01/01/2019 car la date de début de son premier contrat (01/01/2010) est moins récente que la date de début de notre année de soin (2019).
- DATE_FIN = 31/10/2019 car la date de fin de son dernier contrat n'est pas encore déterminée (le contrat étant encore en cours après le 31/12/2019) mais comme il y a une discontinuité de dates entre les deux contrats, afin de garder une présence de Ray à 10 mois sur 12 dans l'année 2019 sans discontinuité, il y a un décalage de deux mois (DATE_FIN ne prend donc pas la valeur du 31/12/2019 mais celle du 31/10/2019, deux mois avant).

Par ailleurs, dans ce cas, EXPO_2018 vaut 1.

Exemple 3 : Selon la base des adhérents, l'adhérent Norman (code de référence 22194) a souscrit à la garantie A du 01/04/2019 au 30/09/2019. Il n'y a pas de discontinuité de dates.

Pour déterminer la valeur prise par EXPO_2019 :

- DATE_DEBUT = 01/04/2019 car la date de début de son contrat (01/04/2019) est plus récente que la date de début de notre année de soin (01/01/2019).
- DATE_FIN = 30/09/2019 car la date de fin de son contrat est moins récente que la date de fin de notre année de règlement (31/12/2019).

Par ailleurs, dans ce cas, EXPO_2018 = 0.

Retour
p.41

Annexe 7 : Liste des variables de l'Open DAMIR

Le tableau ci-dessous est une extraction de l'onglet « Variable par axe analyse » du lexique de l'Open DAMIR. La variable Catégorie a été rajoutée et son remplissage a été fait en confrontant nos propres considérations avec celles prises par Arnold MEKONTSO dans son mémoire *L'open DAMIR : apport à la maîtrise des dépenses de santé* (2018).

Axe	Variable	Libellé	Catégorie
<i>Tables A à partir de 2015</i>			
PERIODE DE TRAITEMENT			
	FLX_ANN_MOI	Année et Mois de Traitement	qualitative ordinale
PRESTATION			
	PRS_NAT	Nature de Prestation	qualitative nominale
	ASU_NAT	Nature d'Assurance	qualitative nominale
	ATT_NAT	Nature de l'Accident du Travail	qualitative nominale
	CPT_ENV_TYP	Type d'Enveloppe	qualitative nominale
	CPL_COD	Complément d'Acte	qualitative nominale
	EXO_MTF	Motif d'Exonération du Ticket Modérateur	qualitative nominale
	PRS_REM_TAU	Taux de Remboursement	qualitative nominale
	PRS_PPU_SEC	Code Secteur Privé/Public	qualitative nominale
	PRS_FJH_TYP	Type de Prise en Charge Forfait Journalier	qualitative nominale
	ETE_IND_TAA	Indicateur TAA Privé/Public	qualitative nominale
	PRS_PDS_QCP	Code Qualificatif Parcours de Soins (sortie)	qualitative nominale
	DRG_AFF_NAT	Nature du Destinataire de Règlement affiné	qualitative nominale
	PRS_REM_TYP	Type de Remboursement	qualitative nominale
ORGANISME			
	ORG_CLE_REG	Région de l'Organisme de Liquidation à partir de 2015 (tables A)	qualitative nominale
PERIODE			
	SOI_ANN	Année de Soins	qualitative ordinale
	SOI_MOI	Mois de Soins	qualitative ordinale
BENEFICIAIRE			
	BEN_SEX_COD	Sexe du Bénéficiaire	qualitative nominale
	AGE_BEN_SNDS	Tranche d'Age Bénéficiaire au moment des soins	qualitative ordinale
	BEN_QLT_COD	Qualité du Bénéficiaire	qualitative nominale
	BEN_RES_REG	Région de Résidence du Bénéficiaire à partir de 2015 (tables A)	qualitative nominale
	MTM_NAT	Modulation du Ticket Modérateur	qualitative nominale
	BEN_CMU_TOP	Top Bénéficiaire CMU-C	qualitative nominale
EXECUTANT			
	PSE_ACT_CAT	Catégorie de l'Exécutant	qualitative nominale
	PSE_SPE_SNDS	Spécialité Médicale PS Exécutant	qualitative nominale
	PSE_ACT_SNDS	Nature d'Activité PS Exécutant	qualitative nominale
	EXE_INS_REG	Région du PS Exécutant à partir de 2015 (tables A)	qualitative nominale
	PSE_STJ_SNDS	Statut Juridique PS Exécutant	qualitative nominale
	MFT_COD	Mode de Fixation des Tarifs Etb Exécutant	qualitative nominale
	ETE_REG_COD	Région d'Implantation Etb Exécutant à partir de 2015 (tables A)	qualitative nominale

ETE_TYP_SNDS	Type Etb Exécutant	qualitative nominale
ETE_CAT_SNDS	Catégorie Etb Exécutant	qualitative nominale
DDP_SPE_COD	Discipline de Prestation Etb Exécutant	qualitative nominale
MDT_TYP_COD	Mode de Traitement Etb Exécutant	qualitative nominale
PRESCRIPTEUR		
PSP_ACT_CAT	Catégorie du Prescripteur	qualitative nominale
PSP_SPE_SNDS	Spécialité Médicale PS Exécutant	qualitative nominale
PSP_ACT_SNDS	Nature d'Activité PS Prescripteur	qualitative nominale
PRE_INS_REG	Région du PS Prescripteur à partir de 2015 (tables A)	qualitative nominale
PSP_STJ_SNDS	Statut Juridique PS Prescripteur	qualitative nominale
ETP_REG_COD	Région d'Implantation Etb Prescripteur à partir de 2015 (tables A)	qualitative nominale
ETP_CAT_SNDS	Catégorie Etb Prescripteur	qualitative nominale
TOP PS5		
TOP_PS5_TRG	Top Périmètre hors CMU C et prestations pour information	qualitative nominale
INDICATEURS		
FLT_ACT_COG	Coefficient Global de la Prestation Préfiltré	quantitative continue
FLT_ACT_NBR	Dénombrement de la Prestation Préfiltré	quantitative discrète
FLT_ACT_QTE	Quantité de la Prestation Préfiltrée	quantitative discrète
FLT_DEP_MNT	Montant du Dépassement de la Prestation Préfiltré	quantitative continue
FLT_PAI_MNT	Montant de la Dépense de la Prestation Préfiltrée	quantitative continue
FLT_REM_MNT	Montant Versé/Remboursé Préfiltré	quantitative continue
PRS_ACT_COG	Coefficient Global	quantitative continue
PRS_ACT_NBR	Dénombrement	quantitative discrète
PRS_ACT_QTE	Quantité	quantitative discrète
PRS_DEP_MNT	Montant du Dépassement	quantitative continue
PRS_PAI_MNT	Montant de la Dépense	quantitative continue
PRS_REM_MNT	Montant Versé/Remboursé	quantitative continue
PRS_REM_BSE	Base de Remboursement	quantitative continue

Tableau 48 : Liste des variables de l'Open DAMIR

Retour
p.29

Retour
p.44

Annexe 8 : Liste des variables retenues ou supprimées (étape 1)

Axe	Variable	Libellé	Catégorie	Commentaire
PERIODE DE TRAITEMENT				
	FLX_ANN_MOI	Année et Mois de Traitement	Gardée	
PRESTATION				
	PRS_NAT	Nature de Prestation	Gardée	
	ASU_NAT	Nature d'Assurance	Gardée	Permet de filtrer sur la branche Maladie
	ATT_NAT	Nature de l'Accident du Travail	Supprimée	La branche AT/DC (prévoyance) n'entre pas dans le cadre des études
	CPT_ENV_TYP	Type d'Enveloppe	Gardée	Permet de filtrer sur le régime général
	CPL_COD	Complément d'Acte	Gardée	À réfléchir sur sa pertinence
	EXO_MTF	Motif d'Exonération du Ticket Modérateur	Supprimée	Non pertinente
	PRS_REM_TAU	Taux de Remboursement	Gardée	Pour déduire un montant RO
	PRS_PPU_SEC	Code Secteur Privé/Public	Supprimée	Non pertinente
	PRS_FJH_TYP	Type de Prise en Charge Forfait Journalier	Gardée	À réfléchir sur sa pertinence
	ETE_IND_TAA	Indicateur TAA Privé/Public	Supprimée	Non pertinente
	PRS_PDS_QCP	Code Qualificatif Parcours de Soins (sortie)	Supprimée	Non pertinente (maille trop fine)
	DRG_AFF_NAT	Nature du Destinataire de Règlement affiné	Supprimée	Non pertinente (maille trop fine)
	PRS_REM_TYP	Type de Remboursement	Gardée	Pour vérifier des quantités qui devraient être à 0
ORGANISME				
	ORG_CLE_REG	Région de l'Organisme de Liquidation à partir de 2015 (tables A)	Supprimée	Non pertinente
PERIODE				
	SOI_ANN	Année de Soins	Gardée	
	SOI_MOI	Mois de Soins	Gardée	
BENEFICIAIRE				
	BEN_SEX_COD	Sexe du Bénéficiaire	Gardée	
	AGE_BEN_SNDS	Tranche d'Age Bénéficiaire au moment des soins	Gardée	
	BEN_QLT_COD	Qualité du Bénéficiaire	Gardée	
	BEN_RES_REG	Région de Résidence du Bénéficiaire à partir de 2015 (tables A)	Gardée	
	MTM_NAT	Modulation du Ticket Modérateur	Supprimée	Non pertinente
	BEN_CMU_TOP	Top Bénéficiaire CMU-C	Gardée	Permet de filtrer sur les non bénéficiaires du CMU-C
EXECUTANT				
	PSE_ACT_CAT	Catégorie de l' Exécutant	Supprimée	Non pertinente
	PSE_SPE_SNDS	Spécialité Médicale PS Exécutant	Supprimée	Non pertinente
	PSE_ACT_SNDS	Nature d'Activité PS Exécutant	Supprimée	Non pertinente
	EXE_INS_REG	Région du PS Exécutant à partir de 2015 (tables A)	Gardée	À titre de sauvegarde
	PSE_STJ_SNDS	Statut Juridique PS Exécutant	Supprimée	Non pertinente
	MFT_COD	Mode de Fixation des Tarifs Etb Exécutant	Gardée	À titre de sauvegarde
	ETE_REG_COD	Région d'Implantation Etb Exécutant à partir de 2015 (tables A)	Supprimée	Non pertinente
	ETE_TYP_SNDS	Type Etb Exécutant	Supprimée	Non pertinente
	ETE_CAT_SNDS	Catégorie Etb Exécutant	Supprimée	Non pertinente
	DDP_SPE_COD	Discipline de Prestation Etb Exécutant	Supprimée	Non pertinente
	MDT_TYP_COD	Mode de Traitement Etb Exécutant	Supprimée	Non pertinente
PRESCRIPTEUR				
	PSP_ACT_CAT	Catégorie du Prescripteur	Supprimée	Non pertinente
	PSP_SPE_SNDS	Spécialité Médicale PS Exécutant	Supprimée	Non pertinente
	PSP_ACT_SNDS	Nature d'Activité PS Prescripteur	Supprimée	Non pertinente

PRE_INS_REG	Région du PS Prescripteur à partir de 2015 (tables A)	Supprimée	Non pertinente
PSP_STJ_SNDS	Statut Juridique PS Prescripteur	Supprimée	Non pertinente
ETP_REG_COD	Région d'Implantation Etb Prescripteur à partir de 2015 (tables A)	Supprimée	Non pertinente
ETP_CAT_SNDS	Catégorie Etb Prescripteur	Supprimée	Non pertinente
TOP PS5			
TOP_PS5_TRG	Top Périmètre hors CMU C et prestations pour information	Gardée	Permet d'enlever les prestations non pertinentes comme celles issues de la CMU-C
INDICATEURS			
FLT_ACT_COG	Coefficient Global de la Prestation Préfiltré	Gardée	
FLT_ACT_NBR	Dénombrement de la Prestation Préfiltré	Gardée	
FLT_ACT_QTE	Quantité de la Prestation Préfiltrée	Gardée	
FLT_DEP_MNT	Montant du Dépassement de la Prestation Préfiltré	Gardée	
FLT_PAI_MNT	Montant de la Dépense de la Prestation Préfiltrée	Gardée	
FLT_REM_MNT	Montant Versé/Remboursé Préfiltré	Gardée	
PRS_ACT_COG	Coefficient Global	Supprimée	Seul le régime obligatoire est étudié
PRS_ACT_NBR	Dénombrement	Supprimée	Seul le régime obligatoire est étudié
PRS_ACT_QTE	Quantité	Supprimée	Seul le régime obligatoire est étudié
PRS_DEP_MNT	Montant du Dépassement	Supprimée	Seul le régime obligatoire est étudié
PRS_PAI_MNT	Montant de la Dépense	Supprimée	Seul le régime obligatoire est étudié
PRS_REM_MNT	Montant Versé/Remboursé	Supprimée	Seul le régime obligatoire est étudié
PRS_REM_BSE	Base de Remboursement	Gardée	

Tableau 49 : Choix de traitement des variables (étape 1)

Annexe 9 : Tableau de correspondance des régions entre VirtuaMut' et Open DAMIR

Département (Wikipédia) Code	Région administrative	Open DAMIR Codage
1	Auvergne-Rhône-Alpes	84
2	Hauts-de-France	32
3	Auvergne-Rhône-Alpes	84
4	Provence-Alpes-Côte d'Azur	93
5	Provence-Alpes-Côte d'Azur	93
6	Provence-Alpes-Côte d'Azur	93
7	Auvergne-Rhône-Alpes	84
8	Grand Est	44
9	Occitanie	76
10	Grand Est	44
11	Occitanie	76
12	Occitanie	76
13	Provence-Alpes-Côte d'Azur	93
14	Normandie	28
15	Auvergne-Rhône-Alpes	84
16	Nouvelle-Aquitaine	75
17	Nouvelle-Aquitaine	75
18	Centre-Val de Loire	24
19	Nouvelle-Aquitaine	75
2A	Corse	93
2B	Corse	93
21	Bourgogne-Franche-Comté	27
22	Bretagne	53
23	Nouvelle-Aquitaine	75
24	Nouvelle-Aquitaine	75
25	Bourgogne-Franche-Comté	27
26	Auvergne-Rhône-Alpes	84
27	Normandie	28
28	Centre-Val de Loire	24
29	Bretagne	53
30	Occitanie	76
31	Occitanie	76
32	Occitanie	76
33	Nouvelle-Aquitaine	75
34	Occitanie	76
35	Bretagne	53
36	Centre-Val de Loire	24
37	Centre-Val de Loire	24
38	Auvergne-Rhône-Alpes	84
39	Bourgogne-Franche-Comté	27
40	Nouvelle-Aquitaine	75
41	Centre-Val de Loire	24
42	Auvergne-Rhône-Alpes	84
43	Auvergne-Rhône-Alpes	84
44	Pays de la Loire	52
45	Centre-Val de Loire	24
46	Occitanie	76
47	Nouvelle-Aquitaine	75
48	Occitanie	76
49	Pays de la Loire	52
50	Normandie	28

Correspondance avec l'Open DAMIR :

BEN_RES_REG	Libellé Région de Résidence du Bénéficiaire
5	Régions et Départements d'outre-mer
11	Ile-de-France
24	Centre-Val de Loire
27	Bourgogne-Franche-Comté
28	Normandie
32	Hauts-de-France - Nord-Pas-de-Calais-Picardie
44	Grand Est
52	Pays de la Loire
53	Bretagne
75	Aquitaine-Limousin-Poitou-Charentes
76	Languedoc-Roussillon-Midi-Pyrénées
84	Auvergne-Rhône-Alpes
93	Provence-Alpes-Côte d'Azur et Corse
99	Inconnu

51	Grand Est	44
52	Grand Est	44
53	Pays de la Loire	52
54	Grand Est	44
55	Grand Est	44
56	Bretagne	53
57	Grand Est	44
58	Bourgogne-Franche-Comté	27
59	Hauts-de-France	32
60	Hauts-de-France	32
61	Normandie	28
62	Hauts-de-France	32
63	Auvergne-Rhône-Alpes	84
64	Nouvelle-Aquitaine	75
65	Occitanie	76
66	Occitanie	76
67	Grand Est	44
68	Grand Est	44
69	Auvergne-Rhône-Alpes	84
70	Bourgogne-Franche-Comté	27
71	Bourgogne-Franche-Comté	27
72	Pays de la Loire	52
73	Auvergne-Rhône-Alpes	84
74	Auvergne-Rhône-Alpes	84
75	Île-de-France	11
76	Normandie	28
77	Île-de-France	11
78	Île-de-France	11
79	Nouvelle-Aquitaine	75
80	Hauts-de-France	32
81	Occitanie	76
82	Occitanie	76
83	Provence-Alpes-Côte d'Azur	93
84	Provence-Alpes-Côte d'Azur	93
85	Pays de la Loire	52
86	Nouvelle-Aquitaine	75
87	Nouvelle-Aquitaine	75
88	Grand Est	44
89	Bourgogne-Franche-Comté	27
90	Bourgogne-Franche-Comté	27
91	Île-de-France	11
92	Île-de-France	11
93	Île-de-France	11
94	Île-de-France	11
95	Île-de-France	11
97	Guadeloupe	5
97	Martinique	5
97	Guyane	5
97	La Réunion	5
97	Mayotte	5

Tableau 50 : Correspondance des régions

Annexe 10 : Classification des actes de l'Open DAMIR selon la segmentation retenue

L'Open DAMIR présentant un grand nombre de codes actes (1 080), afin de limiter l'espace pris pour exposer notre classification dans cette annexe, nous procédons à la numérotation des sous-familles d'actes retenues :

Grands postes	Familles d'actes	Sous-familles d'actes	ID
Soins courants	Honoraires médicaux	Consultations/visites	1
		Actes médicaux (techniques)	2
		Autres honoraires médicaux	3
	Honoraires paramédicaux	Auxiliaires médicaux (kinésithérapeutes, infirmiers, ...)	4
	Analyse et examen de laboratoire	Analyse médicale et examens laboratoire	5
	Imagerie médicale	Actes d'imagerie, de radiologie et ostéodensitométrie	6
	Transport, ambulances	Transport	7
	Pharmacie	Petit appareillage	8
		Grand appareillage	9
		Pharmacie	10
		Vaccins anti-grippes	11
Hospitalisation	Hospitalisation	Frais de séjour	12
		Honoraires et actes chirurgicaux	13
		Forfait journalier	14
		Chambre particulière	15
		Lit accompagnant	16
		Autres - hospitalisation	17
Dentaire	Dentaire	Prothèse	18
		Soins dentaires	19
		Parodontologie	20
		Implantologie	21
		Orthodontie	22
		Autres - dentaire	23
		Optique	Optique
Verres	25		
Chirurgie œil	26		
Lentilles	27		
Autres - Optique	28		
Aides auditives	Audio	Audioprothèse	29
		Pile, accessoire	30
		Autres - audio	31
Autres	Cure thermique	Cure thermique	32
	Médecine douce	Ostéopathie, diététique, chiropractie, ...	33
	Prévention	Prévention	34
	Prestations supplémentaires	Maternité	35
		Autres (aides, inclassable, ...)	36

Tableau 51 : Segmentation des actes de soin

Dans certains cas d'actes d'hospitalisation où la sous-famille était bien trop incertaine, nous avons classé l'acte par sa famille d'actes uniquement (la sous-famille d'actes prend alors une valeur égale à la famille d'actes Hospitalisation). Nous identifions donc « Hospitalisation » par le numéro « 37 ».

Voici donc ci-dessous, à chaque valeur de PRS_NAT, son identifiant dans notre table de segmentation :

PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID
0	36	1102	1	1109	1	1116	1	1123	2	1130	36	1137	1
1096	1	1103	1	1110	1	1117	1	1124	36	1131	1	1138	36
1097	1	1104	1	1111	1	1118	1	1125	1	1132	1	1139	36
1098	1	1105	1	1112	1	1119	1	1126	1	1133	1	1140	1
1099	1	1106	1	1113	1	1120	1	1127	1	1134	1	1141	1
1100	1	1107	1	1114	1	1121	37	1128	36	1135	1	1142	36
1101	1	1108	1	1115	1	1122	2	1129	1	1136	1	1143	36

PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID
1144	36	1228	1	1434	19	1644	36	1908	36	2103	7	2167	17
1145	36	1229	1	1435	36	1645	7	1909	36	2104	7	2168	17
1146	36	1231	1	1436	36	1646	36	1910	36	2105	7	2170	36
1147	36	1232	1	1437	1	1647	36	1911	1	2106	7	2173	1
1148	36	1311	13	1451	19	1648	36	1912	3	2107	7	2176	19
1149	36	1312	2	1452	20	1649	36	1913	1	2108	7	2181	37
1150	36	1313	2	1453	19	1650	36	1914	1	2109	7	2182	37
1152	1	1314	13	1461	18	1651	36	1915	36	2111	12	2183	37
1153	1	1315	13	1462	18	1652	36	1916	36	2112	12	2184	37
1154	1	1316	2	1463	18	1653	36	1917	36	2113	12	2185	37
1155	36	1317	13	1464	18	1701	36	1918	1	2114	37	2186	37
1156	36	1318	2	1465	21	1702	36	1920	36	2115	37	2187	37
1157	1	1319	37	1466	23	1703	36	1921	36	2116	37	2188	37
1158	1	1320	6	1470	19	1704	36	1922	1	2117	37	2189	37
1159	1	1321	13	1471	18	1705	36	1923	1	2118	37	2190	37
1160	36	1322	2	1472	20	1706	36	1924	11	2119	37	2191	37
1161	1	1323	13	1473	18	1707	36	1925	11	2120	37	2195	36
1162	36	1324	6	1474	18	1708	36	1926	11	2121	17	2202	37
1163	36	1331	6	1475	18	1711	36	1931	1	2122	17	2204	36
1164	1	1332	6	1476	18	1712	36	1932	1	2123	17	2206	37
1165	1	1333	10	1477	23	1718	36	1933	1	2124	17	2211	12
1166	1	1335	6	1511	37	1721	36	1934	1	2125	17	2212	37
1167	36	1336	6	1521	36	1722	36	1935	1	2126	17	2213	12
1168	1	1341	5	1522	36	1724	36	1936	1	2127	17	2214	36
1169	36	1342	5	1523	36	1727	36	1937	1	2128	17	2215	36
1170	1	1345	1	1602	36	1728	36	1938	1	2129	17	2221	15
1171	36	1351	6	1603	36	1729	36	1939	1	2131	17	2222	17
1172	36	1352	2	1605	36	1735	36	1940	14	2132	17	2223	17
1173	36	1361	36	1606	36	1744	36	1941	1	2133	17	2224	37
1174	36	1400	18	1607	36	1745	36	1942	1	2134	17	2225	37
1175	1	1401	18	1608	36	1763	36	1943	1	2135	17	2227	37
1176	36	1402	18	1609	36	1764	36	1944	1	2136	17	2229	37
1177	36	1403	23	1610	36	1766	36	1945	1	2137	17	2230	36
1178	36	1404	23	1611	36	1767	36	1951	36	2138	17	2231	17
1179	36	1405	23	1612	36	1771	36	1952	36	2139	17	2232	37
1180	36	1406	18	1613	36	1781	36	1954	37	2140	17	2234	17
1181	36	1407	18	1614	36	1782	36	1955	36	2141	15	2235	37
1182	36	1408	18	1615	36	1783	36	1956	37	2142	17	2236	37
1183	36	1409	18	1616	36	1784	36	1957	36	2143	17	2237	36
1188	36	1410	18	1617	36	1785	36	1960	36	2144	17	2238	1
1191	1	1411	19	1618	36	1786	36	1961	36	2145	17	2239	36
1192	1	1412	18	1619	36	1787	36	1971	36	2146	17	2240	37
1193	1	1413	23	1621	36	1788	36	1972	36	2147	17	2241	37
1194	1	1414	18	1622	36	1811	7	1973	36	2150	6	2242	37
1196	36	1415	18	1623	36	1812	7	1974	36	2151	17	2243	37
1209	1	1416	18	1624	36	1813	7	1975	36	2152	17	2245	37
1210	1	1417	18	1625	36	1814	7	1976	36	2153	17	2246	37
1211	1	1418	18	1627	36	1821	1	1977	36	2154	3	2247	37
1212	1	1419	18	1628	36	1841	4	1978	36	2155	17	2248	36
1213	1	1420	18	1629	36	1842	4	1981	37	2156	17	2249	37
1214	1	1421	18	1630	36	1843	4	1990	36	2157	17	2250	14
1215	1	1422	22	1631	36	1844	4	1991	36	2158	17	2251	14
1216	1	1423	18	1632	36	1845	4	1992	36	2159	1	2252	14
1221	1	1424	22	1633	36	1846	4	1993	36	2160	36	2257	14
1222	1	1425	18	1634	36	1847	4	1994	36	2161	36	2258	14
1223	1	1426	23	1636	36	1903	1	1995	36	2162	36	2259	14
1224	1	1427	23	1640	36	1904	1	1996	36	2163	17	2260	37
1225	1	1431	19	1641	36	1905	1	1997	36	2164	17	2261	17
1226	1	1432	19	1642	36	1906	1	1998	36	2165	37	2262	17
1227	1	1433	19	1643	36	1907	36	1999	36	2166	17	2263	36

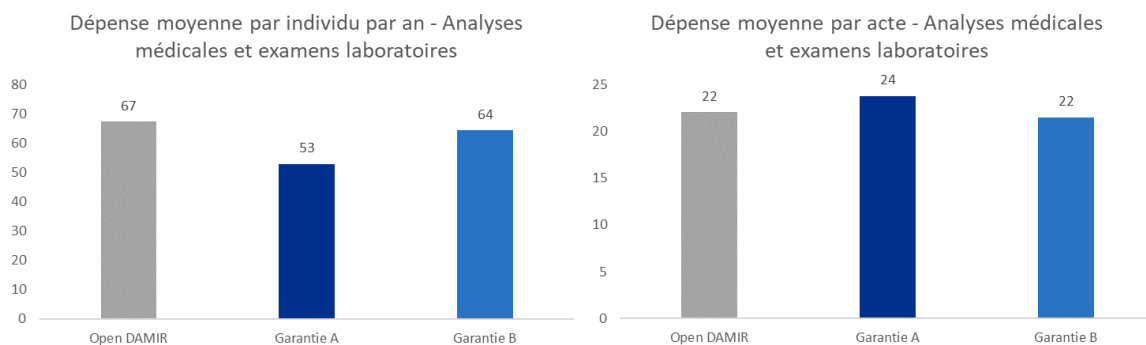
PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID
2264	36	2426	1	3327	33	3515	10	4113	32	4343	36	5118	25
2265	36	2428	6	3328	33	3516	10	4114	32	4351	36	5119	25
2268	36	2501	36	3329	10	3517	10	4131	32	4352	36	5120	25
2271	37	2502	36	3330	10	3518	8	4132	32	4353	36	5201	18
2273	37	2503	36	3331	11	3521	8	4141	32	4359	36	5202	23
2282	3	2504	36	3332	11	3522	8	4142	32	4360	36	5203	18
2283	37	3101	36	3334	11	3523	24	4143	32	4361	36	5204	22
2284	37	3110	36	3335	11	3524	25	4144	32	4363	36	5205	18
2285	12	3111	4	3336	10	3525	25	4145	32	4364	36	5206	22
2321	37	3112	4	3337	37	3526	25	4151	32	4365	36	5401	29
2331	37	3113	4	3338	37	3527	25	4152	32	4366	36	6011	36
2332	37	3115	4	3339	37	3528	25	4153	32	4368	36	6012	36
2333	36	3116	4	3340	37	3529	25	4154	32	4369	36	6013	36
2334	37	3117	4	3341	10	3530	25	4206	7	4370	36	6014	36
2335	37	3118	11	3342	10	3531	28	4207	7	4371	36	6110	36
2336	1	3119	11	3343	10	3532	24	4208	36	4372	36	6111	36
2337	37	3121	4	3351	10	3533	25	4209	36	4375	36	6112	36
2338	4	3122	4	3352	10	3534	28	4210	7	4376	36	6113	36
2339	37	3125	4	3353	10	3535	27	4211	36	4377	36	6114	36
2341	37	3126	36	3354	10	3536	25	4212	7	4378	10	6115	36
2342	1	3127	37	3355	10	3537	25	4213	1	4379	36	6116	36
2343	1	3128	7	3356	10	3538	25	4214	7	4380	36	6117	36
2344	1	3131	4	3357	10	3539	25	4215	36	4381	36	6118	36
2345	1	3132	4	3361	36	3540	29	4216	7	4382	10	6119	36
2346	1	3133	4	3362	36	3541	29	4217	7	4391	36	6120	36
2347	1	3134	4	3363	36	3542	8	4218	7	4392	36	6121	36
2351	6	3135	1	3364	10	3543	8	4219	7	4393	36	6122	36
2352	6	3136	4	3365	10	3544	8	4220	7	4394	36	6123	36
2353	36	3137	13	3366	10	3545	8	4221	7	4395	36	6124	36
2354	6	3138	1	3374	10	3546	8	4222	36	4396	3	6126	35
2355	6	3139	36	3375	10	3547	30	4223	36	4397	3	6127	35
2371	36	3211	5	3379	11	3548	8	4224	36	4411	36	6131	32
2372	36	3212	5	3380	10	3549	8	4225	7	4412	36	6132	32
2380	1	3213	5	3381	10	3550	29	4227	36	4413	36	6133	32
2381	1	3216	5	3382	10	3551	8	4228	36	4414	36	6134	36
2382	1	3221	5	3383	10	3552	8	4229	7	4415	36	6135	36
2383	1	3222	5	3384	10	3553	25	4311	36	4416	4	6210	36
2384	1	3223	5	3385	10	3554	25	4312	36	4417	36	6212	36
2385	1	3225	5	3386	10	3555	25	4313	36	4419	36	6213	36
2386	1	3300	36	3387	10	3556	25	4315	37	4501	36	6214	36
2387	1	3301	36	3388	10	3557	25	4316	36	4511	36	6215	36
2388	1	3304	36	3389	10	3558	29	4317	36	4611	36	6221	36
2389	37	3306	10	3390	10	3561	9	4318	36	4612	36	6222	36
2391	6	3307	10	3391	36	3571	8	4319	36	5101	25	6231	36
2392	6	3311	10	3392	36	3572	36	4320	36	5102	25	6232	36
2411	37	3312	10	3393	36	3573	36	4321	36	5103	25	6233	36
2412	37	3313	10	3394	36	3574	36	4322	36	5104	25	6234	36
2413	37	3314	10	3395	36	3575	36	4323	36	5105	25	6235	36
2414	1	3315	10	3396	36	3581	25	4324	36	5106	25	6236	36
2415	37	3316	10	3397	36	3582	25	4325	36	5107	24	6237	36
2416	37	3317	10	3398	36	3583	24	4326	36	5108	24	6238	36
2417	5	3318	10	3399	11	3591	36	4327	36	5109	25	6239	36
2418	6	3319	10	3411	10	3592	36	4328	36	5110	25	6241	36
2419	37	3320	10	3412	37	3593	36	4329	36	5111	25	6242	36
2420	37	3321	10	3413	36	3594	36	4330	36	5112	25	6251	36
2421	37	3322	10	3414	36	3610	36	4331	36	5113	25	6252	36
2422	4	3323	10	3511	8	3611	36	4332	36	5114	25	6253	36
2423	37	3324	10	3512	8	3612	36	4339	36	5115	25	6257	36
2424	37	3325	10	3513	8	4111	32	4341	36	5116	25	6258	36
2425	37	3326	10	3514	8	4112	32	4342	36	5117	25	6261	36

PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID	PRS_NAT	ID
6262	36	9318	34	9766	36	9880	36
6311	36	9411	23	9767	36	9890	36
7111	36	9412	23	9768	36	9891	36
7112	36	9413	23	9769	36	9894	36
7113	36	9421	36	9771	36	9895	36
7119	36	9422	36	9772	36	9896	36
8111	36	9423	36	9773	36	9901	7
8112	36	9424	36	9774	36	9902	36
8113	36	9429	36	9775	36	9911	36
8114	36	9430	11	9776	36	9912	36
8115	36	9431	36	9777	36	9999	36
8116	36	9432	36	9778	36		
8117	36	9433	28	9802	36		
8118	36	9434	36	9803	36		
8119	36	9511	36	9806	36		
8120	36	9512	34	9808	36		
8121	36	9521	34	9809	36		
8222	36	9566	10	9810	36		
8223	36	9567	34	9812	36		
8224	36	9568	34	9813	36		
8225	36	9569	36	9814	36		
8226	36	9570	36	9815	36		
8227	36	9601	37	9816	36		
9111	36	9701	36	9817	36		
9112	36	9703	36	9820	36		
9113	36	9704	36	9822	36		
9114	36	9705	36	9823	36		
9115	36	9706	36	9825	36		
9116	36	9707	36	9830	36		
9118	36	9708	36	9831	36		
9119	36	9709	36	9832	36		
9121	36	9710	36	9833	36		
9122	36	9711	36	9841	36		
9123	36	9712	36	9842	36		
9129	36	9713	36	9843	36		
9131	36	9714	36	9844	36		
9132	32	9715	36	9845	36		
9133	32	9716	36	9847	36		
9134	36	9717	36	9849	36		
9135	36	9719	36	9851	36		
9141	1	9721	36	9852	36		
9143	37	9722	36	9853	36		
9144	36	9723	17	9854	36		
9151	36	9724	36	9855	36		
9152	36	9725	17	9856	36		
9161	36	9726	36	9857	36		
9162	36	9727	36	9858	36		
9163	36	9731	36	9859	36		
9164	36	9732	36	9860	36		
9167	36	9741	36	9861	36		
9170	36	9742	36	9862	36		
9191	36	9743	36	9863	36		
9201	36	9744	36	9864	36		
9202	36	9751	36	9865	36		
9203	36	9752	36	9866	36		
9211	36	9761	36	9867	36		
9221	36	9762	36	9871	36		
9311	34	9763	36	9872	36		
9312	34	9764	36	9873	36		
9313	36	9765	36	9875	36		

Nous ne sommes pas à l'abri d'une mauvaise classification ou d'une mauvaise extraction de nos résultats de classification. N'hésitez pas à nous contacter si vous voyez des incohérences.

Annexe 11 : Les autres dépenses

11.1. Analyses médicales et examens laboratoires

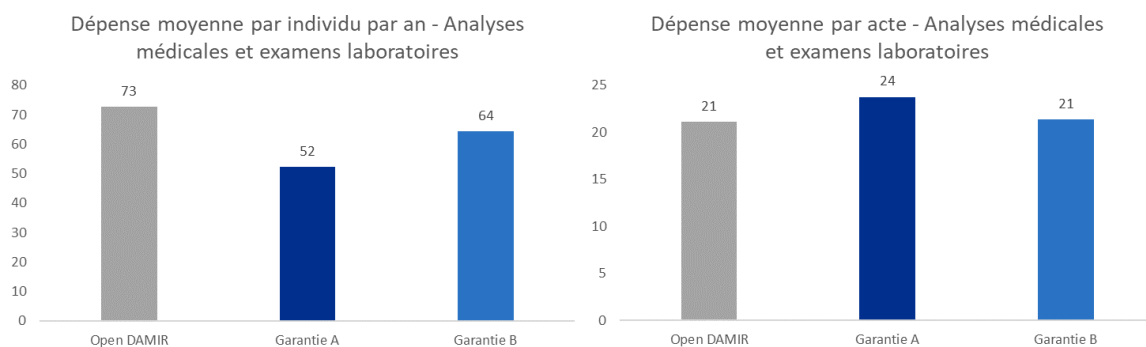


Graphique 28 : Dépense moyenne pour les analyses médicales et examens laboratoires

En moyenne, sur une année, un Français dépense 67 € pour des analyses médicales et des examens laboratoires et le prix moyen d'un de ces actes s'élève à 22 €. Les types d'analyse de prise de sang sont nombreux et peuvent parfois coûter une dizaine ou vingtaine d'euros (voire quelques euros).

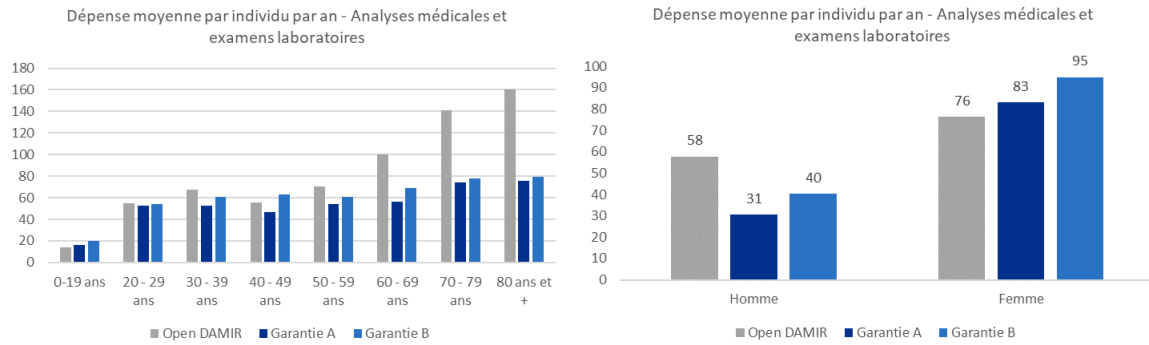
En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 53 € par an pour un prix unitaire de 24 € l'acte. Il dépense donc moins annuellement en moyenne mais légèrement plus par acte.

Les adhérents à la garantie B présentent un coût moyen de l'acte et une dépense moyenne annuelle très proche de ceux observés sur la population française dans son entièreté. La différence avec la garantie A pourrait s'expliquer de la même manière que pour les actes techniques médicaux. En revanche, dans ce cas-ci, la garantie B est représentative des habitudes de consommation des Français pour cette sous-famille.



Graphique 29 : Dépense moyenne pour les analyses médicales et examens laboratoires uniquement sur la région D

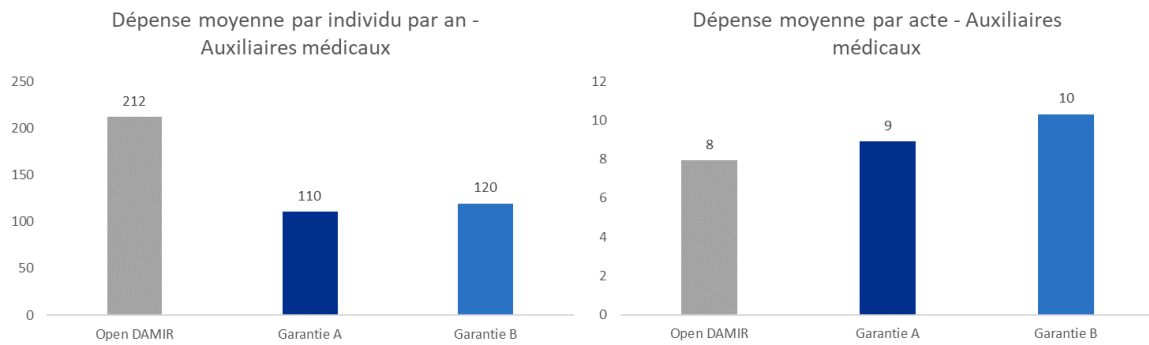
Le graphique 29 indique qu'il y a éventuellement un « effet région » à noter.



Graphique 30 : Dépense moyenne pour les analyses médicales et examens laboratoires selon les âges et le sexe

Le graphique 30 indique que la dépense annuelle moyenne pour ces actes augmente avec l'âge de manière globale et que les femmes consomment plus annuellement.

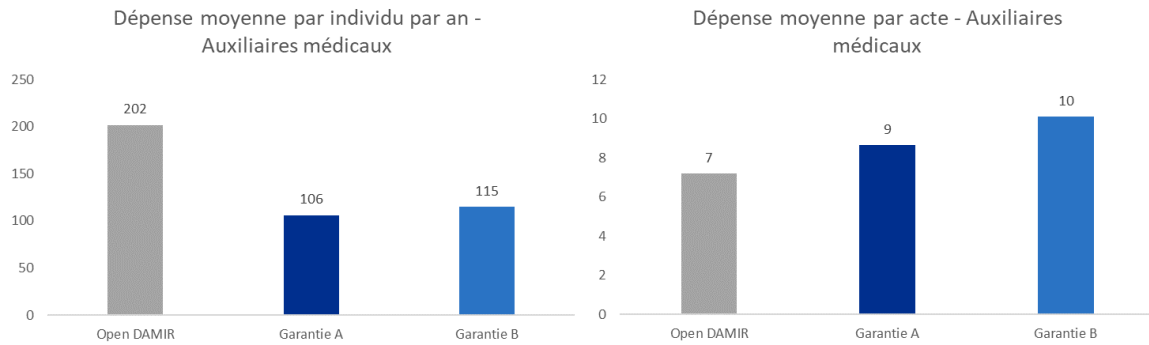
11.2. Auxiliaires médicaux



Graphique 31 : Dépense moyenne pour les auxiliaires médicaux

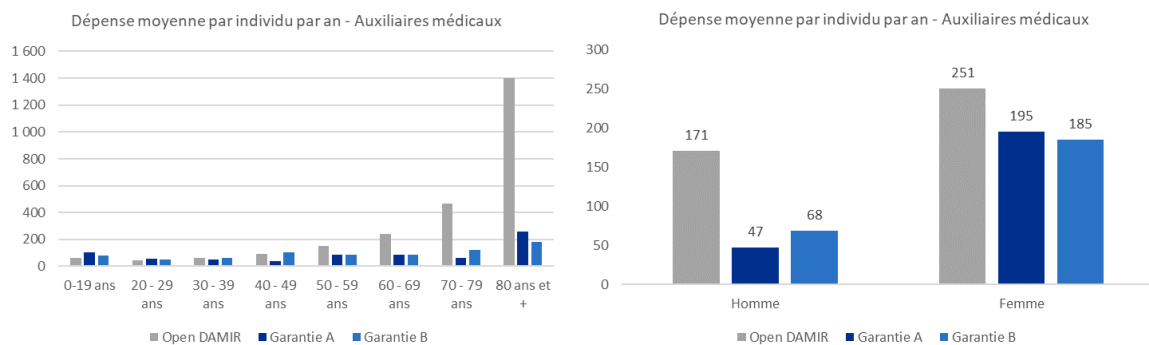
En moyenne, sur une année, un Français dépense 212 € pour des auxiliaires médicaux et le prix moyen d'un de ces actes s'élève à 8 € d'après le graphique 31. À titre indicatif, le prix d'une séance chez un kinésithérapeute est de l'ordre de 16 € et chez l'orthophoniste, le prix est de l'ordre de 25 € à plus de 70 €. Le prix d'une piqûre par une infirmière coûte quelques euros (environ 3 €).

En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 110 € par an (soit près de la moitié comparé à l'Open DAMIR) pour un prix unitaire de 9 € l'acte (montant proche de l'Open DAMIR). Ces différents montants pour la garantie B sont similaires. De ce fait, les adhérents de VirtuaMut', comparés à la population française entière, tendent à consommer moins (moins fréquemment ou sur des actes moins coûteux).



Graphique 32 : Dépense moyenne pour les auxiliaires médicaux uniquement sur la région D

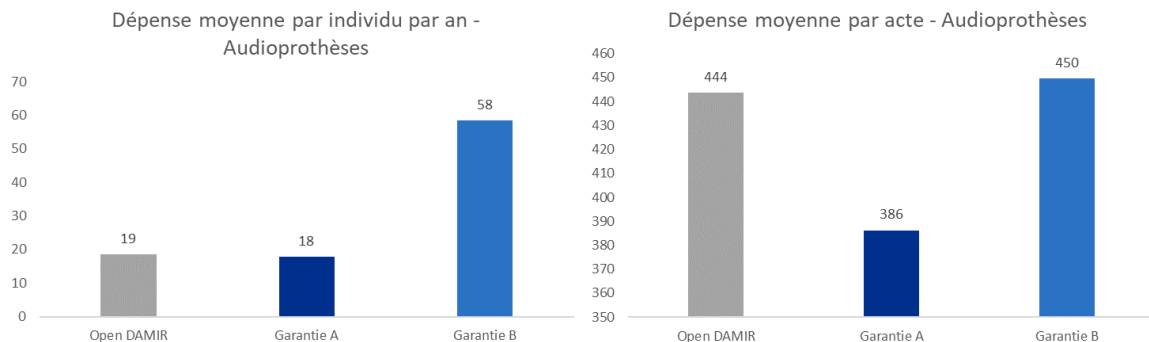
Le graphique 32 indique qu’il y a éventuellement un « effet région » à noter.



Graphique 33 : Dépense moyenne pour les auxiliaires médicaux selon les âges et le sexe

Le graphique 33 indique que la dépense annuelle moyenne pour ces actes augmente avec l’âge de manière globale et que les femmes consomment plus annuellement (cela est d’autant plus visible chez VirtuaMut’). Il y a un pic de dépenses de la population française âgée de plus de 70 ans, ce qui paraît cohérent comme plus une personne est âgée, plus elle aura recours aux auxiliaires médicaux.

11.3. Audioprothèses

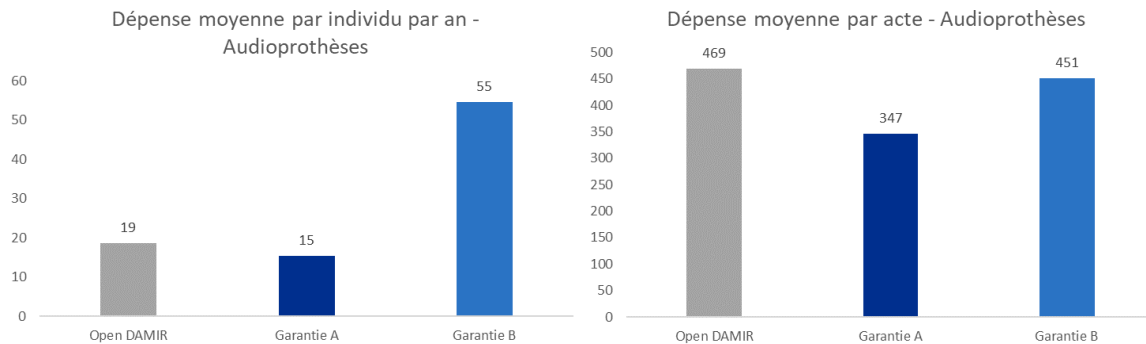


Graphique 34 : Dépense moyenne pour les audioprothèses

En moyenne, sur une année, un Français dépense 19 € pour des audioprothèses. Cette faible dépense s’explique par le fait que « seulement » 16 % de la population française présente un problème d’audition en 2014 (selon une étude de la DREES) et même parmi eux, seulement une portion (moins de 34 %) s’équipe pour y remédier. Le prix moyen d’un de ces actes s’élève à 444 € d’après le graphique 34. Or, le prix moyen d’une audioprothèse est de 1 500 € en 2018 selon le syndicat des audioprothésistes

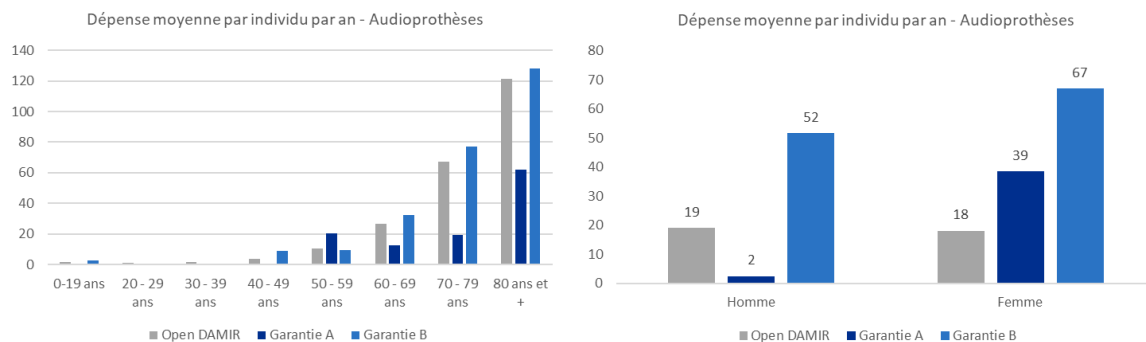
UNSAF. Nous en concluons donc qu'il y a, dans cette sous-famille d'actes, aussi les accessoires auditifs et éventuellement les piles.

En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 18 € par an (très proche de l'Open DAMIR) pour un prix unitaire de 386 € l'acte. Pour la garantie B, la dépense annuelle par individu s'élève à 58 € (beaucoup plus que pour l'Open DAMIR) pour un prix unitaire à 450 € (proche de l'Open DAMIR). Cette différence s'explique par un portefeuille plus âgé pour la garantie B.



Graphique 35 : Dépense moyenne pour les audioprothèses uniquement sur la région D

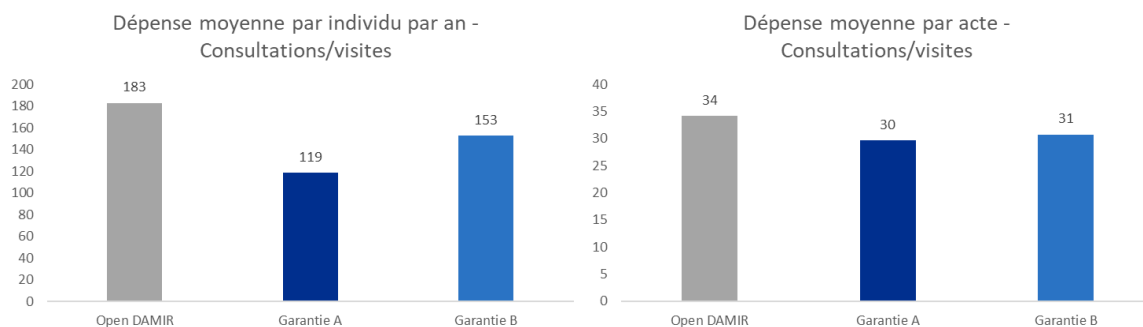
Le graphique 35 indique qu'il n'y a pas, *a priori*, d'« effet région » significatif (sauf peut-être pour le prix unitaire d'une audioprothèse mais nous n'avons pas plus d'informations sur ce qui expliquerait cette différence de 25 €).



Graphique 36 : Dépense moyenne pour les audioprothèses selon les âges et le sexe

Le graphique 36 indique que la dépense annuelle moyenne pour ces actes augmente avec l'âge, ce qui est cohérent avec ce qui est observable dans la population française (les personnes âgées ont tendance à avoir plus de problèmes de surdité). La forte dépense pour les âges élevés et le fait que les individus de la garantie B soient relativement âgés comparés à la population française dans son entièreté pourraient expliquer l'importance de la dépense moyenne par individu par an pour la garantie B, telle que visible en graphique 34. Par ailleurs, il semblerait que les femmes consomment plus annuellement que les hommes chez VirtuaMut' (alors que cela est similaire dans l'Open DAMIR). Cela est sûrement dû à l'effet du volume faible de portefeuille.

11.4. Consultations/visites

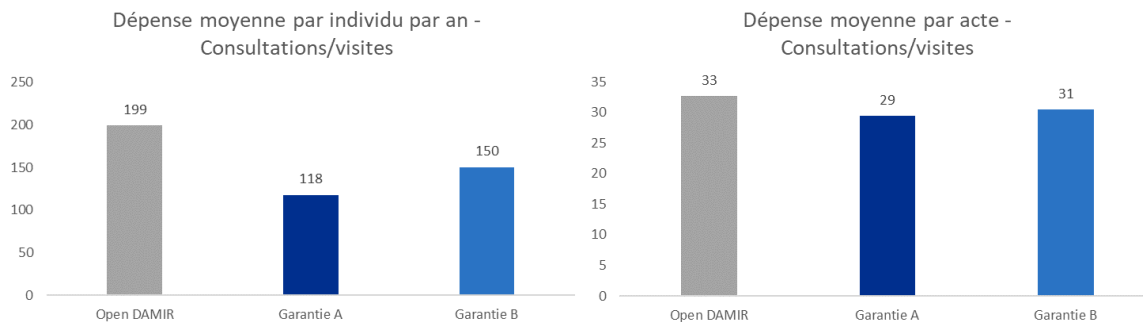


Graphique 37 : Dépense moyenne pour les consultations et visites

En moyenne, sur une année, un Français dépense 183 € pour des consultations ou visites (et autres assimilés) et le prix moyen d'un de ces actes s'élève à 34 €. À titre indicatif, le prix d'une consultation chez le généraliste (secteur 1) est de 25 € depuis le 1^{er} mai 2017. Le prix d'une consultation chez le spécialiste (secteur 1) varie de 28 € à 49 € pour la majorité des cas. Ces montants paraissent donc cohérents d'autant qu'une consultation peut en enclencher une autre.

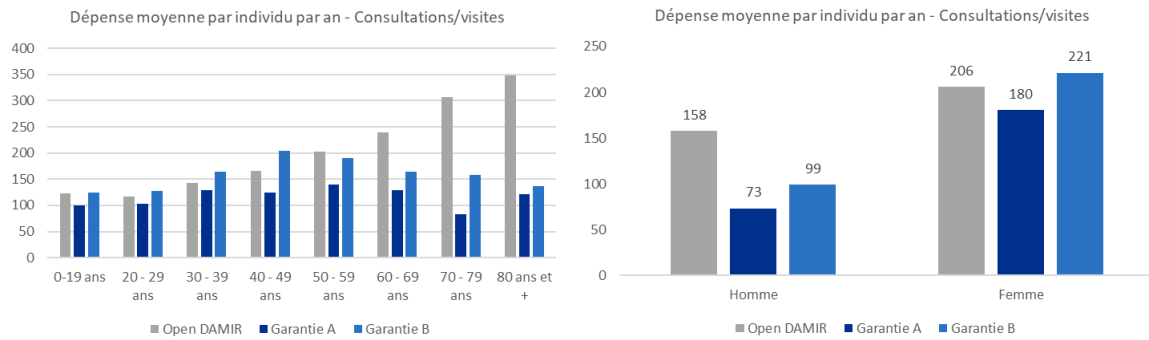
En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 119 € par an pour un prix unitaire de 30 € l'acte. Cela est inférieur à ce qui est visible sur l'Open DAMIR et pourrait s'expliquer par l'argument du « portefeuille restreint ».

Par rapport à la garantie A, les adhérents à la garantie B présentent un coût moyen de l'acte très proche mais une dépense annuelle plus élevée (et plus proche de l'Open DAMIR). L'argument du « portefeuille restreint » tend à se confirmer comme les écarts tendent à diminuer avec l'augmentation de l'effectif du portefeuille de VirtuaMut'.



Graphique 38 : Dépense moyenne pour les consultations et visites uniquement sur la région D

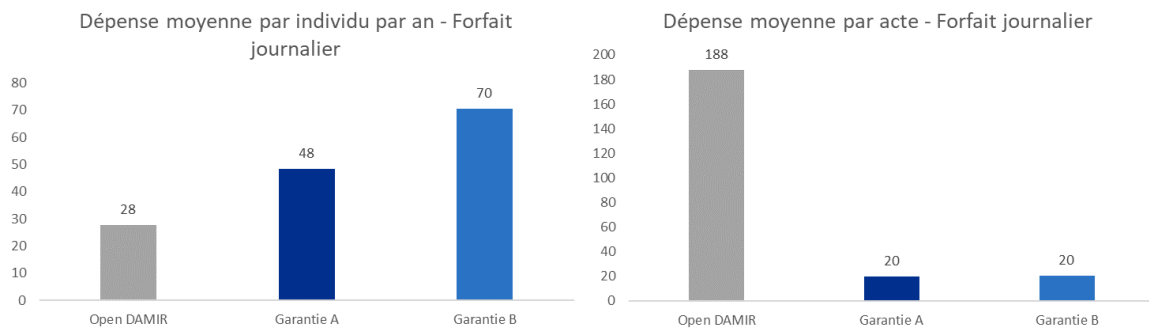
Il y a ici peut-être un « effet région » sur la dépense annuelle moyenne par individu (écart de 16 €) qui peut éventuellement s'expliquer par une fréquence légèrement supérieure de visites ou consultations chez le médecin des individus habitant en région D (tarifs plus bas, habitudes, modes de vie, ...) comme le coût unitaire par acte est légèrement plus bas.



Graphique 39 : Dépense moyenne pour les consultations et visites selon les âges et le sexe

Le graphique 39 indique que la dépense annuelle moyenne pour ces actes augmente avec l'âge dans la base de l'Open DAMIR alors que dans le cas de la mutuelle, cette dépense prend la forme d'une cloche avec des dépenses en consultations et visites plus élevées chez les individus d'âges moyens (40 à 59 ans). Les habitudes de consommation pour cette sous-famille d'actes des adhérents de la mutuelle ne semblent pas représentatives de celles de la population française dans sa globalité. Par ailleurs, il semblerait que les femmes consomment plus annuellement que les hommes et cela est d'autant plus marqué dans le périmètre de la mutuelle.

11.5. Forfaits journaliers

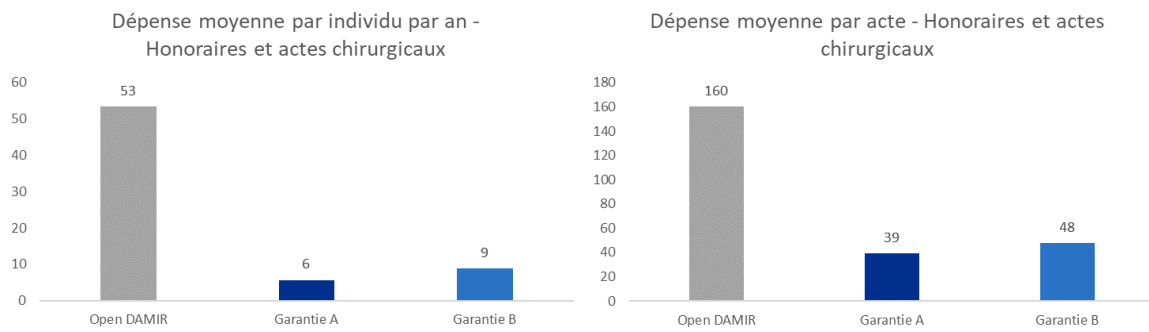


Graphique 40 : Dépense moyenne pour les forfaits journaliers

En moyenne, sur une année, un Français dépense 28 € pour des forfaits journaliers hospitaliers et le prix moyen d'un de ces actes s'élève à 188 € d'après le graphique 40. Ce décalage s'explique par le fait que seulement une partie de la population française consomme : selon la publication de l'agence technique de l'information sur l'hospitalisation, il y aurait eu en 2018, 12,8 millions de patients hospitalisés, soit environ 19 % de la population française.

En revanche, le montant de dépense de 188 € par acte nous semble étrange comme un forfait journalier coûte 20 € (par jour). La quantité d'actes prise en compte est pourtant bien celle de la variable QTE_ACTE (qui décompte le nombre de jours) et non celle de dénombrement (et dans tous les cas, les deux sont proches) et la classification des codes actes de cette sous-famille est assez peu ambiguë au vu de la clarté des libellés. Une erreur de traitement serait possible mais elle se serait répercutée dans le cas des autres sous-familles d'actes. Il était plutôt attendu de retrouver les 20 € qui sont bien présents chez les garanties A et B (dont le niveau de prestation est de 100 % des frais réels). À l'heure actuelle, nous ne trouvons pas d'explication à ce montant, nous décidons donc d'écarter ce segment des segments à tarifier par GLM.

11.6. Honoraires et actes chirurgicaux

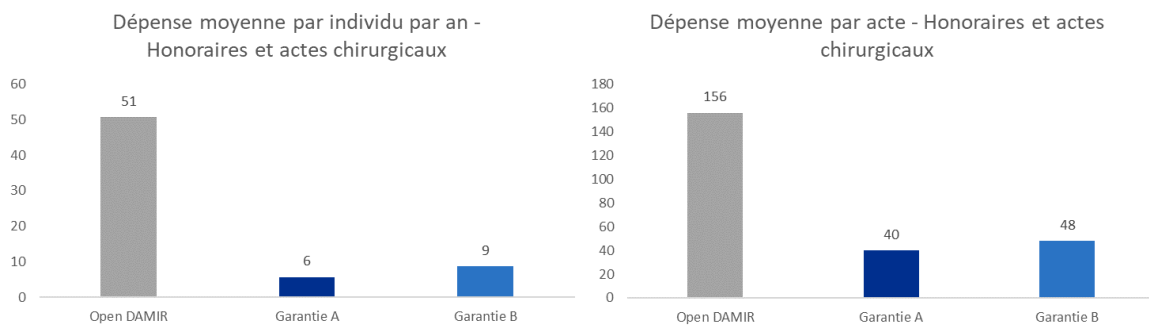


Graphique 41 : Dépense moyenne pour des honoraires et actes chirurgicaux

En moyenne, sur une année, un Français dépense 53 € pour des honoraires et actes chirurgicaux et le prix moyen d'un de ces actes s'élève à 160 € d'après le graphique 41. À titre indicatif, le prix d'une consultation chez le chirurgien varie de 23 € à 60 € selon la durée et la complexité de la consultation. Le prix moyen d'un acte de chirurgie pour une opération de cataracte est de 355 € avec dépassement maîtrisé d'honoraires.

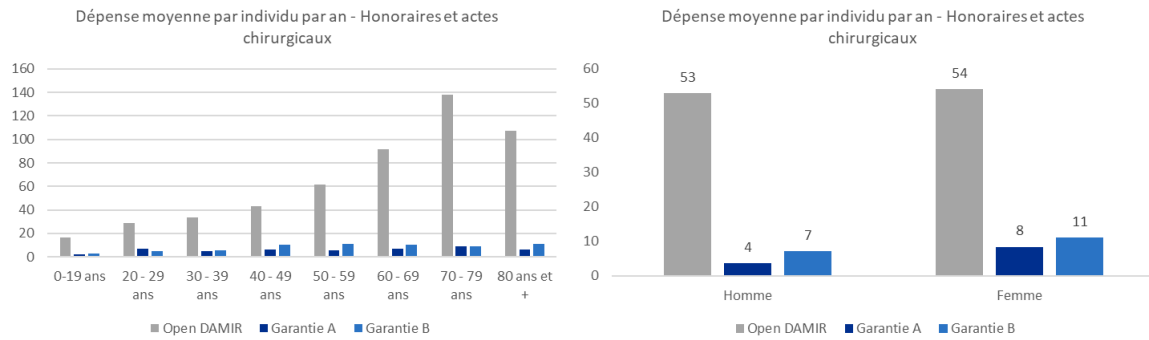
En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 6 € par an pour un prix unitaire de 39 € l'acte. Cela est inférieur à ce qui est visible sur l'Open DAMIR et pourrait s'expliquer par l'argument du « portefeuille restreint » (il y a peut-être eu seulement que des actes spécifiques et non coûteux).

Par rapport à la garantie A, les adhérents à la garantie B présentent un coût moyen de l'acte et une dépense annuelle un peu plus élevés. L'hypothèse du « portefeuille restreint » pourrait être un élément de réponse mais comme la garantie A ne propose pas de remboursement de prestations pour les actes chirurgicaux, nous aurions tendance à penser qu'il y aurait dû y avoir plus d'écarts entre les montants issus de la garantie A et de la B.



Graphique 42 : Dépense moyenne pour des honoraires et actes chirurgicaux uniquement sur la région D

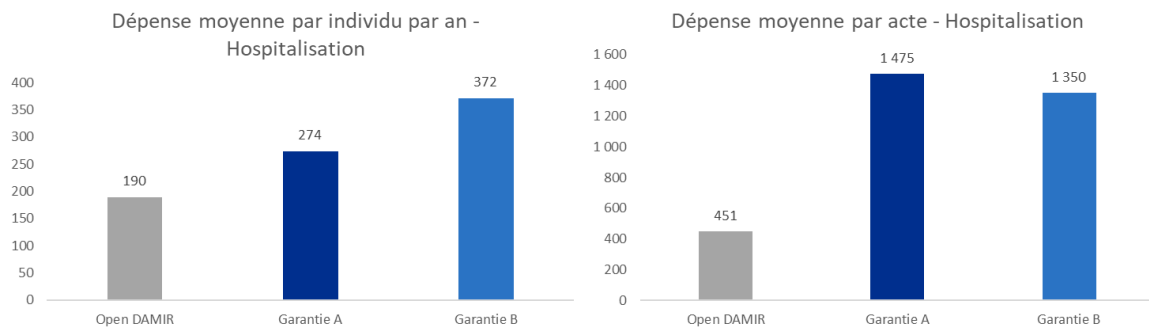
Le graphique 42 indique qu'il n'y a *a priori* pas d'« effet région » significatif.



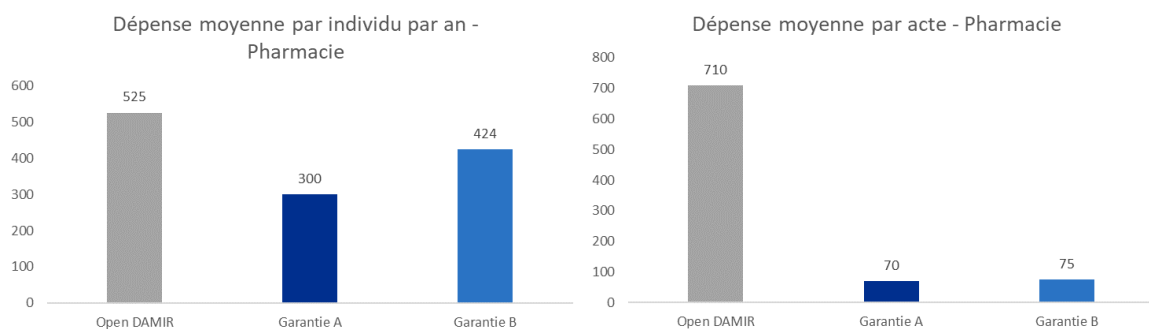
Graphique 43 : Dépense moyenne pour des honoraires et actes chirurgicaux selon les âges et le sexe

Le graphique 43 indique que la dépense annuelle moyenne pour ces actes augmente avec l'âge de manière générale pour l'Open DAMIR (pic à 70 ans). Pour VirtuaMut', la faible valeur de dépenses tend à être volatile. Par ailleurs, les femmes consomment légèrement plus annuellement que les hommes.

11.7. Hospitalisation et pharmacie



Graphique 44 : Dépense moyenne pour les actes hospitaliers



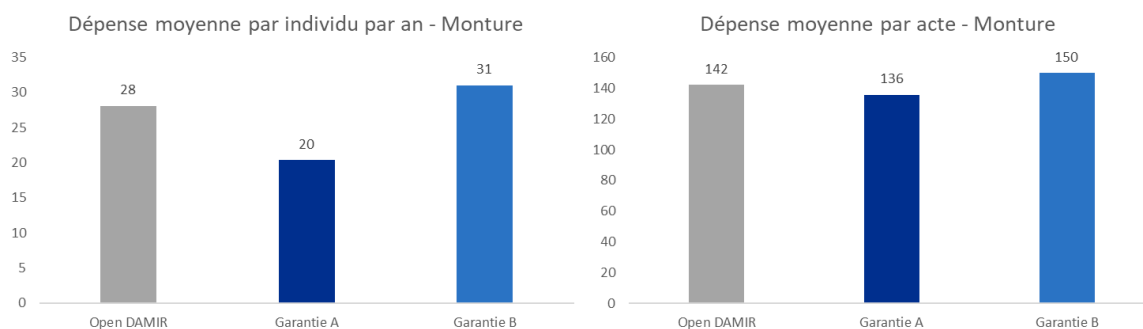
Graphique 45 : Dépense moyenne pour de la pharmacie

Au regard des montants de dépenses faibles en hospitalisation et élevés en pharmacie (notamment les dépenses moyennes par acte) pour l'Open DAMIR, nous en concluons qu'il y a eu une différence de classification en ce qui concerne les grands appareillages : ils sont classés en pharmacie selon notre classification harmonisée mais la mutuelle les a classés en hospitalisation sans indiquer de libellés différents. En effet, nous n'avons pas pu différencier les lignes de grands appareillages des lignes d'hospitalisation dans les données de VirtuaMut' et les avons confondus l'un avec l'autre sans nous en rendre compte ; nous avons même cru qu'il n'y avait pas d'acte de grands appareillages – cette étude était nécessaire pour le réaliser. Comme il n'y a pas moyen d'effectuer une différenciation dans les données de VirtuaMut', les grands appareillages de l'Open DAMIR seront transférés pour la suite des travaux dans le segment « Hospitalisation » le temps de l'élaboration d'un tarif. Comme la grille de

garanties de VirtuaMut' n'indique pas de prestations claires pour les grands appareillages (il est mentionné de remboursement étudié au cas par cas), faute de solution, il sera pris les prestations de l'audioprothèse (comme les grands appareillages comprennent parfois les appareils auditifs).

Par ailleurs, dans la pharmacie, il ne faudra pas oublier que les petits appareillages y sont inclus (ce qui explique une dépense moyenne par acte de 70 €).

11.8. Montures



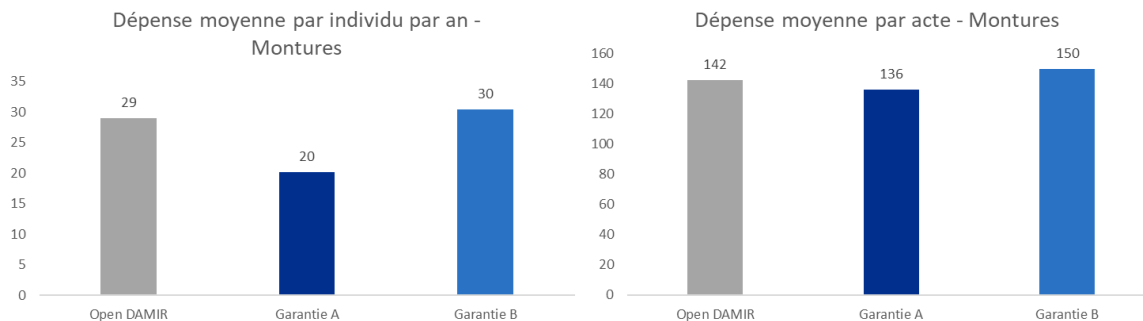
Graphique 46 : Dépense moyenne pour des montures

En moyenne, sur une année, un Français dépense 28 € pour les montures. L'argument est le même que pour les audioprothèses : seule une partie de la population est concernée. Le prix moyen d'un de ces actes s'élève à 142 € d'après le graphique 46. Cela reste cohérent avec les données de marché.

En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 20 € par an pour un prix unitaire de 136 € l'acte. La différence avec l'Open DAMIR pourrait venir d'un léger effet « portefeuille restreint ».

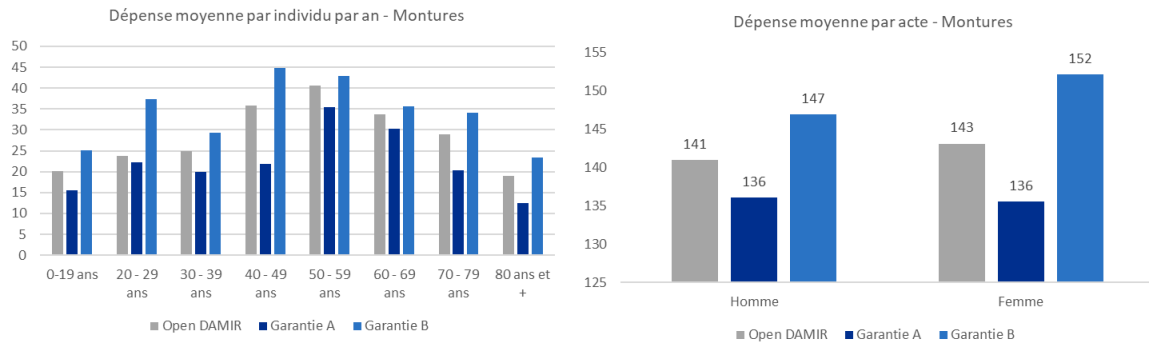
Les adhérents à la garantie B présentent un coût moyen de l'acte à 150 € et une dépense moyenne annuelle de 31 €.

En somme, pour les deux garanties, les dépenses restent suffisamment proches de celles observées sur la population française dans son entièreté. Le portefeuille de VirtuaMut' est donc représentatif des habitudes de consommation des Français pour cette sous-famille.



Graphique 47 : Dépense moyenne pour des montures uniquement sur la région D

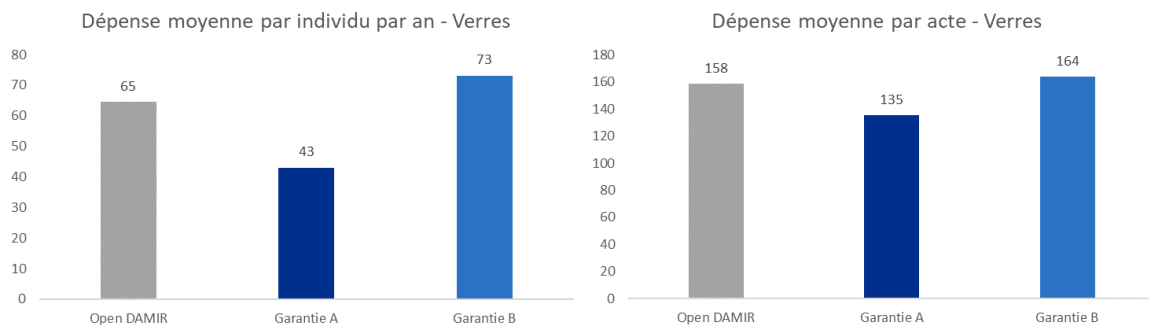
Le graphique 47 indique qu'il n'y a pas, *a priori*, d'effet « région » significatif.



Graphique 48 : Dépense moyenne pour les montures selon les âges et le sexe

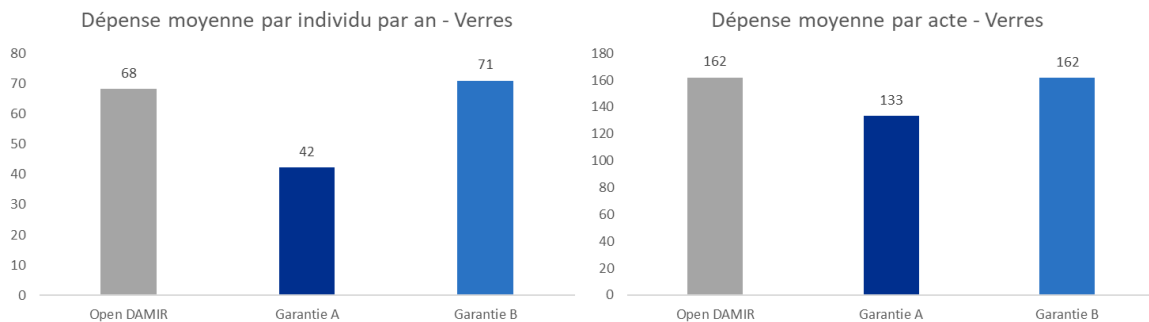
La dépense moyenne par individu par an en montures présente une forme de cloche avec un pic vers les âges moyens (entre 40 et 59 ans) où certains individus commencent à s'équiper. Les hommes et les femmes consomment plus ou moins autant.

11.9. Verres



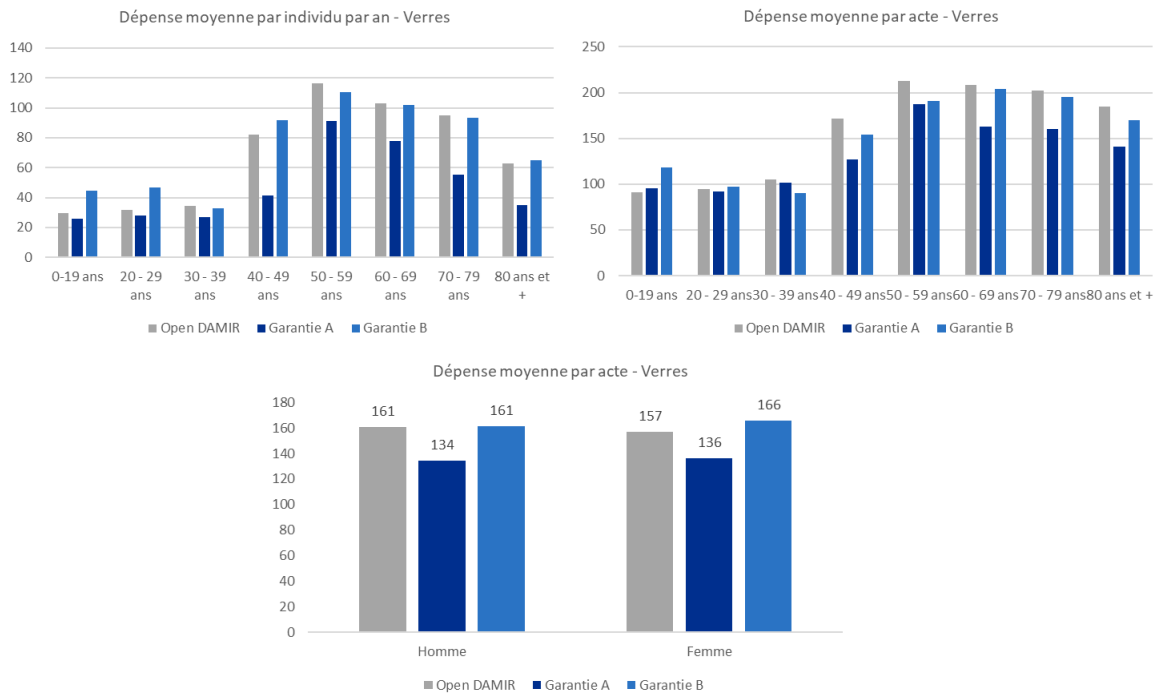
Graphique 49 : Dépense moyenne pour des verres

L'analyse ici sera similaire au cas de la monture (les uns vont souvent avec l'autre), en sachant que des verres simples coûtent en moyenne 143 € les deux verres et 433 € pour des verres progressifs selon les chiffres issus d'une étude menée par le cabinet Alcimed pour la Direction de la Sécurité Sociale (*Analyse économique du secteur des appareillages optiques et auditifs*, mars 2011). Par ailleurs, selon une étude de la DREES datant de 2014, « sept adultes sur dix portent des lunettes », ce qui explique un montant de 65 € de dépense annuelle par individu.



Graphique 50 : Dépense moyenne pour des verres uniquement sur la région D

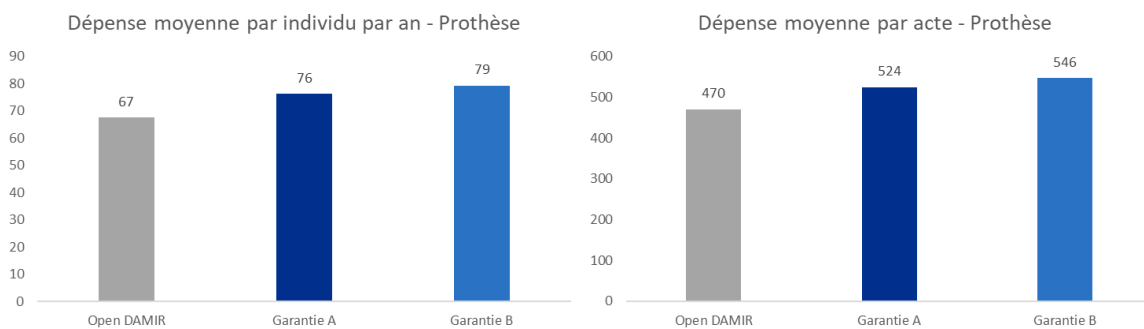
Le graphique 50 indique qu'il n'y a pas, *a priori*, d'effet « région » significatif.



Graphique 51 : Dépense moyenne pour les consultations et visites selon les âges et le sexe

Que ce soit en termes de dépense moyenne par individu ou par acte au fil des âges, une sorte de « palier » peut être observé entre les deux classes d’âge 30-39 ans et 40-49 ans. Cela pourrait s’expliquer par le changement de la nature des verres (passage aux progressifs notamment), par le changement des caractéristiques de verres (vision aggravée donc verres plus épais et plus coûteux à affiner ou prise d’options de plus en plus nombreuses) ou encore, par l’aggravation de l’état des yeux au fil des âges (maladie de cécité). Il n’y a pas non plus de différence notable entre les hommes et les femmes pour cette sous-famille d’actes.

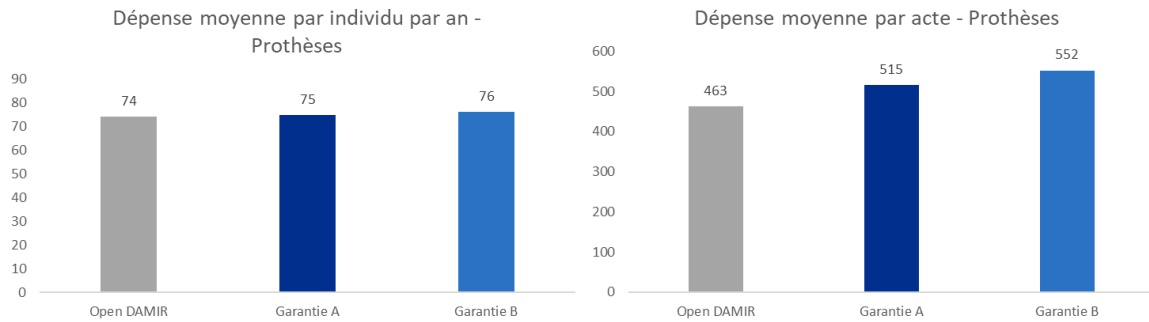
11.10. Prothèses dentaires



Graphique 52 : Dépense moyenne pour des prothèses dentaires

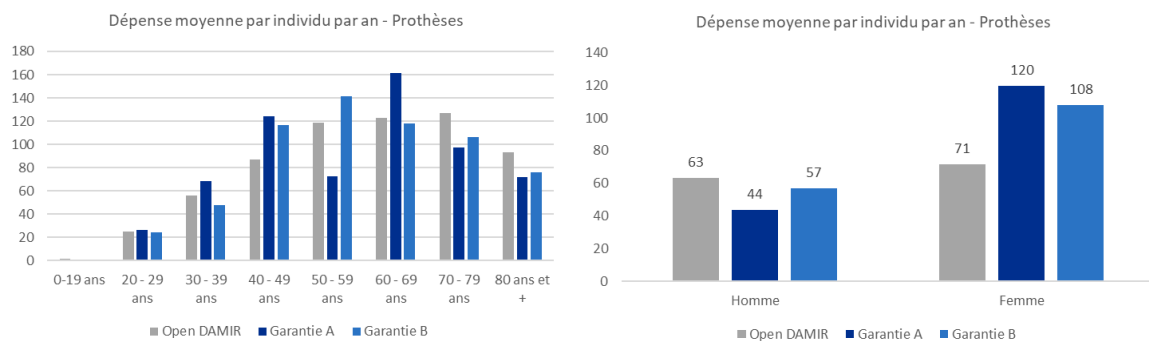
L’analyse ici sera similaire à celle sur les verres en sachant que les prothèses dentaires sont très diversifiées et les coûts dépendent à la fois des dents concernées (leur position), du type de prothèses, des matériaux utilisés. *A priori*, selon le graphique 52, la dépense annuelle moyenne d’un Français en prothèse dentaire serait de 67 € pour un prix unitaire d’acte à 470 €.

Les dépenses plus élevées pour les portefeuilles de VirtuaMut’ pourraient s’expliquer par l’argument du « portefeuille restreint ».



Graphique 53 : Dépense moyenne pour des prothèses dentaires uniquement sur la région D

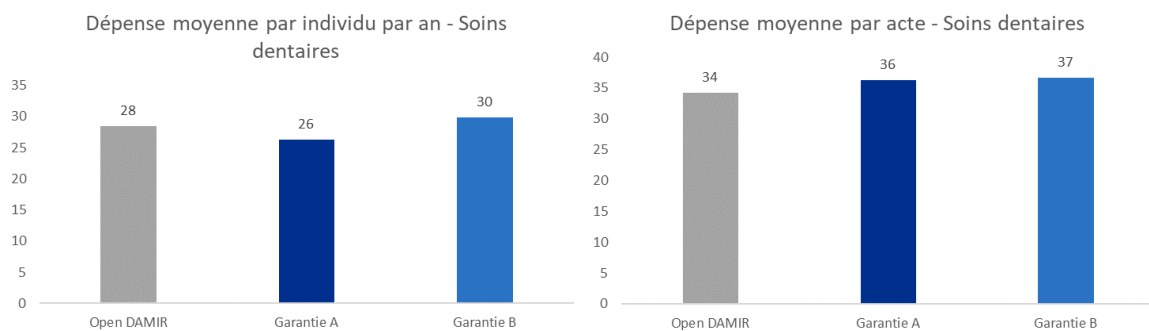
Le graphique 53 indique qu'un « effet région » pourrait éventuellement être à noter.



Graphique 54 : Dépense moyenne pour des prothèses dentaires selon les âges et le sexe

D'une manière similaire aux verres, la dépense moyenne par individu par an prend la forme d'une cloche. Cela traduit aussi le fait que pour ces actes (prothèses dentaires, verres, audioprothèses), les dépenses sont fortement corrélées aux âges et à l'évolution de la santé due à la vieillesse. Il est observé ici un pic pour les individus de 60 à 69 ans et cela est cohérent avec les besoins d'un être humain au fil du temps en matière de dentition (perte de dents au fur et à mesure du temps). Ainsi, pour les individus âgés de 0 à 19 ans, les prothèses dentaires ne sont que peu usitées. Par ailleurs, il est observé que les femmes ont tendance à coûter plus que les hommes et cela est d'autant plus marqué chez VirtuaMut' que dans l'Open DAMIR.

11.11. Soins dentaires

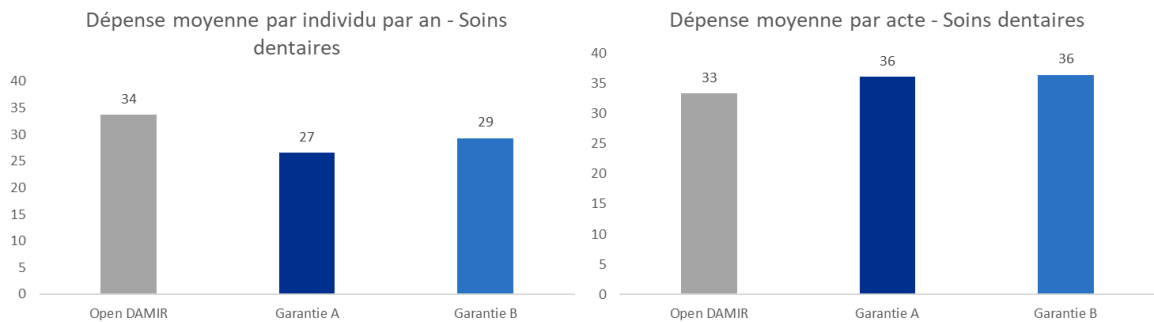


Graphique 55 : Dépense moyenne pour des soins dentaires

En moyenne, sur une année, un Français dépense 28 € pour les soins dentaires (détartrage, traitement de carie, ...) et le prix moyen d'un de ces actes s'élève à 34 € d'après le graphique 55. À titre indicatif, le

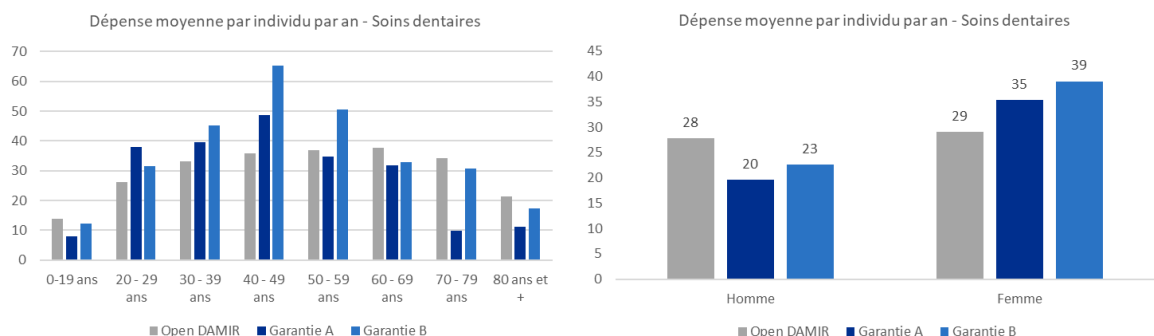
prix d'un détartrage est de l'ordre de 29 € et le traitement d'une carie peut aller de 17 à plus de 60 €. Une dévitalisation, notamment d'une molaire par exemple peut coûter proche de 82 €.

En ce qui concerne les adhérents de VirtuaMut', un adhérent à la garantie A dépense en moyenne 26 € par an pour un prix unitaire de 36 € l'acte. Les adhérents à la garantie B présentent un coût moyen de l'acte à 30 € et une dépense moyenne annuelle de 37 €. En somme, pour les deux garanties, les dépenses restent suffisamment proches de celles observées sur la population française dans son entièreté. Le portefeuille de VirtuaMut' est donc représentatif des habitudes de consommation des Français pour cette sous-famille.



Graphique 56 : Dépense moyenne pour des soins dentaires uniquement sur la région D

Le graphique 56 indique qu'un « effet région » pourrait éventuellement être à noter.



Graphique 57 : Dépense moyenne pour des soins dentaires selon les âges et le sexe

L'analyse est similaire au cas des verres. À noter cependant ici que pour l'Open DAMIR, la dépense des hommes et des femmes est similaire.

Retour p.66

Annexe 12 : Démonstration – Pour la prime pure

Ce qui est démontré :

$$\text{Prime pure} = \mathbb{E}[X_k] = \mathbb{E}[S] \times \mathbb{E}[N]$$

Avec

$$X_k = \sum_{i=1}^N S_i$$

Où :

- $N \in \mathbb{N}$ est la variable aléatoire représentant le nombre total de soins sur la période considérée ;
- $S_i \in \mathbb{R}^{+*}$ est la variable aléatoire représentant le montant du $i^{\text{ème}}$ soin avec $i \in \{1, \dots, N\}$;
- Les S_1, S_2, \dots, S_N sont i.i.d. à une variable S et indépendants de N ;
- Avec la convention que $X_k = 0$ si $N = 0$.

On conditionne X_k par N :

$$\mathbb{E}(X_k|N) = \mathbb{E}\left(\sum_{i=1}^N S_i \mid N\right) = \sum_{i=1}^N \mathbb{E}(S_i|N) \stackrel{(A)}{=} \sum_{i=1}^N \mathbb{E}(S) = N\mathbb{E}(S)$$

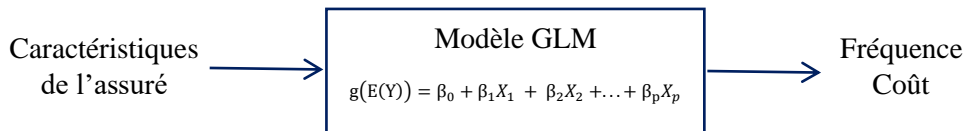
(A) Indépendance de S_i et N et les S_i sont i.i.d. à S

Or :

$$\mathbb{E}(X_k) = \mathbb{E}(\mathbb{E}(X_k|N)) = \mathbb{E}(N\mathbb{E}(S)) = \mathbb{E}(S)\mathbb{E}(N)$$

Annexe 13 : Modèles linéaires généralisés (GLM)

Les modèles linéaires généralisés permettent d'estimer la fréquence et le coût moyen par le biais de variables explicatives qui permettront le pilotage de ces deux termes. Ces variables explicatives sont ici par exemple le sexe ou la région de résidence de l'assuré. La fréquence et le coût moyen dépendent alors des caractéristiques de l'adhérent et chaque type d'adhérent aura sa prime pure propre. Les adhérents ayant les mêmes caractéristiques se verront attribuer la même prime pure.



13.1. Généralités

Présentés pour la première fois sous ce nom par John NELDER et Robert WEDDERBURN en 1972, puis exposés de façon complète par Peter MC CULLAGH et John NELDER en 1989, les modèles linéaires généralisés (MLG ou par anglicisme, GLM) sont des modèles statistiques qui généralisent la régression linéaire (comme leur nom l'indique) et qui avaient pour vocation l'unification de modèles statistiques variés (dont la régression de Poisson ou celle logistique). Ils répondent en outre aux limites des modèles linéaires gaussiens qui sont :

- La variable à expliquer Y suit supposément une loi normale. Or, dans le cas par exemple des montants de soins, il est observé une asymétrie de la loi du montant qui est donc incompatible avec l'hypothèse de normalité.
- Les variables qualitatives ne sont pas explicables via ces modèles.
- La variance de la variable à expliquer est constante et ne dépend pas des variables explicatives (c'est la notion d'homoscédasticité).

Les GLM outrepassent ces trois limites citées tout en restant un modèle de type expliqué/explicatif.

Un GLM se compose de trois composantes :

- 1) La composante aléatoire du modèle : c'est la variable à expliquer Y , i.e. la variable réponse. Sa loi doit appartenir à la famille exponentielle.
Ex : Y représente le montant total des soins remboursés par la mutuelle pendant l'année.
- 2) La composante déterministe du modèle : ce sont les variables explicatives X_1, X_2, \dots, X_p .
 $X = (X_1, X_2, \dots, X_p)$ est le vecteur explicatif du modèle.
Ex : X_1 représente la classe d'âges notée 10 de l'assuré.
- 3) La fonction lien g du modèle : elle exprime la relation fonctionnelle qui existe entre la composante aléatoire du modèle et la composante déterministe du modèle. C'est une fonction déterministe, strictement monotone et définie sur \mathbb{R} .
Ex : g est la fonction lien logarithmique.

Il existe cependant des hypothèses à respecter pour pouvoir utiliser un tel modèle :

- 1) Soit Y_1, \dots, Y_n les valeurs prises par la variable à expliquer Y . Les Y_1, \dots, Y_n sont i.i.d.
- 2) La loi de Y n'est pas nécessairement la loi normale mais elle doit appartenir à la famille exponentielle.
- 3) Un GLM ne suppose pas de relation linéaire entre la variable à expliquer et les variables explicatives mais suppose une relation linéaire entre les variables explicatives et la transformation par la fonction lien de l'espérance conditionnée de la variable à expliquer. Cette

nuance implique que les variables explicatives peuvent se retrouver dans une écriture de modèle comme étant des termes d'exposants.

- 4) La variance dépend des variables explicatives et l'homoscédasticité est parfois difficile à atteindre via les modèles.
- 5) Les coefficients associés aux variables explicatives sont approximés par le « maximum de vraisemblance » (et non grâce à une méthode des moindres carrés ordinaires). De ce fait, les estimations sont réalisées sur des échantillons de taille supérieure.

Ainsi, un GLM permet d'étudier la liaison qu'il existe entre une variable à expliquer Y et un ensemble de variables explicatives $X = (X_1, X_2, \dots, X_p)$. Par exemple, dans nos travaux, nous chercherons à expliquer le coût moyen de remboursement de la mutuelle par des variables tels que l'âge de l'assuré ou encore, sa région de résidence, ...

Le modèle se pose alors comme suit : $g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$

où les paramètres β_0, \dots, β_p sont estimés par la méthode de « maximum de vraisemblance ».

13.2. La loi de la variable à expliquer

La loi de la variable réponse Y appartient à la famille exponentielle. Par conséquent, elle peut s'écrire sous la forme qui suit :

$$f_{\theta, \phi}(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \text{ avec } y \in S$$

Où :

- S est un sous-ensemble de \mathbb{N} ou de \mathbb{R} ;
- $\theta \in \mathbb{R}$ est appelé « paramètre canonique » ou « paramètre de la moyenne » et est inconnu ;
- $\phi \in \mathbb{R}$ est le « paramètre de dispersion » supposé connu ;
- $a(\cdot)$ est une fonction définie sur \mathbb{R} et est non nulle ;
- $b(\cdot)$ est une fonction définie sur \mathbb{R} , non nulle, deux fois dérivable et à dérivée première injective ;
- $c(\cdot)$ est une fonction définie sur \mathbb{R}^2 .

La loi normale, la loi binomiale, la loi de Poisson, la loi Gamma et la loi de Gaussienne inverse appartiennent à la famille exponentielle.

Bien que ϕ soit supposé connu, ce n'est pas toujours le cas et ce paramètre est alors appelé « paramètre de nuisance ». Il sera alors estimé et considéré par la suite comme étant connu.

Généralement, il est d'usage de considéré que :

$$a(\phi) = \frac{\phi}{\omega_i}$$

où ω_i est un poids connu *a priori* affecté aux observations Y_i de Y .

Nous supposons par la suite que $\omega_i = 1$.

Ci-dessous un tableau présentant les paramètres de la famille exponentielle pour des lois de probabilités usuelles.

Loi de probabilité	$\theta(\mu)$	$b(\theta)$	$a(\phi)$
Normale $\mathcal{N}(\mu, \sigma^2)$	μ	$\frac{\theta^2}{2}$	σ^2
Bernoulli $\mathcal{B}(1, \mu)$	$\log\left(\frac{\mu}{1-\mu}\right)$	$\log(1 + e^\theta)$	1
Poisson $\mathcal{P}(\mu)$	$\log(\mu)$	e^θ	1
Gamma $\mathcal{G}(\mu, \nu)$	$-\frac{1}{\mu}$	$-\log(-\theta)$	1
Gaussienne Inverse $\mathcal{L}\mathcal{G}(\mu, \sigma^2)$	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2

Tableau 52 : Paramètres de la famille exponentielle pour des lois usuelles

13.3. Moyenne et variance de la variable réponse

Pour une variable aléatoire Y dont la densité peut s'écrire sous forme exponentielle et avec l'hypothèse que $\omega_i = 1$, ses deux premiers moments peuvent s'écrire via les formules :

$$E[Y] = b'(\theta)$$

$$Var[Y] = b''(\theta) a(\phi)$$

Dans la formule de la variance, $b''(\theta)$ est appelée la fonction variance. Si nous notons $\mu = E[Y] = b'(\theta)$ alors $b''(\theta) = V(\mu)$.

Loi de probabilité	E[Y]	Var[Y]
Normale $\mathcal{N}(\mu, \sigma^2)$	θ	σ^2
Bernoulli $\mathcal{B}(1, \mu)$	$\frac{e^\theta}{1 + e^\theta}$	$\mu(1 - \mu)$
Poisson $\mathcal{P}(\mu)$	e^θ	μ
Gamma $\mathcal{G}(\mu, \nu)$	$-\frac{1}{\theta}$	$\frac{\mu^2}{2}$
Gaussienne Inverse $\mathcal{L}\mathcal{G}(\mu, \sigma^2)$	$\frac{1}{\sqrt{-2\theta}}$	$\mu^3 \sigma^2$

Tableau 53 : Espérance et variance de la famille exponentielle

13.4. Fonction de lien

La troisième composante des modèles linéaires généralisés est la fonction de lien. Cette fonction est déterministe, strictement monotone, inversible, définie sur \mathbb{R} telle que

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Cette équation peut être simplifiée par une forme vectorielle :

$$g(\mu) = \beta_0 + \sum_{i=1}^p \beta_i X_i$$

$$g(\mu) = X' \beta$$

$$g(\mu) = \eta[X]$$

η peut être vue comme une « variable synthétique », un résumé linéaire des variables explicatives appelé « score »

Chacune des lois de la famille exponentielle possède une fonction de lien spécifique, appelée fonction de lien canonique, permettant de relier l'espérance au paramètre θ . Ce paramètre est alors appelé paramètre canonique. Le choix de la fonction de lien dépend de la variable à expliquer et ne se porte pas toujours sur la fonction de lien canonique, il dépend de la variable à expliquer, du contexte des travaux, des connaissances *a priori* sur la nature des données. Ainsi, la fonction lien canonique n'est pas priorisée.

Le lien est tel que $g(\mu) = \theta$. Or, $\mu = b'(\theta)$ nous avons donc $g^{-1}(\cdot) = b'(\cdot)$.

Loi de probabilité	Type de données	Nom du lien	Fonction de lien
Normale $\mathcal{N}(\mu, \sigma^2)$		Lien identité	$g(\mu) = \mu$
Binomiale $\mathcal{B}(n, \mu)$	Pourcentage	Lien logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
Poisson $\mathcal{P}(\mu)$	Comptage	Lien log	$g(\mu) = \log(\mu)$
Gamma $\mathcal{G}(\mu, \nu)$	Durée	Lien inverse	$g(\mu) = \frac{1}{\mu}$
Gausse Inverse $\mathcal{LG}(\mu, \sigma^2)$			$g(\mu) = \frac{1}{\mu^2}$

Tableau 54 : Fonctions de lien

13.5. Estimation des paramètres par la méthode du maximum de vraisemblance

Dans les modèles linéaires généralisés, la méthode du maximum de vraisemblance est souvent utilisée pour estimer les paramètres. Considérons $Y = (Y_1, Y_2, \dots, Y_n)$ le vecteur à expliquer, X_1, X_2, \dots, X_p les vecteurs explicatifs et g la fonction de lien. θ_i peut être expliqué par les paramètres β_i via :

$$\theta_i = (b' \cdot g)^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})$$

Cette équation permet de lier les paramètres à estimer à la fonction de vraisemblance. Une fois que les β_i seront estimés, Y pourra être expliquée par $\hat{\beta}_i$.

La densité de chaque observation i s'écrit sous la forme suivante :

$$f_{\theta, \phi}(y_i) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right)$$

En supposant que toutes les observations i sont indépendantes et en tenant compte du fait que θ dépend de β , nous pouvons écrire la fonction de vraisemblance sous la formule :

$$\begin{aligned} \mathcal{L}(\theta(\beta), y, \phi) &= \ln\left(\prod_{i=1}^n f_{\theta, \phi}(y_i)\right) \\ &= \sum_{i=1}^n \ln(f_{\theta, \phi}(y_i)) \\ &= \sum_{i=1}^n \frac{\omega_i (y_i \theta_i - b(\theta_i))}{\phi} + \sum_{i=1}^n c(y_i, \phi) \end{aligned}$$

Il est souhaité de chercher les valeurs du vecteur β qui maximisent la vraisemblance de l'équation précédente. Cela revient à chercher les paramètres β_i tel que :

$$\frac{\partial \mathcal{L}(\theta(\beta), y, \phi)}{\partial \beta_{ij}} = 0, \quad j = 0, \dots, p$$

Autrement dit :

$$\sum_{i=1}^n \frac{\partial \ln(f_{\theta, \phi}(y_i))}{\partial \beta_{ij}} = 0$$

$$\sum_{i=1}^n \frac{\partial}{\partial \beta_j} \left(\frac{\omega_i(y_i \theta_i - b(\theta_i))}{\phi} + c(y_i, \phi) \right) = 0$$

Avec l'hypothèse d'indépendance, pour chaque i :

$$\frac{\partial \ln(f_{\theta, \phi}(y_i))}{\partial \beta_j} = \frac{\partial \ln(f_{\theta, \phi}(y_i))}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$

Les 4 composantes de cette équation sont calculées séparément et il est obtenu :

$$\frac{\partial \ln(f_{\theta, \phi}(y_i))}{\partial \theta_i} = \frac{\omega_i(y_i - b'(\theta_i))}{\phi}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

Les équations de vraisemblance retenues sont données par :

$$\frac{\partial \mathcal{L}(\theta(\beta), y, \phi)}{\partial \beta_{ij}} = \sum_{i=1}^n \frac{\omega_i(y_i - b'(\theta_i))x_{ij}}{\phi b''(\theta_i)g'(\mu_i)} = 0, \quad j = 0, \dots, p$$

Ce qui revient à :

$$\sum_{i=1}^n \frac{\omega_i(y_i - b'(\theta_i))x_{ij}}{b''(\theta_i)g'(\mu_i)} = 0, \quad j = 0, \dots, p$$

En général, les équations de la plupart des modèles linéaires généralisés qui déterminent leurs paramètres par maximum de vraisemblance ne sont pas linéaires. De plus, les estimateurs des paramètres ne présentent pas de formule fermée et facilement solvable. Il n'y a donc pas de solution explicite. Pour trouver les estimateurs β_i , il est possible d'utiliser des méthodes itératives comme Newton-Raphson ou la méthode du Score de Fisher. Des approximations successives permettent de s'approcher des estimations au sens du maximum de vraisemblance.

Annexe 14 : Mesures de corrélation

Afin de pouvoir utiliser la méthode GLM fréquence x coût moyen, il a d'abord été testé, pour chaque segment de tarification concerné, l'indépendance entre la variable de fréquence et de coût. Pour cela, plusieurs mesures de corrélation ont été mentionnées :

- Le coefficient de corrélation linéaire (non utilisée)
- Le coefficient de Pearson
- Le Tau de Kendall (τ de Kendall)
- Le Rho de Spearman (ρ de Spearman)

14.1. Coefficient de corrélation linéaire

Ce coefficient se définit comme suit :

$$c_{lin}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x \sigma_y}$$

Où

- X et Y sont deux variables aléatoires (par exemple, X est la fréquence et Y le coût moyen) ;
- n est le nombre d'observations considérées ;
- x_i désigne une observation (i.e. une valeur prise) de X et y_i désigne une observation (i.e. une valeur prise) de y ;
- \bar{x} est la moyenne des valeurs prises par X et \bar{y} est la moyenne des valeurs prises par Y ;
- σ_x et σ_y sont les écarts-types de X et Y respectivement.

Ce coefficient présente quelques inconvénients tels que :

- La possibilité de ne capter seulement que les corrélations linéaires ;
- Il nécessite une variance finie pour X et Y
- Il existe des cas où le coefficient de corrélation linéaire peut être proche de 0 alors que la corrélation est forte (notamment dans les cas où la variance de X est élevée alors que celle de Y est faible) ;
- Il dépend des lois marginales (loi de X et de Y).

14.2. Coefficient de Pearson

Ce coefficient est un estimateur empirique du coefficient de corrélation linéaire. Il se définit de la manière suivante :

$$c_{lin}(\widehat{X}, \widehat{Y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Avec les mêmes notations que précédemment.

Plus ce coefficient est proche de 0 et plus il est justifié de pencher vers l'indépendance des variables X et Y. Étant donné qu'il reste un estimateur empirique du coefficient précédemment présenté, les mêmes remarques d'inconvénients s'appliquent.

14.3. Tau de Kendall

Ce coefficient se définit comme suit :

$$\tau(X, Y) = \underbrace{\mathbb{P}[(X - \tilde{X})(Y - \tilde{Y}) > 0]}_{\text{probabilité de concordance}} - \underbrace{\mathbb{P}[(X - \tilde{X})(Y - \tilde{Y}) < 0]}_{\text{probabilité de discordance}}$$

Où

- X et Y sont deux variables aléatoires (par exemple, X est la fréquence et Y le coût moyen) ;
- (\tilde{X}, \tilde{Y}) est une copie indépendante de (X, Y)

La version empirique du Tau de Kendall est la suivante :

$$\hat{\tau}(X, Y) = \frac{\text{nombre de paires d'observations concordantes} - \text{nombre de paires d'observations discordantes}}{\text{nombre de paires d'observations}}$$

Deux paires d'observations (X, Y) et (\tilde{X}, \tilde{Y}) sont concordantes quand, par exemple, $X > \tilde{X}$ et $Y > \tilde{Y}$. Elles sont discordantes quand, par exemple $X > \tilde{X}$ et $Y < \tilde{Y}$.

Il peut y avoir une occurrence de problème si, par exemple, X égalise avec \tilde{X} et ainsi, sous-estimer la corrélation.

14.4. Rho de Spearman

Ce coefficient se définit comme suit :

$$\rho(X, Y) = c_{lin}(F_X(X), F_Y(Y))$$

Où

- X et Y sont deux variables aléatoires (par exemple, X est la fréquence et Y le coût moyen) ;
- F_X et F_Y sont les fonctions de répartition respectivement de X et Y.

Le coefficient de Spearman permet de détecter des tendances monotones. Lorsque la tendance est affine, il se comporte de façon similaire au coefficient de Pearson.

Annexe 15 : Autres compléments à la méthode GLM

Il s'agit ici d'aborder toutes les notions complémentaires qui ne font pas l'objet d'une annexe qui leur est propre et qui interviennent dans les étapes d'une tarification par un GLM.

15.1. Variance Inflation factor (VIF) et Generalized Variance Inflation factor (GVIF)

Appelé aussi facteur d'inflation de la variance, le VIF ne permet avant tout qu'à détecter un problème de multicolinéarité (et non pas à quantifier de manière exacte la corrélation entre des variables). Il estime de combien la variance d'un coefficient (d'un modèle de régression) associé à une variable explicative est « augmentée » du fait de la relation linéaire de cette variable avec les autres du modèle. Plus concrètement, un VIF égal à 5 pour un facteur de risque donné signifie que la variance du coefficient estimé qui lui est associé dans le modèle de régression est 5 fois supérieure à la variance qui aurait dû être observée si ce facteur de risque n'était pas corrélé aux autres facteurs.

À chaque variable explicative peut donc être associé un VIF. Ainsi, pour la variable explicative X_j , le VIF est défini de la manière suivante :

$$VIF_j = \frac{1}{1 - R_j^2}$$

Où R_j^2 est le coefficient de détermination du modèle de régression linéaire où X_j est expliquée par toutes les autres variables explicatives (et où la variable à expliquer n'intervient pas).

Le VIF est compris entre 1 et $+\infty$. Tout comme pour les coefficients de corrélation de Kendall, Pearson, Spearman, il n'y a pas de consensus clair sur la valeur seuil au-delà de laquelle il faut considérer l'existence d'un problème de multicolinéarité. Il est cependant d'usage d'interpréter le VIF [18] de la manière suivante :

- Un VIF égal à 1 (cas idéal car $R_j^2 = 0$) indique qu'il n'y a pas de corrélation entre une variable explicative donnée et n'importe laquelle des autres variables explicatives du modèle ;
- Un VIF compris entre 1 et 5 indique qu'il existe une corrélation modérée entre une variable explicative donnée et n'importe laquelle des autres variables explicatives du modèle. Pour la plupart du temps, cela n'est pas suffisamment significatif pour que les résultats soient impactés ;
- Un VIF supérieur à 5 indique qu'il y a potentiellement une forte corrélation entre une variable explicative donnée et n'importe laquelle des autres variables explicatives du modèle. Dans ce cas-ci, les coefficients estimés ainsi que les p-value des tests de significativité issus de la modélisation ne sont pas fiables.

Sur *R*, le VIF est calculé via la fonction *vif()* de la librairie *car*. Cette fonction permet de calculer le VIF à partir d'un modèle de régression linéaire.

Il est néanmoins important de noter que le calcul du VIF tel que présenté précédemment n'est faisable que dans le cas de variables quantitatives. Dans les travaux, les variables explicatives sont qualitatives et les modèles utilisés sont des GLM (et non des modèles linéaires gaussiens). Il faut dans ce cas-là utiliser la version généralisée du VIF, telle qu'explicitée par John FOX et Georges MONETTE (*Generalized Collinearity Diagnostics, Journal of the American Statistical Association*, 87, pages 178 à 183, 1992).

Dans *R*, cette version généralisée est déjà implémentée dans la fonction *vif()*, qui s'adapte automatiquement aux modèles linéaires généralisés.

Le GVIF pour les variables quantitatives continues est égal au VIF.

Il peut être associé un GVIF à chaque variable catégorielle. L'interprétation des valeurs du VIF peut être transposée à l'identique en considérant $\left(GVIF_{2DF}^{\frac{1}{2}}\right)^2$ [19]. Cela signifie que l'interprétation pour « VIF > 5 » est équivalente à celle pour « $\left(GVIF_{2DF}^{\frac{1}{2}}\right)^2 > 5$ ».

Retour
p.79

15.2. Déviance

La déviance est un indicateur qui permet d'évaluer la qualité d'ajustement d'un modèle linéaire généralisé aux données observées.

L'idée est de comparer le modèle estimé au modèle avec le meilleur ajustement possible (celui qui aurait autant d'observations que de paramètres à estimer), soit, au modèle dit saturé. Ce dernier modèle n'est cependant pas souvent intéressant à modéliser car il reproduit la réalité au lieu de la résumer.

La comparaison du modèle estimé (via le GLM) au modèle saturé repose sur la comparaison de leurs vraisemblances respectives, notées \mathcal{L} et \mathcal{L}_{sat} . Plus ces dernières sont proches et plus l'ajustement du modèle est de qualité. Le rapport de vraisemblance suivant est alors considéré :

$$\lambda = \frac{\mathcal{L}_{sat}}{\mathcal{L}}$$

Ou encore, de manière équivalente, en utilisant la log-vraisemblance :

$$\ln(\lambda) = \ln(\mathcal{L}_{sat}) - \ln(\mathcal{L})$$

En particulier, pour tester l'ajustement, il est généralement considéré deux statistiques :

- La déviance réduite : $D = 2\ln(\lambda)$ (la statistique d'intérêt)
- La déviance non réduite : $D^* = \phi D$ où ϕ est le paramètre de dispersion de la famille exponentielle liée à la variable à expliquer Y dans le modèle estimé considéré.

Une déviance proche de zéro (i.e. \mathcal{L}_{sat} proche de \mathcal{L}) signifie que le modèle saturé n'apporte pas tant plus d'informations que le modèle estimé. Au contraire, une déviance élevée signifie que le modèle estimé est loin du modèle saturé et qu'il n'est donc pas adéquat.

Par ailleurs, comme sous H_0 , $D \xrightarrow{\mathcal{L}} \chi_{n-p-1}^2$ (quantile d'une loi du Khi-deux avec n-p-1 degrés de liberté, où n est le nombre d'observations et p le nombre de variables explicatives ou de modalités de variables explicatives), il est formellement admis que la qualité d'ajustement du modèle estimé est mauvaise si la condition suivante est vraie :

$$D_{obs} > \chi_{n-p-1, 1-\alpha}^2$$

Où :

- D_{obs} est la valeur empirique observée de D ;
- $\chi_{n-p-1, 1-\alpha}^2$ est le quantile d'ordre $1 - \alpha$ d'une loi de χ^2 avec n-p-1 degrés de liberté.

Une règle empirique est parfois aussi considérée : l'ajustement du modèle est bon si $\frac{D_{obs}}{n-p-1} \approx 1$.

En pratique, dans R, `summary()` affiche deux types de déviances :

- La déviance nulle est celle qui se calcule en considérant le modèle saturé et le modèle dit « nul » (celui où il n'y a qu'un paramètre à estimer et où la variable d'intérêt serait expliquée par β_0).

Une faible déviance nulle signifie que le modèle nul présente une bonne qualité d'ajustement aux données.

- La déviance résiduelle est celle qui se calcule en considérant le modèle saturé et le modèle estimé par le GLM. C'est la déviance qui a été présentée en début de sous-partie.

La différence entre la déviance nulle et la déviance résiduelle suit une loi de χ^2 avec comme degrés de liberté la différence de degrés de liberté entre le modèle estimé et le modèle « nul ».

Néanmoins, ajouter une variable au modèle entraîne une augmentation mécanique de la déviance (il y a plus d'information dans le modèle). Il est donc préférable de comparer des modèles ayant le même nombre de variables explicatives.

Ce rapport de vraisemblance peut être aussi utilisé pour tester la significativité des coefficients estimés par un GLM (en comparant la vraisemblance du modèle avec la variable associée au coefficient testé et la vraisemblance du modèle sans).

Retour

p.79

Retour

p.82

15.3. Critère BIC

Le critère BIC (Bayesian Information Criterion) ou le critère AIC (Akaike Information Criterion) sont deux critères qui permettent de choisir un modèle parmi plusieurs en fournissant une base sur laquelle les comparer. Ils constituent tous deux des mesures d'ajustement de modèle et de sa qualité sur la base de la log-vraisemblance.

Le BIC est défini par :

$$BIC = 2 \ln(\mathcal{L}) + k \ln(n)$$

Où :

- \mathcal{L} est la vraisemblance du modèle estimé ;
- n est le nombre d'observations ;
- k est le nombre de paramètres.

Plus le BIC est faible et meilleure est l'adéquation du modèle.

Contrairement à la déviance, le critère BIC pénalise aussi la complexité du modèle (via le paramètre k qui est introduit dans sa formule de calcul) i.e. plus le nombre de paramètres est élevé, plus le BIC sera grand. Le critère AIC pénalise aussi la complexité du modèle mais de manière moins importante que le critère BIC. Le BIC est privilégié pour de la prédiction.

Retour

p.79

Retour

p.86

15.4. Test d'indépendance du Khi-deux

Ce test permet de déterminer le lien existant entre deux variables qualitatives, notées X et Y , ayant respectivement p et q modalités.

L'hypothèse nulle du test est H_0 : X et Y sont indépendantes. Soit n_{ij} le nombre d'observations présentant la modalité i pour la variable X et la modalité j pour la variable Y . On note :

$$n_{i.} = \sum_{j=1}^q n_{ij} \quad n_{.j} = \sum_{i=1}^p n_{ij} \quad n = \sum_{i,j} n_{ij}$$

On note aussi, pour $i \in [1, p]$ et $j \in [1, q]$:

$$\begin{aligned} p_i^X &= \mathbb{P}(X = i) \\ p_j^Y &= \mathbb{P}(Y = j) \\ p_{ij} &= \mathbb{P}(X = i, Y = j) \end{aligned}$$

Sous H_0 , on a que $p_{ij} = p_i^X p_j^Y$.

Sous H_0 , l'effectif théorique attendu présentant les modalités i et j pour respectivement les variables X et Y est $\frac{n_i n_j}{n}$.

La statistique du test est la suivante :

$$Z^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}} \sim \chi^2_{(p-1)(q-1)}$$

Cela revient en fait à calculer génériquement :

$$Z^2 = \sum \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

Si $Z^2 > \chi^2_{(p-1)(q-1), 1-\alpha}$ (où $\chi^2_{(p-1)(q-1), 1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi de χ^2 avec $(p-1)(q-1)$ degrés de liberté), alors il y a rejet de l'hypothèse nulle

Retour
p.80

15.5. Tests d'adéquation pour les lois discrètes

Le test considéré est le test d'adéquation du Khi-deux (non-paramétrique) qui permet de comparer la distribution théorique envisagée à la distribution empirique des données.

L'hypothèse nulle H_0 est la suivante : la variable aléatoire X suit une loi de probabilité \mathcal{P} . X est par exemple la fréquence, le coût moyen ou encore, la consommation.

Pour cela, il faut répartir les n observations de X en k classes et définir p_j la probabilité qu'une observation donnée de X appartienne à la $j^{\text{ème}}$ classe ($j \in \llbracket 1, \dots, k \rrbracket$). Chaque classe présente un effectif noté n_j . Ainsi :

$$n = \sum_{j=1}^k n_j$$

Le principe intrinsèque reste le même que pour le test d'indépendance du Khi-deux, à savoir, la statistique suivante :

$$Z^2 = \sum \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

Cependant, cette fois-ci, les effectifs théoriques attendus sont les np_j . Ainsi :

$$Z^2 = \sum_{j=1}^k \frac{(n_j - np_j)^2}{np_j}$$

Si $Z^2 > \chi^2_{k-1, 1-\alpha}$ (où $\chi^2_{k-1, 1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi de χ^2 avec $k-1$ degrés de liberté), alors il y a rejet de l'hypothèse nulle.

Il existe cependant pour ce test une condition d'application : pour tout j , $np_j \geq 5$. Son principal inconvénient est qu'il nécessite des échantillons grands de données.

Retour
p.78

15.6. Tests d'adéquation pour les lois continues

Deux tests d'adéquation de lois ont été considérés : le test de Kolmogorov-Smirnov et le test de Cramer-von Mises.

Pour les deux tests, l'hypothèse nulle H_0 est la suivante : $F = \widehat{F}_n$

Où :

- F est la fonction de répartition théorique de la loi candidate à tester ;
- \widehat{F}_n est la fonction de répartition empirique des données observées.

Pour rappel :

Soit X la variable aléatoire d'intérêt (par exemple, la fréquence ou le coût moyen). Soit $\{X_1, X_2, \dots, X_n\}$ des observations de X (suite de variables i.d.d. à X).

Alors la fonction de répartition empirique de X est définie de la manière suivante :

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \text{ avec } x \in \mathbb{R}$$

Dans le cas du test de Kolmogorov-Smirnov, la statistique du test est la suivante :

$$KS = \sqrt{n} \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)|$$

Ou plus concrètement :

$$KS = \max_{i=1, \dots, n} \left(\left| F(X_{(i)}) - \frac{i}{n} \right|, \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right)$$

Où :

- $X_{(i)}$ sont les statistiques d'ordre des observations de X (valeurs rangées par ordre croissant) ;
- n le nombre d'observations considérées.

L'inconvénient de ce test est cependant sa sensibilité aux valeurs extrêmes.

Le test de Cramer-von Mises y est moins sensible et c'est une alternative au test de Kolmogorov-Smirnov. Les deux se basent sur l'écart entre les deux fonctions de répartition d'intérêt (celle théorique et celle empirique) mais le test de Kolmogorov-Smirnov se base sur la valeur maximale de cette différence alors que le test de Cramer-von Mises se base sur la somme des différences.

Sa statistique de test est la suivante :

$$CvM = n \int_{-\infty}^{+\infty} (\widehat{F}_n(X) - F(X))^2 dF(X)$$

Ou encore :

$$CvM = \sum_{i=1}^n \left(F(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$$

Avec les mêmes notations que pour le test de Kolmogorov-Smirnov.

Ces statistiques sont ensuite comparées à une valeur critique pour déterminer le rejet ou non de l'hypothèse nulle.

15.7. Test de significativité des coefficients estimés

Il existe plusieurs manières de réaliser des tests de significativité mais le principe reste le même : déterminer si un coefficient β_i estimé dans un GLM est significativement différent de 0 ou non. Si le coefficient est non significatif dans le modèle, il est possible d'abandonner la variable du modèle qui lui est associée quand cette dernière est quantitative, ou de regrouper des modalités dans le cas de variables qualitatives.

Lorsque la fonction `glm()` dans R est utilisée et qu'un `summary()` est ensuite appliqué au modèle, deux informations sont visibles pour chaque coefficient estimé :

- La *z-value* ou *t-value* (interprétable comme la statistique de test) ;
- La *p-value* ($\Pr(> |z|)$ ou $\Pr(> |t|)$).

L'affichage de la *z-value* ou *t-value* dépend de l'estimation ou non du paramètre ϕ de surdispersion par le GLM (en cas d'estimation du paramètre, c'est la *t-value* qui est affichée). Selon la valeur affichée, la *p-value* indiquée est soit reliée à un test de Student (dans le cas d'une *t-value*), soit à un test basé sur une loi Normale (dans le cas d'une *z-value*).

Plus concrètement, l'hypothèse nulle du test de significativité est $H_0 : \beta_i = 0$ ($i = 0, 1, \dots, p$).

Sous H_0 , dans le cas où la *z-value* est affiché sous R, il est considéré une variable aléatoire Z suivant une loi normale centrée réduite.

Ainsi :

$$p_{value} = 2 \mathbb{P}(Z > |z|)$$

Où z est la statistique du test i.e. $z = \frac{\beta_i}{\sigma(\beta_i)}$ avec $\sigma(\beta_i)$ l'erreur standard du coefficient estimé.

z est en fait le quotient de « *Estimate* » par « *Std.Error* ».

Similairement, sous H_0 , dans le cas où la *t-value* est affiché sous R, il est considéré une variable aléatoire T suivant une loi de Student avec $n-p-1$ degré de liberté (n étant le nombre d'observations et p le nombre de variables explicatives ou de modalités de variables explicatives).

Ainsi :

$$p_{value} = 2 \mathbb{P}(T > |t|)$$

Où t est la statistique du test i.e. $t = \frac{\beta_i}{\sigma(\beta_i)}$ avec $\sigma(\beta_i)$ l'erreur standard du coefficient estimé.

t est en fait le quotient de « *Estimate* » par « *Std.Error* ».

Pour déterminer le rejet ou non de l'hypothèse nulle, il convient de comparer la *p-value* calculée avec le niveau de confiance α considéré ($\alpha = 0,05$ par exemple).

Retour
p.82

15.8. Les résidus

Les notations utilisées dans l'Annexe 13 détaillant le modèle linéaire généralisé sont reprises ici.

Il existe différents types de résidus. Il s'agit de définir ceux utilisés dans le mémoire.

- Les résidus de Pearson sont définis par :

$$r_i^p = \frac{\sqrt{\omega_i}(y_i - \mu_i)}{\sqrt{V(\mu_i)}}$$

Pour rappel, $\mu_i = E(Y_i)$, V est la fonction variance et ω_i est souvent considéré égal à 1.

Pour ces résidus, $E(r_i^p) = 0$ et $Var(r_i^p) = \frac{\omega_i Var(Y_i)}{V(\mu_i)} = \frac{\omega_i (V(\mu_i) \frac{\phi}{\omega_i})}{V(\mu_i)} = \phi$ (homoscédasticité).

- Les résidus de déviance sont définis par :

$$r_i^D = \text{signe}(y_i - \mu_i) \sqrt{d_i}$$

Où d_i est la contribution de la $i^{\text{ème}}$ observation y_i à la déviance D du modèle.

La déviance D a été définie comme une mesure de la qualité d'ajustement du modèle. Il peut être imaginé que chaque observation y_i contribue à hauteur d'une quantité d_i à la déviance D de telle sorte que :

$$D = \sum_{i=1}^n d_i$$

Ainsi,

$$D = \sum_{i=1}^n (r_i^D)^2$$

Si r_i^D est grand (i.e. si $|r_i^D| > 1$), cela signifie que la $i^{\text{ème}}$ observation contribue au mauvais ajustement du modèle.

Annexe 16 : Lois usuelles

16.1. Loi de Poisson

Soit N une variable aléatoire de comptage dont la distribution suit une loi de Poisson de paramètre $\lambda > 0$. On note $N \sim Poi(\lambda)$.

$$p_k = \mathbb{P}(N = k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ avec } k \in \mathbb{N}$$

$$\mathbb{E}(N) = \lambda$$

$$Var(N) = \lambda$$

Propriété notable :

Si, N_1, N_2, \dots, N_m sont des variables indépendantes suivant des lois de Poisson de paramètre $\lambda_1, \lambda_2, \dots, \lambda_m$ respectivement, alors :

$$N = \sum_{i=1}^m N_i \sim Poi\left(\sum_{i=1}^m \lambda_i\right)$$

16.2. « Loi » quasi-Poisson (ou Poisson surdispersée)

La loi de Poisson présente l'hypothèse sous-jacente de l'égalité entre espérance et variance.

En cas de surdispersion constatée des données (i.e. $Var(N) \geq \mathbb{E}(N)$), il est possible de considérer une « loi » quasi-Poisson (qui n'est en réalité pas une loi en tant que telle) au lieu de la loi de Poisson.

On dit que la variable aléatoire N suit une loi quasi-Poisson de paramètres $\lambda > 0$ et $\phi \in \mathbb{R}$ si N/ϕ suit une loi de Poisson de paramètre λ/ϕ .

Celle-ci a ainsi pour particularité de considérer que :

$$Var(N) = \phi \mathbb{E}(N)$$

Où ϕ est le paramètre de surdispersion de la famille exponentielle associée à la variable à expliquer.

Par ailleurs, il est possible de montrer que les coefficients estimés dans les modèles linéaires généralisés via la « loi » quasi-Poisson sont exactement les mêmes que ceux via la loi Poisson avec des écarts-types différents.

16.3. Loi binomiale négative

Soit N une variable aléatoire de comptage dont la distribution suit une loi binomiale négative de paramètres $\alpha > 0$ et $\beta \in [0,1]$. On note $N \sim NB(\alpha, \beta)$.

$$p_k = \mathbb{P}(N = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(k + 1)} (1 - \beta)^\alpha \beta^k \text{ avec } k \in \mathbb{N}$$

$$\mathbb{E}(N) = \frac{\alpha\beta}{1 - \beta}$$

$$Var(N) = \frac{\alpha\beta}{(1 - \beta)^2}$$

Propriété notable :

Si, N_1, N_2, \dots, N_m sont des variables indépendantes suivant des lois binomiales négatives de paramètres $(\alpha_1, \beta), (\alpha_2, \beta), \dots, (\alpha_m, \beta)$ respectivement, alors :

$$N = \sum_{i=1}^m N_i \sim NB \left(\sum_{i=1}^m \alpha_i, \beta \right)$$

La loi binomiale négative fait bien partie de la famille exponentielle avec $\theta = \ln(\beta)$, $b(\theta) = -a \ln(\beta)$ et $a(\phi) = 1$ [20].

Retour

p.78

16.4. Loi Gamma

Soit X une variable aléatoire qui suit une loi Gamma de paramètres $\alpha > 0$ et $\beta > 0$. On note $X \sim \Gamma(\alpha, \beta)$.

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ pour tout } x \geq 0$$

$$\mathbb{E}(X) = \frac{\alpha}{\beta}$$

$$\text{Var}(X) = \frac{\alpha}{\beta^2}$$

16.5. Loi log-normale

Soit X une variable aléatoire qui suit une loi log-normale de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 > 0$. On note $X \sim \mathcal{LN}(\mu, \sigma^2)$.

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \text{ pour tout } x > 0$$

$$\mathbb{E}(X) = e^{\mu + \frac{\sigma^2}{2}}$$

$$\text{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

La loi log-normale appartient à la famille exponentielle généralisée [21] : c'est le logarithme de X qui suit une loi normale de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ et qui appartient à la famille exponentielle.

16.6. Loi de Weibull

Soit X une variable aléatoire qui suit une loi de Weibull de paramètres $\alpha > 0$ et $\beta > 0$. On note $X \sim \text{Weib}(\alpha, \beta)$.

$$f_X(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha} \text{ pour tout } x > 0$$

$$\mathbb{E}(X) = \beta \Gamma\left(1 + \frac{1}{\alpha}\right)$$

$$\text{Var}(X) = \beta^2 \left(1 + \frac{2}{\alpha}\right) - E(X)^2$$

La loi de Weibull à paramètre de forme α fixé appartient à la famille exponentielle [22].

Retour

p.85

16.7. Loi du χ^2

Soit X une variable aléatoire qui suit une loi du Khi-deux avec $k \in \mathbb{N}^*$ degrés de liberté. On note $X \sim \chi_k^2$.

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \text{ pour tout } x \geq 0$$

La loi du Khi-deux avec k degrés de liberté est la loi de la somme de carrés de k lois normales centrées réduites indépendantes.

$$\mathbb{E}(X) = k$$

$$\text{Var}(X) = 2k$$

16.8. Famille de distribution Tweedie [20]

Tweedie (1984) a suggéré la famille suivante :

$$f(y, \mu, \phi) = A(y, \phi) e^{\frac{1}{\phi}[(y\theta(\mu) - \kappa(\theta(\mu)))]}$$

Où :

$$\theta(\mu) = \begin{cases} \frac{\mu^{1-\gamma}}{1-\gamma} & \gamma \neq 1 \\ \ln(\mu) & \gamma = 1 \end{cases} \text{ et } \kappa(\theta(\mu)) = \begin{cases} \frac{\mu^{2-\gamma}}{2-\gamma} & \gamma \neq 2 \\ \ln(\mu) & \gamma = 2 \end{cases}$$

La loi de Y est de ce fait une loi de Poisson composée, avec des sauts Gamma. On note :

$$Y \sim \mathcal{CPoi}\left(\mu^{2-\gamma} \phi(2-\gamma), \Gamma\left(-\frac{2-\gamma}{\phi(1-\gamma)}, \phi(2-\gamma)\mu^{\gamma-1}\right)\right)$$

Où $\gamma \in [1,2]$ et $\mu = E(Y)$.

La fonction de variance est de la forme $V(\mu) = \phi\mu^\gamma$.

Cette famille est une sous-classe de la famille exponentielle.

Les distributions de Tweedie sont entièrement définies par deux paramètres (l'espérance et la variance). Par ailleurs, selon le paramètre γ , des lois usuelles sont retrouvables :

- Quand $\gamma \rightarrow 0$, la distribution de Tweedie se comporte comme une loi normale ;
- Quand $\gamma \rightarrow 1$ (ou $\alpha \rightarrow \infty$), la distribution de Tweedie se comporte comme une loi de Poisson ;
- Quand $\gamma \rightarrow 2$ (ou $\alpha \rightarrow 0$), la distribution de Tweedie se comporte comme une loi Gamma.

De plus :

- Quand $1 < \gamma < 2$, la distribution de Tweedie se comporte comme une loi composée Poisson-Gamma ;
- Quand $2 < \gamma < 3$, la distribution de Tweedie se comporte comme une distribution positive stable (appelée aussi distribution de Lévy tronquée α -stable) ;
- Quand $\gamma \rightarrow 3$, la distribution de Tweedie se comporte comme une loi Gaussienne inverse.

Annexe 17 : Résultats du benchmarking

17.1. Pour la garantie A

Coefficients de déformation effet région - Garantie A - Idée 1								
Classe âge	0	20	30	40	50	60	70	80
Région O1	64%	68%	72%	70%	72%	75%	81%	90%
Région O2	93%	93%	97%	95%	95%	94%	95%	99%
Région O3	87%	89%	87%	87%	87%	88%	90%	89%
Région O4	91%	93%	92%	91%	91%	90%	92%	92%
Région O5	91%	87%	88%	88%	88%	88%	90%	91%
Région O6	96%	89%	92%	93%	94%	96%	99%	106%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	93%	90%	88%	88%	89%	88%	91%	93%
Région O8	94%	89%	87%	86%	86%	85%	87%	91%
Région O9	99%	96%	94%	92%	93%	94%	96%	99%
Région O10	100%	98%	96%	95%	95%	95%	98%	106%
Région O11	99%	99%	95%	95%	95%	95%	96%	98%
Région O12	104%	106%	104%	102%	101%	99%	99%	110%

Coefficients de déformation effet région - Garantie A (GLM non optimisés) - Idée 2								
Classe âge	0	20	30	40	50	60	70	80
Région O1	64%	71%	74%	71%	75%	83%	92%	115%
Région O2	88%	92%	96%	94%	93%	98%	98%	119%
Région O3	84%	88%	85%	85%	87%	89%	92%	90%
Région O4	85%	92%	90%	90%	91%	91%	93%	95%
Région O5	86%	86%	87%	87%	88%	90%	94%	94%
Région O6	97%	88%	91%	92%	95%	94%	101%	111%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	93%	90%	86%	85%	88%	90%	93%	95%
Région O8	91%	90%	86%	83%	85%	85%	87%	88%
Région O9	99%	96%	94%	91%	93%	99%	101%	106%
Région O10	106%	101%	99%	99%	97%	108%	108%	127%
Région O11	105%	101%	96%	96%	94%	100%	98%	104%
Région O12	113%	111%	113%	109%	106%	114%	109%	139%

Exemple 1 - Assureur 1								
Classe âge	0	20	30	40	50	60	70	80
Région O1								
Région O2	108%	110%	108%	107%	107%	108%	107%	107%
Région O3	100%	100%	97%	98%	98%	98%	98%	98%
Région O4	100%	100%	100%	100%	100%	100%	100%	100%
Région O5	100%	100%	97%	98%	98%	98%	98%	98%
Région O6	100%	100%	100%	100%	100%	100%	100%	100%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	100%	100%	97%	98%	98%	98%	98%	98%
Région O8	100%	100%	97%	98%	98%	98%	98%	98%
Région O9	100%	100%	97%	98%	98%	98%	98%	98%
Région O10	104%	103%	103%	102%	102%	103%	101%	103%
Région O11	100%	100%	100%	100%	100%	100%	100%	100%
Région O12	104%	103%	103%	102%	102%	103%	101%	103%

Exemple 2 - Assureur 2								
Classe âge	0	20	30	40	50	60	70	80
Région O1	96%	96%	94%	97%	96%	95%	95%	96%
Région O2	100%	100%	100%	100%	100%	100%	100%	100%
Région O3	91%	93%	91%	92%	91%	91%	90%	90%
Région O4	100%	100%	100%	100%	100%	100%	100%	100%
Région O5	83%	82%	80%	82%	82%	81%	81%	81%
Région O6	83%	82%	80%	82%	82%	81%	81%	81%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	83%	82%	80%	82%	82%	81%	81%	81%
Région O8	83%	82%	80%	82%	82%	81%	81%	81%
Région O9	91%	93%	91%	92%	91%	91%	90%	90%
Région O10	96%	96%	94%	97%	96%	95%	95%	96%
Région O11	91%	93%	91%	92%	91%	91%	90%	90%
Région O12	91%	93%	91%	92%	91%	91%	90%	90%

Tableau 55 : Plusieurs tables de coefficients de déformation pour la garantie A

17.2. Pour la garantie B

Coefficients de déformation effet région - Garantie B - Idée 1								
Classe âge	0	20	30	40	50	60	70	80
Région O1	61%	69%	74%	74%	72%	74%	77%	86%
Région O2	104%	106%	111%	107%	106%	104%	104%	104%
Région O3	87%	88%	88%	91%	90%	90%	90%	89%
Région O4	89%	92%	91%	93%	92%	91%	92%	93%
Région O5	88%	85%	87%	90%	89%	88%	89%	90%
Région O6	93%	87%	91%	94%	94%	96%	98%	104%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	89%	88%	87%	90%	89%	88%	89%	92%
Région O8	90%	86%	85%	87%	85%	84%	85%	89%
Région O9	97%	95%	94%	95%	94%	94%	96%	98%
Région O10	96%	95%	95%	95%	95%	95%	97%	104%
Région O11	102%	103%	100%	100%	100%	99%	100%	101%
Région O12	103%	106%	105%	102%	102%	100%	100%	110%

Coefficient de déformation effet région - Garantie B (GLM non optimisés) - Idée 2								
Classe âge	0	20	30	40	50	60	70	80
Région O1	60%	71%	78%	77%	75%	84%	92%	113%
Région O2	98%	101%	108%	105%	106%	106%	108%	133%
Région O3	86%	91%	88%	92%	91%	93%	93%	92%
Région O4	88%	93%	92%	93%	93%	93%	94%	98%
Région O5	87%	87%	91%	92%	91%	93%	94%	95%
Région O6	94%	87%	92%	95%	94%	97%	100%	107%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	91%	91%	90%	91%	91%	93%	94%	97%
Région O8	91%	88%	87%	88%	86%	86%	87%	86%
Région O9	97%	99%	98%	97%	96%	101%	102%	107%
Région O10	99%	106%	105%	104%	103%	111%	113%	132%
Région O11	102%	105%	101%	102%	101%	102%	103%	109%
Région O12	105%	111%	111%	109%	108%	112%	114%	147%

Exemple 1 - Assureur 1								
Classe âge	0	20	30	40	50	60	70	80
Région O1								
Région O2	126%	129%	127%	108%	128%	129%	129%	129%
Région O3	100%	100%	100%	85%	100%	100%	100%	100%
Région O4	108%	110%	110%	92%	109%	110%	109%	109%
Région O5	100%	100%	100%	85%	100%	100%	100%	100%
Région O6	108%	110%	110%	85%	116%	110%	109%	110%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	100%	100%	100%	85%	100%	100%	100%	100%
Région O8	100%	100%	100%	85%	100%	100%	100%	100%
Région O9	100%	100%	100%	85%	100%	100%	100%	100%
Région O10	116%	117%	118%	99%	117%	118%	118%	118%
Région O11	108%	110%	110%	92%	109%	110%	109%	109%
Région O12	126%	129%	127%	108%	128%	129%	129%	129%

Exemple 2 - Assureur 2								
Classe âge	0	20	30	40	50	60	70	80
Région O1	97%	96%	95%	87%	95%	95%	95%	95%
Région O2	100%	100%	100%	92%	100%	100%	100%	100%
Région O3	92%	91%	90%	84%	90%	90%	98%	90%
Région O4	100%	100%	100%	92%	100%	100%	100%	100%
Région O5	81%	83%	81%	75%	81%	81%	81%	81%
Région O6	81%	83%	81%	92%	81%	81%	81%	81%
Région D	100%	100%	100%	100%	100%	100%	100%	100%
Région O7	81%	83%	81%	75%	81%	81%	81%	81%
Région O8	81%	83%	81%	75%	81%	81%	81%	81%
Région O9	92%	91%	90%	84%	90%	90%	91%	90%
Région O10	97%	96%	95%	87%	95%	95%	95%	95%
Région O11	92%	91%	90%	84%	90%	90%	91%	90%
Région O12	92%	91%	90%	84%	90%	90%	91%	90%

Tableau 56 : Plusieurs tables de coefficients de déformation pour la garantie B