Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Madame Eugénie OUDIN

Titre du mémoire : L'influence de la sélection médicale en temps de pandémie sur la sinistralité en prévoyance individuelle

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

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

Membres présents du jury de la filière : Signature :

Membres présents du jury de l'Institut des Actuaires : Signature :

Entreprise : SwissLife Santé & Prévoyance

Nom : DROUET D'AUBIGNY Delphine

Signature :

Directeur de mémoire en entreprise

Nom : MONTAGNE Pierre

Signature :

Invité :

Nom : MARJOLLET Pierre

Signature :

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 : MONTAGNE Pierre

Signature du candidat :

Résumé

Mots clés : *Prévoyance Individuelle, Sélection Médicale, Ratio de Morbidité, GLM, Arbres CART, Forêts Aléatoires, GBM, ORSA*

Première étape obligatoire lors de la souscription d'un contrat d'assurance en prévoyance individuelle, la sélection médicale permet à l'assureur d'acquérir une connaissance plus approfondie du futur assuré destiné à être couvert par le contrat.

Toutefois, l'état de santé d'une personne n'étant pas figé dans le temps, l'évaluation établie lors de la souscription est amenée à évoluer, rendant ainsi obsolètes après plusieurs années les informations recueillies relatives à la personne. Il devient alors nécessaire de mettre en place des solutions pour dépasser les limites de la sélection médicale. Dans ce mémoire, une solution sera proposée afin d'atténuer cette limitation dans le temps au travers de la modélisation d'une revalorisation s'appuyant sur la théorie des GLM.

Par ailleurs, face à l'apparition de nouveaux enjeux et notamment de nouvelles maladies l'assureur est obligé de sans cesse s'adapter. Ainsi, en parallèle d'un effet limité dans le temps, la sélection médicale doit elle aussi évoluer afin de demeurer efficace. Une illustration de l'évolution du questionnaire médical dans le cas de la Covid sera ainsi proposée dans ce mémoire, nécessitant le recours à différentes méthodes de Machine Learning telles que les arbres CART, les forêts aléatoires et les algorithmes de Gradient Boosting.

Enfin, en complément de l'évolution du système de sélection médicale, dans le cas où cette solution ne serait pas retenue, les conséquences de ces nouveaux enjeux seront mesurées dans le cadre d'un scénario ORSA.

Abstract

Key words : *Individual Cover, Medical Selection, Morbidity Rate, GLM, Decision Trees, Random Forests, GBM, ORSA*

Medical selection is the first mandatory step in taking out an individual insurance contract. It allows insurer to acquire a more in-depth knowledge of the future insured to be covered by the contract.

However, since a person's state of health is not fixed in time, the assessment made at the time of underwriting is likely to evolve, making the information collected about the person obsolete after several years. Therefore, solutions must be put in place. In this paper, a solution will be proposed to mitigate this limitation effect through the modeling of a revaluation based on the GLM theory.

Furthermore, the insurers must adapt their medical selection methods to new issues like new diseases. Thus, in parallel with a limited effect in time, medical selection must also evolve in order to remain effective. An illustration of the evolution of the medical questionnaire in the case of Covid will be proposed in this paper, requiring the use of different Machine Learning methods such as CART trees, random forests and Gradient Boosting algorithms.

Finally, in addition to the evolution of the medical selection system, the impact of these new issues will be measured in different ORSA scenarios.

Note de synthèse

Mots clés : *Prévoyance Individuelle, Sélection Médicale, Ratio de Morbidité, GLM, Arbres CART, Forêts Aléatoires, GBM, ORSA*

Introduction

Mécanisme obligatoire en prévoyance individuelle, la sélection médicale est le moyen privilégié pour l'assureur de prévenir l'antisélection. Au moyen détourné de nombreuses questions, ce dernier est alors en mesure de dresser un tableau de l'état de santé d'un futur assuré.
Cependant, nul ne peut ignorer que le profil de risque établi lors de la souscription sera amené à évoluer au cours du temps. Une question générale émane alors de ce constat : quelle est l'influence de la sélection médicale sur la sinistralité d'une compagnie d'assurance ? De cette question peuvent aussi découler d'autres discussions relatives à son efficacité et son évolution au cours du temps.

Le temps, un mécanisme limitant l'effet de la sélection médicale

Bien qu'étant obligatoire en prévoyance individuelle, il est intéressant de se pencher sur l'étude de l'impact de la sélection médicale sur la sinistralité. Une manière de procéder parmi tant d'autres est alors de regarder une adaptation du ratio standard de mortalité (SMR) à savoir, le ratio de morbidité.

Construit en s'appuyant sur la théorie sous-jacente du ratio SMR, le ratio de morbidité permet la comparaison de la sinistralité de deux populations différentes : une population de référence et une population d'étude.

Pour mener à bien l'analyse, trois variables méritent alors d'être définies.

$\boldsymbol{S_x}$ **:** le nombre d'arrêts de travail concernant des assurés d'âge $\boldsymbol{x}$ de la population d'étude

$\boldsymbol{i_x}$ **:** la probabilité d'incidence en arrêt de travail pour la population de référence selon l'âge $\boldsymbol{x}$

$\boldsymbol{E_x}$ **:** le nombre d'assurés d'âge $\boldsymbol{x}$ de la population d'étude exposés à un arrêt de travail

Grâce aux trois variables exposées ci-dessus, deux nouvelles quantités font alors leur apparition.

- $\sum_x \boldsymbol{S_x}$: le nombre d'arrêts de travail observés dans la population d'étude
- $\sum_x \boldsymbol{i_x} \times \boldsymbol{E_x}$: le nombre d'arrêts de travail attendus pour les assurés d'âge $\boldsymbol{x}$ si l'incidence était identique à celle de la population de référence

Rapportées l'une à l'autre, ces deux quantités permettent d'obtenir la statistique de test suivante :

$$t = \frac{\sum_x S_x}{\sum_x i_x \times E_x}$$

L'étude appliquée dans le cas de la population TNS (Travailleurs Non Salariés), individus constituant le portefeuille d'étude SwissLife et soumis à la sélection médicale, comparés aux individus TNS de la population française non soumis au processus de sélection, a permis d'aboutir aux conclusions suivantes.

La sélection médicale a bien un effet sur la sinistralité. En effet, les valeurs inférieures à 1 de la statistique de test t, visibles sur le graphique ci-dessous, au cours des premières années d'ancienneté dans le risque indiquent que la sélection médicale permet de se ramener à une population avec une sinistralité plus faible qu'une population générale non soumise à l'étape de sélection.

Toutefois, au bout de deux années d'ancienneté, la sinistralité de la population soumise à la sélection médicale vient à doubler par rapport à celle non soumise au processus de sélection, laissant ainsi penser que l'effet de la sélection médicale s'estompe au cours du temps. Afin de prévenir d'éventuelles dérives en termes de sinistralité, l'assureur est alors contraint de mettre en place des solutions dans le but de contrebalancer l'effet de la sélection médicale diminuant au cours du temps.

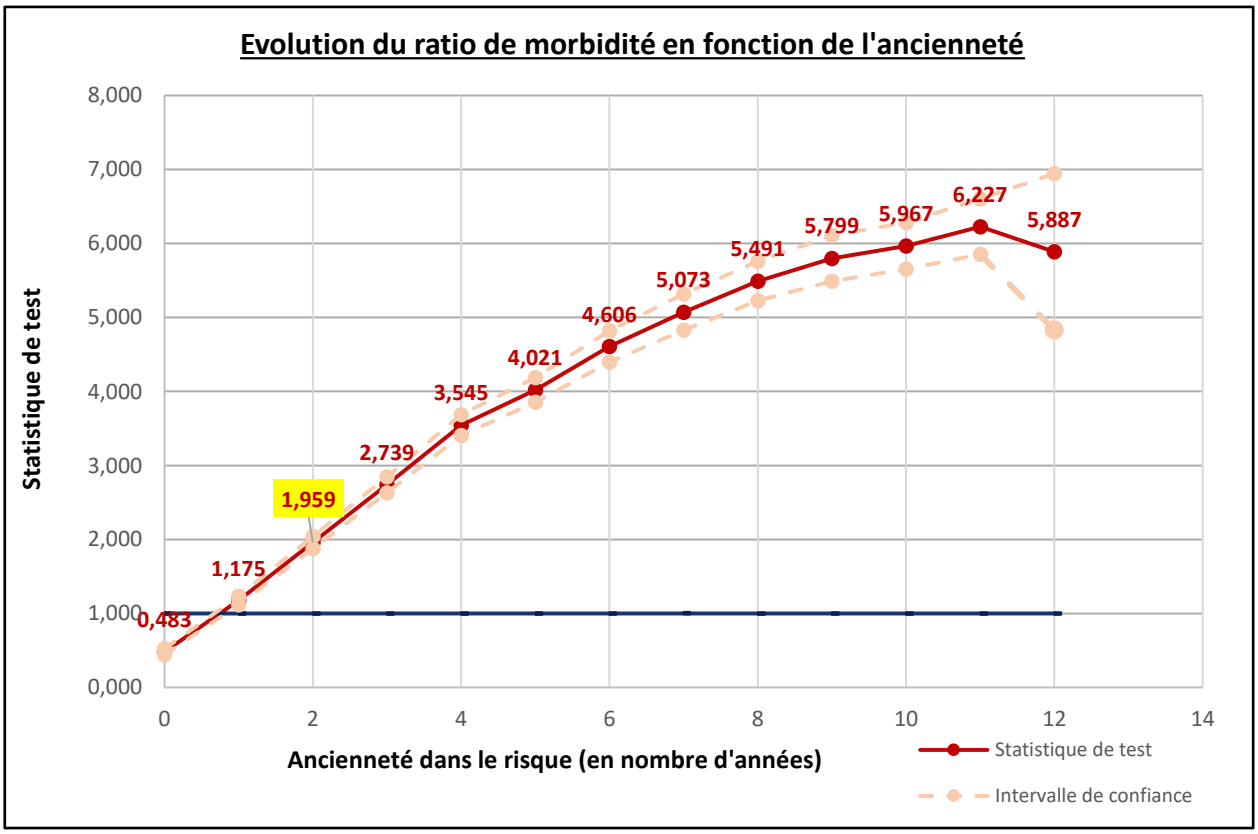


Figure 1 - [Note de Synthèse] - Effet de la sélection médicale au cours du temps

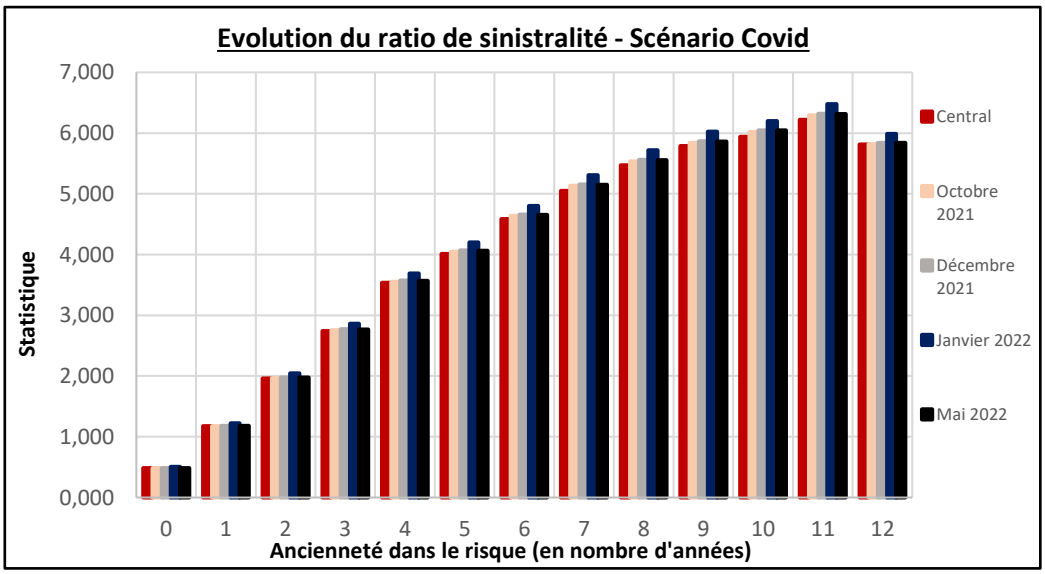
Une solution proposée ici aux dérives de sinistralité éventuelles est la mise en place d'une revalorisation. En effet, la majoration tarifaire compte parmi les leviers à disposition de l'assureur. Cependant, cette dernière ne peut être appliquée arbitrairement. Différentes variables apparaissent comme indispensables dans la détermination de la revalorisation.

Parmi elles, et comme démontré auparavant, l'**ancienneté** dans le risque. Par ailleurs, la **catégorie socio professionnelle** ne peut être mise de côté. En effet, le même ratio de morbidité ventilé selon les professions conduit à la conclusion que toutes les activités professionnelles ne sont pas équitablement sujettes aux questions de morbidité. La catégorie professionnelle relative aux professions paramédicales telles que les infirmières en est un parfait exemple. Enfin, la **sinistralité**, au travers du ratio S/P permet de compléter la liste non exhaustive des variables explicatives nécessaires à la mise en place d'une revalorisation.

Grâce à celles-ci et à l'appui de la théorie des GLM Gaussien et Tweedie notamment, une revalorisation peut être mise en place, offrant la possibilité à l'assureur de soumettre chaque assuré du portefeuille à une seconde sélection médicale déguisée.

Le temps, un mécanisme conduisant à l'évolution de la sélection médicale

Ces dernières années, le monde de l'assurance a vu émerger de nouveaux enjeux. Parmi eux, l'apparition de nouvelles maladies telles que la Covid-19, venue accroître considérablement le nombre d'arrêts de travail courts.
Face à ces derniers enjeux, et bien qu'ayant déjà une incidence matérielle sur la sinistralité, la sélection médicale ne peut se permettre d'être figée dans le temps. L'application du ratio de morbidité dans le cas de la propagation de la Covid selon différentes périodes, plus ou moins soumises au virus, en est un exemple comme l'illustre le graphique ci-dessous.



	Incidence			
Age	06-12/10/21	01-07/12/21	19-25/01/22	04-10/05/22
18-39	0,06%	0,60%	4,82%	0,41%
40-49	0,05%	0,60%	4,43%	0,41%
50-60	0,04%	0,42%	2,47%	0,43%
>60	0,03%	0,23%	1,24%	0,40%

Figure 2 - [Note de synthèse] - Ratio de sinistralité - Prise en compte de l'incidence Covid

En effet, comme il est possible de le lire sur le graphique ci-dessus, des périodes de forte incidence de la Covid, comme « Janvier 2022 », conduisent à une augmentation non négligeable de la sinistralité, modélisée ici au travers de la statistique de test.

Face à ces nouveaux défis, les mécanismes de sélection médicale se doivent d'être repensés.
Par conséquent, il devient pertinent de se questionner sur la refonte du questionnaire de sélection médicale, en intégrant notamment des informations relatives à ces nouveaux enjeux.

Pour y parvenir, un travail préalable de reconstruction de la méthode d'acceptation du risque pour le produit étudié est donc primordial. Ce travail, permettant la modélisation de la variable **risque**, représentant le fait d'accepter ou non un individu à l'issue du questionnaire de sélection médicale, a été mené à l'aide des variables explicatives que voici :

- **L'âge**
- **Le poids**
- **La taille**
- **Le rapport poids taille (RPT)** défini par $RPT = 1 + \frac{Poids}{Taille^2} \times 100$
- **Fumeur**
- **L'asthme**
- **La thyroïde**
- **La thyroïde cancéreuse**
- **L'hypertension artérielle**
- **Le cholestérol**

Pour information, les cinq dernières variables listées permettent de représenter, à l'aide d'un « Oui » ou d'un « Non », si l'individu est concerné par l'une de ces pathologies.

Ainsi, grâce à une étape de reconstruction de ces variables et aux techniques de Machine Learning telles que les arbres CART, les forêts aléatoires ou encore les algorithmes de Gradient Boosting, l'acceptation d'un individu dans le portefeuille a pu être modélisée.

Dans ce contexte, les résultats obtenus pour des algorithmes lancés sur une population de 318 060 individus sont les suivants :

Oui	Refus	Study	TOTAL
221 371	20 093	76 596	**318 060**

Figure 3 - [Note de synthèse] - Répartition du risque avant l'intégration de l'information Covid

A l'issue de ces premiers algorithmes ne prenant pas en compte l'information Covid, 70% des individus sont directement acceptés dans le risque pour 6% refusés. Par ailleurs, 24% d'entre eux sont classés dans la catégorie « Study », indiquant qu'un examen complémentaire devra être effectué.

Dans un second temps, dans l'optique de faire évoluer le questionnaire de sélection médicale et de le rendre toujours plus compétitif, le même travail a pu être appliqué à une base de données complétée de trois variables relatives à la Covid.

La première, portant le nom de **covid** permet de renseigner si une personne a eu ou non la Covid. La seconde, sous le nom de **hospi**, permet quant à elle de savoir si l'assuré a été hospitalisé ou non en réanimation alors qu'il était atteint de la Covid. Enfin, la dernière permet de savoir si l'individu étudié présente encore des symptômes un mois après son infection. Celle-ci portera le nom de **covid_long**.

En s'appuyant une fois de plus sur les méthodes de Machine Learning, un nouveau modèle de prédiction conduisant aux résultats suivants est obtenu.

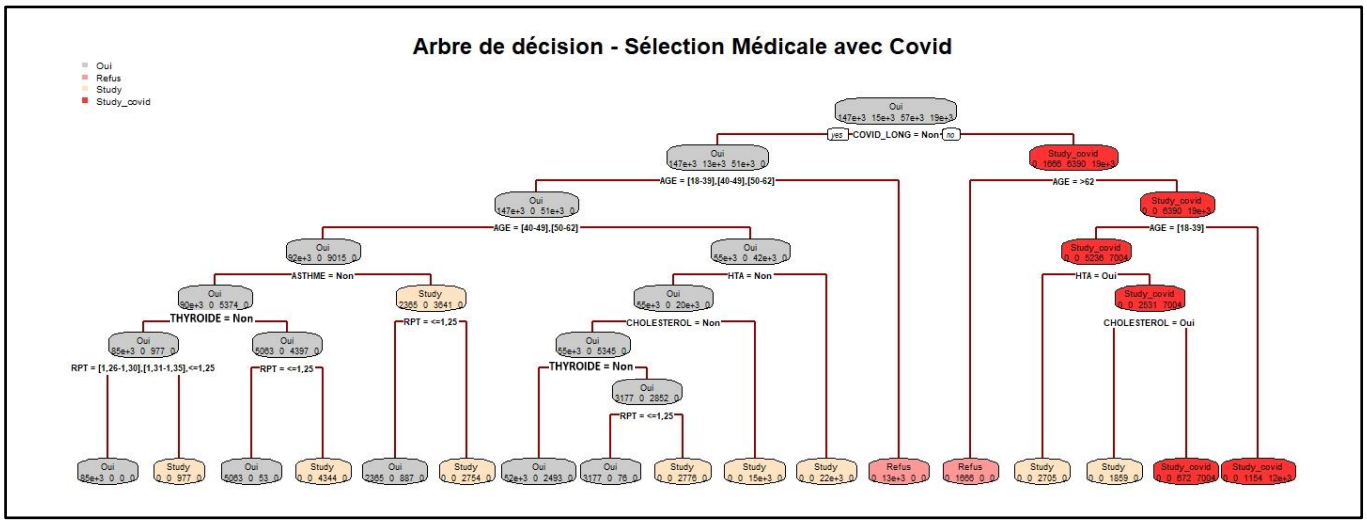


Figure 4 - [Note de synthèse] - Arbre optimal de décision - Sélection Médicale avec Covid

Grâce à sa représentation visuelle explicite, il est possible de remarquer que l'arbre de décision ainsi obtenu se retrouve alors modifié avec la création d'une nouvelle classe pour la variable **risque.** Cette dernière possède désormais quatre classes avec l'ajout de la classe « Study_covid », permettant de modéliser le passage d'individus initialement acceptés dans le risque à une classe d'étude complémentaire propre à la Covid. Pas moins de 8% des individus sont alors concernés et nécessitent désormais une étude complémentaire.

Oui	Refus	Study	Study_covid	TOTAL
196 221	20 093	76 596	25 150	**318 060**

Figure 5 - [Note de synthèse] - Répartition du risque après l'intégration de l'information Covid

Cet exercice appliqué dans le cas de la Covid, mais qui pourrait se révéler pertinent dans le cas d'autres maladies, permet de mettre en lumière l'impact d'une telle maladie sur des pathologies initialement connues et classiques. La mise sous silence de ces nouvelles pathologies dans le questionnaire médical conduirait alors l'assureur à accroître son niveau d'antisélection, niveau qui est pourtant destiné à être réduit par la sélection médicale.

Certaines pathologies ayant de fortes corrélations avec la Covid, telles que le cholestérol ou encore l'hypertension artérielle, seraient alors négligées, pouvant conduire l'assureur à des dérives de sinistralité non anticipées.

Toutefois, la modification de la méthode de sélection médicale n'est pas l'unique façon d'anticiper des dérives de sinistralité. La mise en place de scénarios au travers de l'ORSA est également une solution pouvant être choisie. Grâce à sa faculté d'analyse prospective des risques, l'ORSA permet, de par son scénario pandémie, de mettre en lumière qu'un tel évènement comme la Covid et ses confinements successifs n'est pas sans conséquences pour l'assureur au niveau de son chiffre d'affaires notamment.

Conclusion

Mise en place lors de la souscription, la sélection médicale permet à l'assureur d'établir un premier profil de risque de l'individu souhaitant souscrire un contrat en prévoyance individuelle. Cependant, le diagnostic ainsi établi est voué à se dégrader au cours du temps à mesure que l'effet de la sélection médicale s'estompe en fonction de l'ancienneté. L'assureur est alors contraint de réagir au moyen de différents leviers, comme la mise en place d'une revalorisation pouvant être modélisée à l'appui de la théorie des GLM par exemple.

En plus de la limitation de son effet dans le temps, la sélection médicale doit également faire face à son obsolescence au vu de l'apparition de nouvelles maladies. Pour y remédier, le questionnaire de sélection médicale, modélisé ici à l'aide des méthodes de Machine Learning, doit sans cesse évoluer pour ne pas laisser des dérives de sinistralité s'installer.

Synthesis Note

Key words : *Individual Cover, Medical Selection, Morbidity Rate, GLM, Decision Trees, Random Forests, GBM, ORSA*

Introduction

Medical selection is a mandatory mechanism in individual insurance allowing the reduction of anti-selection. By asking several questions, the insurer is able to draw up a summary of the health of a future insured.
However, no one can ignore the fact that the risk profile established at the time of underwriting will change over time. Therefore, a general question arises from this observation: what is the impact of medical selection on an insurance company's claims experience? And from this, other discussions can also arise regarding its effectiveness and its evolution over time.

Time, a mechanism killing the effect of medical selection

Although mandatory in individual insurance, it is interesting to study the impact of medical selection on claims experience. A way to do it is to look at an adaptation of the standard mortality ratio (SMR), namely the morbidity ratio.

Built on the underlying theory of the SMR, the morbidity ratio allows the comparison of claims experience between two different populations: a reference population and a study population.

To carry out the analysis, three variables need to be defined.

$\boldsymbol{S_x}$ **:** the number of sick leaves involving insured persons of age $\boldsymbol{x}$ in the study population

$\boldsymbol{i_x}$ **:** the probability of sick leaves for the reference population according to age $\boldsymbol{x}$

$\boldsymbol{E_x}$ **:** the number of insured persons of age $\boldsymbol{x}$ in the study population exposed to a sick leave

By means of the three variables exposed above, two new quantities then appear.

- $\sum_x \boldsymbol{S_x}$: the number of sick leaves observed in the study population
- $\sum_x \boldsymbol{i_x} \times \boldsymbol{E_x}$: the number of sick leaves expected for insured persons of age $\boldsymbol{x}$ if the incidence were identical to that of the reference population

When related to each other, these two quantities give the following test statistic:

$$t = \frac{\sum_x S_x}{\sum_x i_x \times E_x}$$

The study applied to two self-employed populations: one subject to medical selection and the other not, led to the following conclusions. Medical selection does have an effect on claims experience. Indeed, the values below 1 of the t-test statistic during the first years of risk exposure indicate that medical selection leads to a population with a lower loss ratio than a general population not subject to the selection step.

However, after two years of experience, the claims experience of the medically screened population doubles that of the unscreened population, suggesting that the effect of medical selection fades over time. In order not to face drift in terms of claims experience, the insurer is then forced to put in place means to counteract the diminishing effect of medical selection over time.

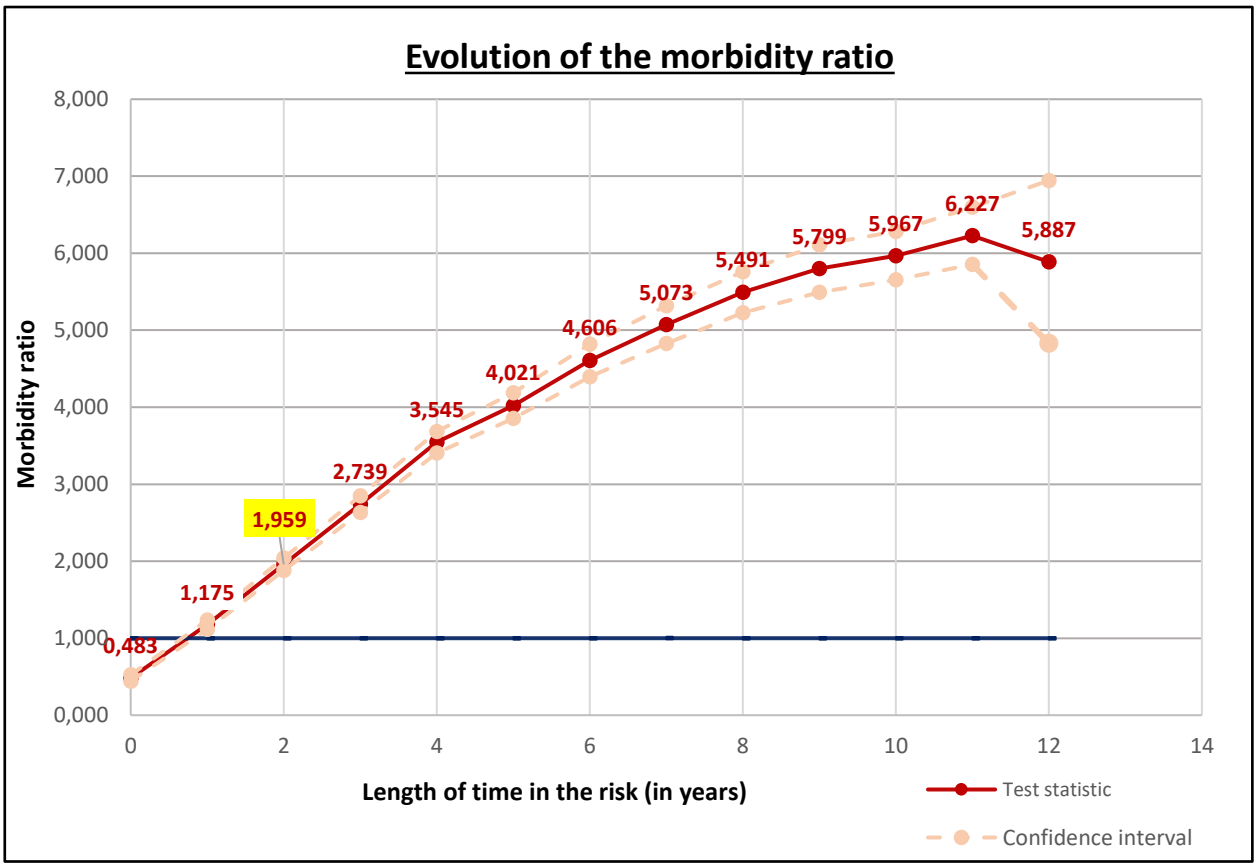


Figure 6 - [Synthesis Note] - Effect of medical selection over time

One solution proposed here to possible claims excesses is the implementation of a revaluation. Indeed, the tariff increase is one of the levers available for the insurer. However, it cannot be applied arbitrarily. Various variables appear to be essential in determining the increase.

Among them, and as previously demonstrated, **seniority** in the risk. In addition, the **socio-occupational category** cannot be ignored. Indeed, the same morbidity ratio broken down by socio-occupational category leads to the conclusion that not all professional activities are equally subject to morbidity issues. The occupational category for paramedical professions such as nurses is a perfect example. Finally, **claims experience**, through the loss ratio, completes

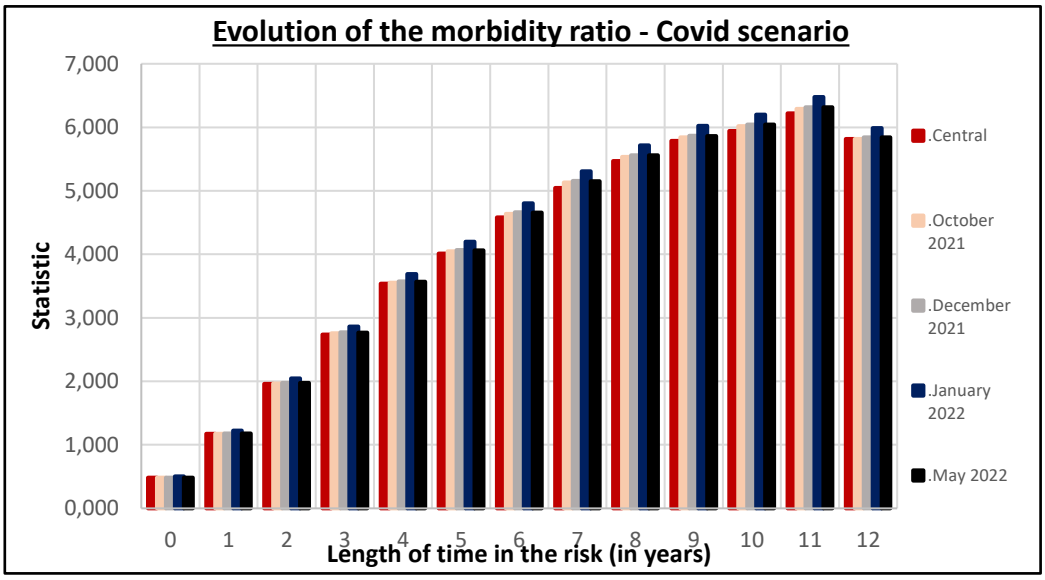
the non-exhaustive list of explanatory variables necessary for the implementation of a revaluation.

Thanks to these variables and to the support of the Gaussian and Tweedie GLM theory, a revaluation can be set up, offering the insurer the possibility of subjecting each insured in the portfolio to a second disguised medical selection.

Time, a mechanism leading to the evolution of medical selection

In recent years, new challenges have emerged in the insurance world. Among them, the appearance of new diseases such as Covid-19, which have considerably increased the number of work stoppages.

With these, and although they have a real impact on the claims experience, medical selection cannot be fixed in time. The application of the morbidity ratio in the case of the Covid spread according to different periods, more or less subject to the virus, is an example of this, as illustrated in the graph below.



	Incidence			
Age	06-12/10/21	01-07/12/21	19-25/01/22	04-10/05/22
18-39	0,06%	0,60%	4,82%	0,41%
40-49	0,05%	0,60%	4,43%	0,41%
50-60	0,04%	0,42%	2,47%	0,43%
>60	0,03%	0,23%	1,24%	0,40%

Figure 7 - [Synthesis Note] - Loss Ratio - Considering the Covid Incidence

Indeed, as can be seen from the graph above, periods of high Covid incidence, such as "January 2022", lead to a non-negligible increase in the claims experience, modelled here through the test statistic.

Faced with these new challenges, medical selection mechanisms need to be rethought.

Consequently, it becomes relevant to consider the redesign of the medical selection questionnaire, in particular by integrating information relating to these new issues.

To achieve this, it is essential to first reconstruct the risk acceptance method for the product under study.

This work, allowing the modelling of the risk variable, representing the fact of accepting or not an individual at the end of the medical selection questionnaire, was carried out using the following explanatory variables.

- **Age**
- **Weight**
- **Height**
- **The ratio weight to height (RWH)** defined as $1 + \frac{Weight}{Height^2} \times 100$
- **Smoker**
- **Asthma**
- **The thyroid**
- **Cancerous thyroid**
- **High blood pressure (HBP)**
- **Cholesterol**

For information, the last five variables listed make it possible to represent, with the help of a "Yes" or a "No", whether the individual is concerned by one of these pathologies.

Thus, thanks to a step of reconstruction of these variables and to Machine Learning techniques such as CART trees, random forests or Gradient Boosting algorithms, the acceptance of an individual in the portfolio could be modelled.

In this context, the results obtained for algorithms run on a population of 318 060 individuals are as follows:

Accepted	Refused	Study	TOTAL
221 371	20 093	76 596	**318 060**

Figure 8 - [Synthesis Note] - Distribution of risk before including Covid information

As a result of these first algorithms, which do not take into account the Covid information, 70% of them are directly accepted in the risk and 6% are refused. Furthermore, 24% of them are classified in the "Study" category indicating that a complementary examination should be performed.

In a second phase, with a view to developing the medical selection questionnaire and making it ever more competitive, the same work was applied to a database supplemented with three variables relating to Covid.

The first, called **covid**, indicates whether or not a person has had Covid. The second, called **hospi**, indicates whether or not the insured person was hospitalised in intensive care while suffering from Covid. Finally, the last one makes it possible to know if the individual studied still presents symptoms one month after his infection. This one will be called **covid_long**.

Once again using Machine Learning methods, a new prediction model leading to the following results is obtained.

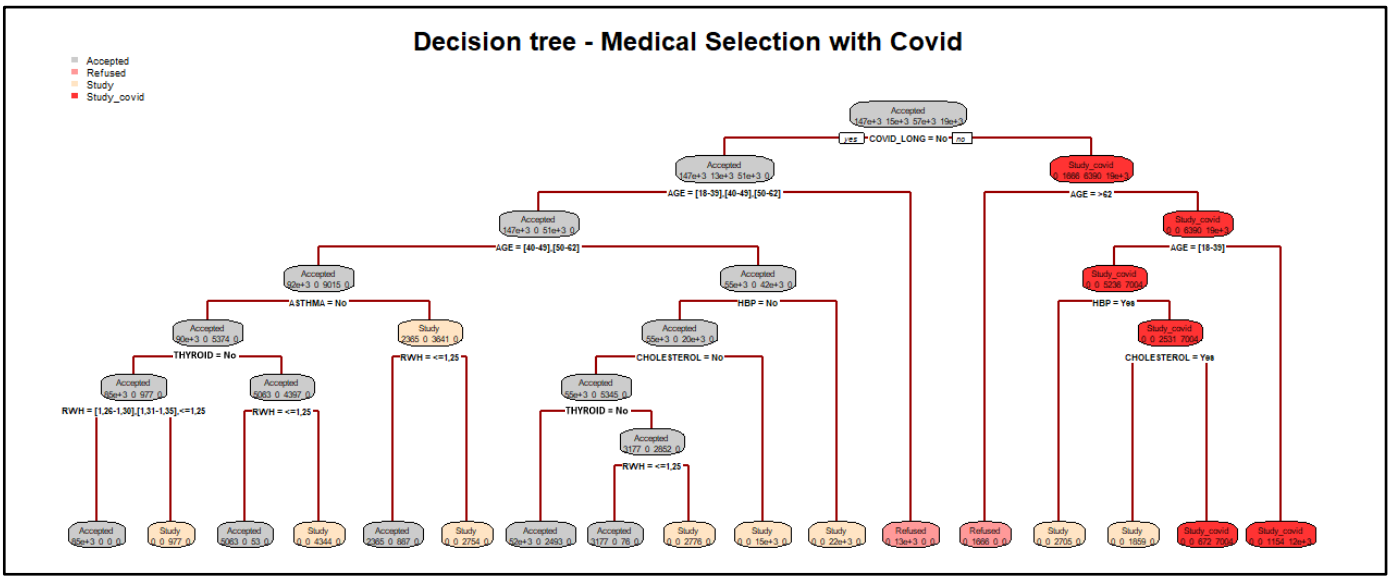


Figure 9 - [Synthesis Note] - Optimal Decision Tree - Medical Selection including Covid

Thanks to its explicit visual representation, it is possible to notice that the decision tree obtained is then modified with the creation of a new class for the variable risk. This one has now four classes with the addition of the "Study_covid" class, which models the passage of individuals initially accepted in the risk to a complementary study class specific to the Covid. No less than 8% of individuals are now affected and require further study.

Accepted	Refused	Study	Study_covid	TOTAL
196 221	20 093	76 596	25 150	**318 060**

Figure 10 - [Synthesis Note] - Distribution of risk after including Covid information

This exercise applied in the case of Covid, but which could be true in the case of other diseases, highlights the impact of such a disease on initially known and classic pathologies. If these new diseases were not included in the medical questionnaire, the insurer would increase its level of anti-selection, a level that is intended to be reduced by medical selection.
Some diseases with strong correlations with Covid, such as cholesterol or high blood pressure, would then be neglected, which could lead the insurer to unanticipated claims drift.

However, changing the medical selection method is not the only way to proceed before anticipating claims drifts. The implementation of scenarios through ORSA is also a solution that can be chosen. Thanks to its ability to analyse risks prospectively, ORSA makes it possible, through its pandemic scenario, to highlight that such an event as Covid and its successive confinements are not without consequences for the insurer, particularly in terms of its revenue.

Conclusion

The medical selection process, which is implemented at the time of underwriting, enables the insurer to establish an initial risk profile of the individual wishing to take out an individual insurance policy.
However, the diagnosis established will deteriorate over time, as the effect of medical selection fades with age. The insurer is then obliged to react by means of various levers, such as the implementation of a revaluation that can be modelled using GLM theory.

In addition to this time-limited effect, medical selection also has to cope with obsolescence in view of the emergence of new diseases. To remedy this, the medical selection questionnaire, modelled using Machine Learning methods, must constantly evolve in order not to let claims drift.

Remerciements

En préambule de ce mémoire, je tiens à remercier **Maud Thomas**, maître de conférences à Sorbonne Université pour son encadrement pédagogique dans la réalisation de ce mémoire.

Je remercie également **Pierre Marjollet** et **Pierre Montagne**, membres de l'équipe Inventaire chez SwissLife, pour leur accompagnement et précieux conseils tout au long de ces travaux.

Je remercie aussi **Lauriane Chatelet**, membre de l'équipe Technique Produits chez SwissLife pour sa transmission de connaissances au sujet du produit SLPI étudié dans ce mémoire.

Enfin, j'étends également mes remerciements à toute l'équipe ainsi qu'à toutes les personnes avec qui j'ai pu échanger lors de cette année en apprentissage.

Table des matières

Index - Abréviations utilisées

AGGIR : *Autonomie Gérontologique Groupes Iso-Ressources*

AGIS : *Association Générale Interprofessionnelle de Solidarité*

BGS : *Besoin Global de Solvabilité*

CART : *Classification And Regression Trees*

CNAMTS : *Caisse Nationale d'Assurance Maladie des Travailleurs Salariés*

CNAVPL : *Caisse Nationale d'Assurance Vieillesse des Professions Libérales*

CPAM : *Caisse Régionale d'Assurance Maladie*

CRAM : *Caisse Régionale d'Assurance Maladie*

CSP : *Classe Socio Professionnelle*

GBM : *Gradient Boosting Machine*

GIR : *Groupes Iso-Ressources*

GLM : *Generalized Linear Models*

IJ : *Indemnité Journalière*

INED : *Institut National d'Etudes Démographiques*

INSEE : *Institut National de la Statistique et des Etudes Economiques*

LOO : *Leave-One-Out*

MSA : *Mutualité Sociale Agricole*

OMS : *Organisation mondiale de la Santé*

OOB : *Out-Of-Bag*

ORSA : *Own Risk and Solvency Assessment*

PTIA : *Perte Totale et Irréversible d'Autonomie*

RPT : *Ratio Poids Taille*

RSI : *Régime Social des Indépendants*

S/P : *Ratio Sinistres sur Primes*

SCR : *Solvency Capital Requirement*

SMR : *Standard Mortality Ratio*

SSI : *Sécurité Sociale Indépendants*

TNS : *Travailleurs Non Salariés*

Introduction

« La prévoyance des maux est le grand art de les affaiblir avant qu'ils n'arrivent ».

Bien que plus de deux siècles et demi d'histoire se soient écoulés depuis cette phrase prononcée par Voltaire, rien ne semble avoir altéré la justesse de la définition du philosophe des Lumières.

Dans un monde où tout doit être prévu, connu par avance, mais dont l'équilibre est de plus en plus menacé par des pandémies, le vieillissement de la population ou encore des départs à la retraite retardés, seul l'ingénu ou le fort optimiste refusera aujourd'hui de se couvrir lui et ses proches contre les aléas de la vie.

Ainsi, bien que la prévoyance individuelle ne date pas d'hier, elle apparaît comme étant toujours synonyme de nouveaux défis pour les assureurs, notamment pour les statuts particuliers comme ceux des indépendants, dont il est bien connu qu'une grande partie n'est encore pas couverte par un contrat de prévoyance.

Toutefois, nouveaux enjeux et nouveaux horizons riment aussi avec de potentiels risques inconnus et impacts sur la solvabilité des assureurs. Par conséquent, dans le but d'éviter de voir affluer un grand nombre de candidats aux profils risqués et peu connus, un écrémage de ces derniers est donc indispensable. Une sélection médicale pour les contrats de type individuel sera donc de mise.

Comme tous les assureurs du marché, SwissLife ne peut se soustraire à cette phase de sélection médicale lors de la souscription en prévoyance individuelle. Faire souscrire un contrat pour protéger contre des risques lourds tels qu'un décès ou un arrêt de travail en omettant de mettre en place une sélection médicale serait fort périlleux pour l'entreprise.

Ainsi, une fois les nombreuses incertitudes recontextualisées, il est intéressant de s'interroger sur l'impact d'une telle sélection médicale sur la sinistralité d'un portefeuille en prévoyance individuelle. Cette dernière a-t-elle une réelle influence sur la sinistralité ?

De cette question générale peuvent découler de nombreuses sous-questions :

Ses effets sont-ils réellement perceptibles ? A-t-elle une efficacité limitée dans le temps, et si oui comment y remédier ? Est-elle figée dans le temps ou au contraire peut-elle évoluer en fonction de nouvelles pathologies pour demeurer efficace ?

Autant de questions qui devront être élucidées au cours de ce mémoire. Pour illustrer les méthodes mises en place afin d'y répondre, un portefeuille propre à un produit SwissLife a été choisi : à savoir un produit de prévoyance destiné aux indépendants.

L'étude s'articulera en cinq parties. La première partie, et la plus générale de toutes, aura pour objectif de rappeler un certain nombre de points au sujet de la prévoyance individuelle en France. Dans une seconde partie, une digression sera apportée afin de présenter dans de plus amples détails le produit SwissLife sélectionné. La troisième partie sera entièrement dédiée à la présentation des éléments théoriques utilisés lors des différentes modélisations.
Les deux parties suivantes seront quant à elles des parties de mise en application de différentes méthodes.

De manière plus spécifique, la quatrième partie aura pour but de démontrer l'effet de la sélection médicale au travers d'un ratio de sinistralité, mais aussi ses limites dans le temps. Par conséquent, une solution à ces limites sera apportée au travers de la mise en place d'une revalorisation à l'appui des modèles linéaires généralisés.

La dernière partie sera orientée autour de la réflexion suivante : l'apparition de nouvelles maladies telles que la Covid rime-t-elle avec refonte de la sélection médicale ?
Afin de développer ce point, un premier travail de modélisation de la sélection médicale chez SwissLife devra être entrepris grâce à des techniques de classification en Machine Learning telles que les arbres de décision ou encore les forêts aléatoires. Une fois cette étape réalisée, l'impact de la prise en compte de la pathologie Covid dans le questionnaire de sélection médicale pourra être étudié.

Enfin, en supposant que la pathologie Covid ne soit plus prise en compte dans le questionnaire de sélection médicale, donc d'un point de vue assuré, quelles seraient les conséquences pour l'entreprise ? Cette réflexion côté assureur sera menée au travers d'un scénario ORSA.

I. Qu'est-ce que la prévoyance individuelle ?

Avec la création de la Sécurité Sociale en 1945 post Seconde Guerre Mondiale, un certain nombre d'organismes se développent dans le domaine de la protection sociale. Leur but est simple : venir suppléer et compléter les remboursements effectués par la Sécurité Sociale.

De manière générale, la prévoyance vise à prévenir l'apparition de risques, et dans le cas où ces derniers apparaitraient, y faire face en versant des prestations.

Le versement de ces prestations en France est divisé en plusieurs niveaux. Dans un premier temps, il s'agit des régimes de base obligatoires. Pour illustrer ces derniers, peut être cité le régime de la Sécurité Sociale, au sein duquel plusieurs sous-régimes cohabitent. Le plus connu d'entre eux est le Régime Général. Mais d'autres régimes existent également, comme le régime agricole qui est propre aux salariés et exploitants agricoles. Il existe également le régime des non-salariés non agricoles pour les commerçants, artisans et professions libérales, ainsi que des régimes particuliers et des régimes spéciaux (agents de la SNCF, agents EDF, fonctionnaires ...).

En complément de ces régimes de base obligatoires, viennent les régimes complémentaires obligatoires. À eux deux, ils constituent l'ensemble des régimes obligatoires.
Parmi ces régimes complémentaires obligatoires, sont à retrouver notamment la plupart des régimes d'entreprise à caractère collectif et à adhésion obligatoire. De plus, pour les travailleurs non-salariés (TNS), d'autres régimes particuliers font foi, il s'agit des régimes professionnels des TNS.

Enfin, les régimes supplémentaires et non obligatoires complètent ce triptyque des différents niveaux de prévoyance en France. Ce sont au sein de ces derniers que des particuliers peuvent souscrire des contrats auprès de sociétés d'assurance, d'instituts de prévoyance ou encore de mutuelles.

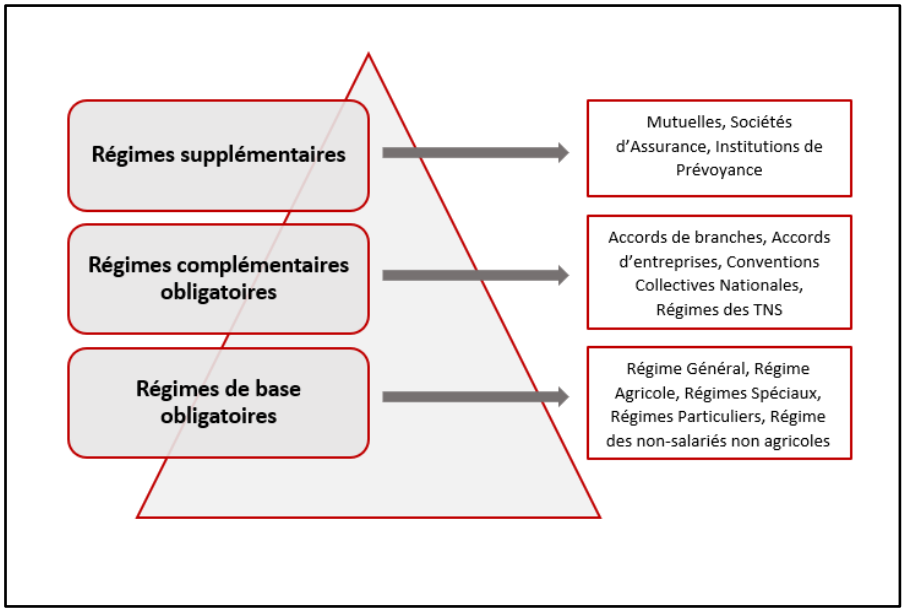


Figure 11 - Pyramide des régimes en France

Après avoir fait un focus sur l'architecture de la prévoyance en France, il est intéressant de se pencher sur l'objectif de celle-ci afin de savoir ce qu'elle couvre.

Avant toute chose, il est important de savoir que la prévoyance vise à couvrir uniquement les risques liés à une personne. Sont donc omis des couvertures sur les risques liés aux biens.

La prévoyance a donc pour objectif la prise en charge des conséquences liées au décès, à des arrêts de travail, à savoir accidents ou maladies, et aux répercussions de la vieillesse au travers de la dépendance.

A présent, un détail plus précis de chacune de ces trois garanties va être apporté afin d'en révéler les sous-garanties et spécificités.

1 Les grands principes généraux de la prévoyance en France

1.1 Les garanties liées au décès

Parmi les garanties liées au décès, deux sous-catégories de garanties sont à distinguer. D'une part, les capitaux décès, et d'autre part, les rentes.

1.1.1 Capitaux décès

Le capital décès sera versé dès lors que le décès de l'assuré survient. Toutefois, celui-ci pourra être versé de manière anticipée si l'assuré se retrouve dans une situation de perte totale et irréversible d'autonomie, souvent appelée PTIA.

Le montant du capital décès versé résultera en général de la rémunération annuelle brute de l'assuré. Le plus souvent, il s'agira d'une fraction exprimée en pourcentage de la rémunération. Le montant pourra également être amené à fluctuer en fonction de la situation familiale de l'assuré : marié, célibataire, veuf ou divorcé.

Par ailleurs, des majorations du capital initialement déterminé pourront être apportées en cas de décès accidentel par exemple, ou encore de décès lié à un accident de la circulation. Dans certains cas, un doublement ou même un triplement des capitaux pourra être appliqué.

De plus, certaines garanties annexes ou options peuvent venir compléter la garantie standard liée au décès. Parmi ces dernières, la garantie frais d'obsèques est un exemple. Celle-ci, déclenchée en cas de décès de l'assuré ou tout simplement d'un membre de sa famille (enfants ou conjoint par exemple), prévoit le versement d'un montant supplémentaire dans le but de prendre en charge les frais liés aux obsèques.

Enfin, une garantie dite de « double-effet » existe. Grâce à celle-ci, en cas de décès simultané de l'assuré et de son conjoint, ou bien du décès du conjoint dans un intervalle de temps défini à la suite du décès de l'assuré, une somme forfaitaire peut être versée aux enfants à charge.

1.1.2 Les rentes

Le versement d'un capital n'est pas l'unique moyen d'être indemnisé en cas de décès. Les rentes font notamment partie des autres moyens existants. Dans le cadre du décès, deux types de rentes sont à distinguer : les rentes de conjoint et les rentes éducation.

- Les rentes de conjoint

Versées en cas de décès de l'assuré, ces dernières visent à contrebalancer la perte d'un revenu au sein d'un foyer par exemple.

Elles peuvent se matérialiser sous deux formes :

- Un versement de manière viagère
- Un versement temporaire jusqu'à 62 ans par exemple, correspondant à l'âge probable de départ à la retraite sauf départ éventuellement anticipé ou retardé

Généralement, ces deux rentes sont calculées en fonction du dernier salaire de l'assuré qui décède.

Par ailleurs, dans le cas où le conjoint de l'assuré décédé décèderait aussi, certaines garanties proposent de continuer de verser la rente aux enfants alors orphelins. Le plus souvent, il ne s'agit pas du montant de la rente totale mais plutôt d'une fraction de cette dernière, 50% par exemple.

- Les rentes éducation

L'autre forme de rente généralement présente dans les contrats de garanties liées au décès, est la rente éducation. Comme son nom l'indique, cette rente a pour objectif de permettre aux enfants de l'assuré décédé de poursuivre leurs études.

De manière générale, cette rente est versée jusqu'à 18 ans et peut être prolongée jusqu'à 26 ans, dans le cas où les enfants à charge justifient d'une poursuite d'études.

Par ailleurs, elle peut être croissante en fonction de l'âge des enfants et également fractionnée par paliers. Par exemple, le bénéficiaire se verra recevoir 15% du dernier salaire de l'assuré jusqu'à 18 ans puis 25% en cas de poursuite d'études au-delà de cet âge.

A noter, que ces deux types de rentes peuvent être déclenchés dans le cas d'une perte totale et irréversible d'autonomie de la part de l'assuré. Le décès seul n'est pas l'unique facteur de déclenchement.

1.2 Les garanties liées aux arrêts de travail

Lorsque sont évoquées les garanties liées aux arrêts de travail, deux catégories sont à distinguer : les arrêts de travail pour incapacité et les arrêts de travail pour invalidité.

1.2.1 Les arrêts de travail en cas d'incapacité

L'incapacité peut être vue comme l'inaptitude temporaire pour un assuré d'effectuer ses missions dans le cadre de son travail. Cette inaptitude peut être à la fois physique et psychologique. Par ailleurs, l'incapacité peut être complète ou partielle. Les deux cas peuvent donner lieu à une indemnisation.

Ainsi, la garantie incapacité en prévoyance va permettre de compenser la perte de salaire potentielle de l'assuré. Des indemnités journalières (IJ) seront versées aux assurés en complément de celles versées par la Sécurité Sociale.

En général, plusieurs facteurs conditionnent le montant de la prestation versée. Le premier paramètre à considérer est la franchise. Elle peut être de différents types (ferme ou continue, rétroactive ou discontinue) et d'une durée variable. Dans un second temps, le montant des indemnités journalières est conditionné au montant du salaire. Il s'agit généralement d'un pourcentage de celui-ci. Enfin, la durée d'indemnisation est également un paramètre à prendre en compte. De manière générale, la durée maximale de versement des prestations est de 36 mois, autrement dit trois ans de prestations.

1.2.2 Les arrêts de travail en cas d'invalidité

Contrairement à la définition de l'incapacité, l'invalidité peut être définie comme l'inaptitude quasi irréversible pour un assuré d'effectuer ses activités professionnelles. Pour être déclaré invalide, un assuré devra justifier d'une réduction d'au moins deux tiers de sa capacité de travail ou de gain. Là encore, l'invalidité peut être soit psychique soit physique, ou encore, les deux à la fois.

De plus, trois catégories d'invalidité sont à distinguer :

- **L'invalidité de 1ère catégorie :** tout en étant invalide, être capable d'exercer une activité rémunérée.
- **L'invalidité de 2ème catégorie :** être invalide et incapable d'exercer une activité rémunérée.
- **L'invalidité de 3ème catégorie :** être invalide, incapable d'exercer une activité rémunérée et avoir absolument besoin de l'aide d'une tierce personne pour effectuer les actes de la vie quotidienne.

En ce qui concerne le mode d'indemnisation, à la différence de la garantie incapacité, les prestations se feront sous forme de rentes et non plus sous forme d'IJ.

Le montant de ces dernières peut être soit fixe ou alors dépendre d'un pourcentage du dernier salaire de la personne invalide.
Sans surprise, le montant de prestation sera aussi conditionné par la catégorie d'invalidité.

D'autre part, dans le cas d'un accident survenant sur le lieu de travail ou d'une maladie professionnelle, un taux d'incapacité sera calculé. En fonction de la valeur de ce taux, inférieure ou supérieure à 10%, un capital ou une rente sera versé à l'assuré.

A noter qu'en général, un assuré même invalide passera rarement directement au stade d'invalidité. Il sera dans un premier temps classé en incapacité avant de basculer en invalidité le temps que son état soit consolidé.
Par ailleurs, le versement de la rente est en général suspendu au moment de l'âge théorique de départ à la retraite de la personne.

1.3 Les garanties liées au vieillissement

Dernière garantie du triptyque proposé : la garantie liée au vieillissement.

Les garanties couvrant la dépendance ont pour objectif le versement d'une prestation en cas de perte totale et irréversible d'autonomie. Cette perte d'autonomie est généralement le résultat du vieillissement et que très rarement d'accidents.

Afin de quantifier la perte d'autonomie, deux définitions existent en France.

La première consiste à se référer aux actes de la vie quotidienne tels que se laver, se nourrir, se déplacer et s'habiller. Parmi ces derniers, il conviendra de déterminer ceux pour lesquels l'assuré est dans l'incapacité de les réaliser seul. En fonction du résultat, un niveau de dépendance, partielle ou totale, sera défini.
En complément de cette analyse, un test neuropsychologique, le test de Folstein peut être effectué.

La seconde définition repose quant à elle sur la grille AGGIR. Cette dernière définit 6 niveaux dont 4 pour la dépendance : les niveaux GIR1, GIR2, GIR3 et GIR4.
Les niveaux GIR1 et GIR2 correspondent à un niveau de dépendance totale alors que les niveaux GIR3 et GIR4 correspondent quant à eux à des niveaux de dépendance partielle.

Généralement, les prestations couvrant les garanties liées au vieillissement se font sous forme de rentes viagères. Toutefois, un capital peut être versé pour adapter le logement de l'assuré par exemple.

Pour information, cette garantie ne sera que très peu détaillée ici. En effet, la garantie dépendance n'est pas prise en compte dans le produit SwissLife servant d'étude dans ce mémoire.

1.4 Les modes d'indemnisation

Lorsque qu'un sinistre survient, qu'il soit la conséquence du vieillissement, d'un décès ou d'un arrêt de travail, une prestation est versée. Cette prestation peut se matérialiser sous deux formes : en nature ou en espèces.

Les prestations en nature recouvrent l'ensemble des remboursements de soins y compris de médicaments. Ces frais de santé, comme il peut être facile de le deviner, résultent d'une maladie, d'un arrêt de travail ou encore d'une maternité par exemple.

De l'autre côté, les prestations en espèces font face aux prestations en nature. Ces dernières, ont pour but de remplacer la perte de revenus faisant suite à une incapacité de travailler. Parmi elles, trois sous-catégories pourront être distinguées.

D'une part, il peut s'agir des indemnités journalières. Comme détaillé auparavant, celles-ci sont versées en cas d'arrêt de travail et notamment en cas d'incapacité.

D'autre part, des prestations liées à la maternité peuvent exister. Ces dernières ont pour objectif de compléter ou de venir remplacer le salaire de la mère lors de son congé parental.

Enfin, des prestations sous forme de rentes peuvent être versées en cas d'invalidité.

1.5 La sélection médicale en prévoyance individuelle

Contrairement à la prévoyance collective, l'adhésion à un contrat d'assurance n'est pas automatique en prévoyance individuelle. Peu importe son profil, tout futur assuré devra se soumettre à un questionnaire relatif à son état de santé. C'est ce mécanisme qui est appelé sélection médicale.

Cette pratique a pour objectif de prévenir l'antisélection. En effet, lorsqu'un assuré décide de se couvrir contre un risque, ce dernier détient bien plus d'informations à propos de son état de santé que l'assureur. Ainsi, dans le but d'éviter d'intégrer dans son portefeuille des personnes déjà atteintes de maladies ou ayant de lourds antécédents médicaux, l'assureur soumet le futur assuré à un questionnaire de sélection médicale.

De manière générale, voici un panel de questions pouvant être posées aux futurs assurés :

- Quel est votre âge ?
- Êtes-vous une femme ou un homme ?
- Quel est votre poids, quelle est votre taille ?
- Quelle est votre profession ?
- Pratiquez-vous une activité sportive régulière ?
- Etes-vous fumeur ?
- Avez-vous une maladie chronique ?
- Avez-vous été hospitalisé dernièrement ?

Dans certains cas, ce questionnaire initial devra être complété par un second questionnaire plus précis et adapté en fonction de la pathologie. De plus, l'avis d'un médecin pourra être nécessaire.

La sélection médicale aura également pour but de déterminer un certain nombre d'exclusions et de pathologies non couvertes par le contrat de prévoyance.

Grâce à ces informations, l'assureur, en fonction de sa propre grille d'évaluation bien entendu, sera en mesure de déterminer un tarif optimal.

A noter qu'en général, la sélection médicale n'est effectuée qu'une unique fois lors de la souscription pour l'entière durée de vie du contrat. En sachant que l'état de santé d'une personne reste rarement stable au cours d'une vie, certains enjeux pour l'assureur en termes de sinistralité semblent déjà se dessiner.

2 Zoom sur la prévoyance de certains secteurs professionnels

Dans cette partie, des spécificités en matière de prévoyance concernant certains secteurs d'activités seront mises en avant. En effet, comme exposé précédemment, certaines professions ne sont pas rattachées au régime de base de la sécurité sociale mais à des régimes spécifiques. C'est notamment le cas des TNS ou des travailleurs non-salariés agricoles ainsi que des salariés du monde agricole.

2.1 La prévoyance des TNS ou travailleurs non-salariés agricoles

Une des complexités avec la population des travailleurs non-salariés est qu'elle regroupe un grand nombre de professions très hétérogènes. Derrière l'acronyme TNS peuvent se cacher à la fois des artisans, des professions libérales mais aussi des commerçants et des industriels.
C'est une des raisons pour laquelle, les TNS ne sont pas rattachés directement au régime de base de la Sécurité Sociale mais au régime de la Sécurité Sociale Indépendants (SSI), anciennement connu sous le nom de Régime Social des Indépendants (RSI) jusqu'en 2018. Ce dernier constitue leur régime de base.

Des prestations en nature seront prises en charge par le SSI pour l'ensemble des TNS, alors que les prestations en espèces seront uniquement garanties pour les activités industrielles, artisanales et commerciales.

En supplément de ce régime de base, une seconde couche de régimes obligatoires intervient. Il s'agit des régimes professionnels qui ont pour objectif la couverture des prestations retraite et prévoyance. Pour illustrer ce propos, différentes caisses de retraite comme CANCAVA ou ORGANIC pour les artisans et les commerçants peuvent être citées. Pour les activités libérales, davantage de caisses sont éligibles.

Enfin, comme dans le cas classique des assurés affiliés au régime de base, ces derniers ont la possibilité en supplément de ces deux régimes obligatoires de souscrire un contrat de prévoyance individuelle auprès d'un assureur.

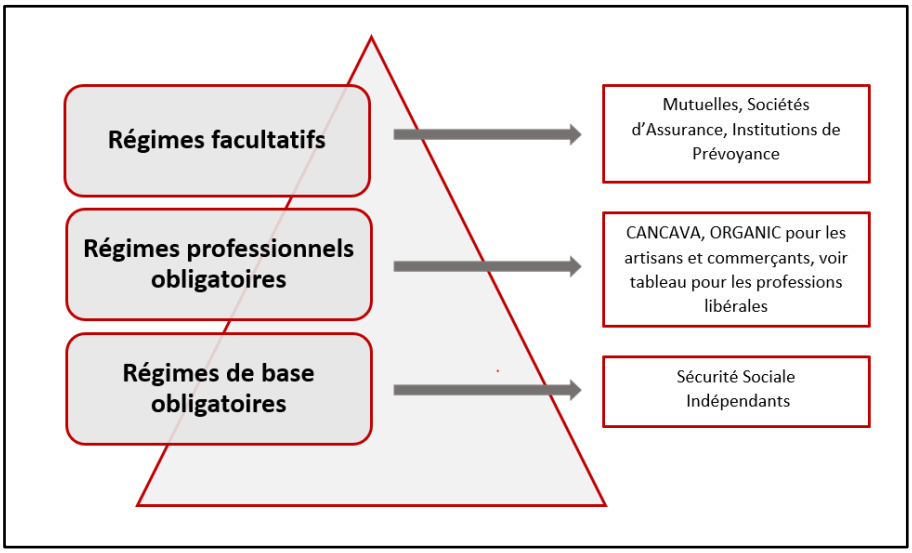


Figure 12 - Pyramide de la prévoyance dans le cas des TNS & non-salariés agricoles

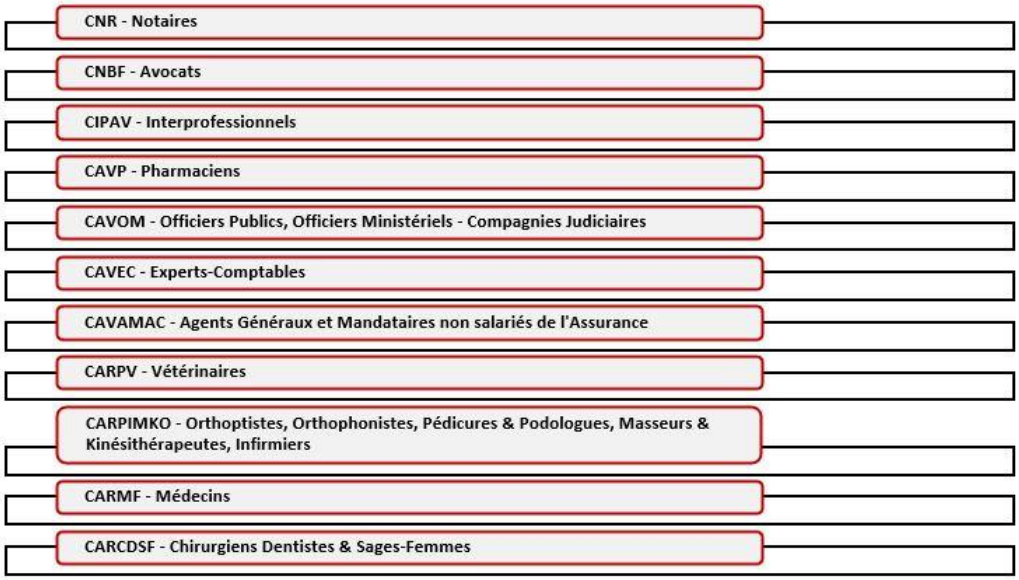


Figure 13 - Liste des régimes de base pour les professions libérales

2.2 La prévoyance des salariés du commerce, de l'industrie, des professions libérales et agricoles, et de l'artisanat

A la différence de leurs homologues non-salariés, les salariés du commerce, de l'industrie, des professions libérales, et de l'artisanat ont pour régime de base le Régime Général de la Sécurité Sociale.

Ce régime a la particularité d'être géré de manière graduelle avec différents niveaux.

Au niveau le plus local, il s'agit de la Caisse Primaire d'Assurance Maladie, plus connue sous l'acronyme de CPAM. Au niveau régional, le relai sera assuré par la Caisse Régionale d'Assurance Maladie (CRAM). Et enfin, le dernier niveau est quant à lui géré par la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS).

Cette gestion à différents niveaux a notamment été établie car le Régime Général de la Sécurité Sociale est le régime qui couvre la plus grande fraction de la population française. Une gestion déléguée était donc nécessaire.

Les salariés du secteur agricole ne sont quant à eux pas rattachés au Régime Général de la Sécurité Sociale. Ces derniers ont pour régime de base le régime de la Mutualité Sociale Agricole (MSA).

Bien que le régime de base de ces deux sous-populations diffère dans le nom, les prestations en espèces versées par la MSA sont quasiment équivalentes à celles versées par le Régime Général de la Sécurité Sociale.

De plus, comme déjà mentionné à plusieurs reprises, en supplément de ces régimes de base, des régimes complémentaires sont obligatoires. Par ailleurs, des régimes facultatifs peuvent aussi être proposés aux salariés de l'industrie, du commerce, des professions libérales et agricoles, et de l'artisanat.

2.3 Des obligations réglementaires

2.3.1 La loi Madelin

Etablie en 1994, la loi Madelin a pour objectif principal d'inciter les travailleurs non-salariés à se protéger contre les risques d'arrêts de travail, d'invalidité et de décès. En effet, les travailleurs non-salariés donc indépendants, sont en général très mal couverts. De plus, les prestations versées par la SSI sont insuffisantes pour se couvrir entièrement.

Dans un objectif d'incitation, la loi Madelin permet donc certains avantages fiscaux, notamment en termes de fiscalité des cotisations. En souscrivant un contrat de type Madelin, l'assuré aura la possibilité de déduire fiscalement ses cotisations versées de son résultat imposable. Ce mécanisme prévaut pour les commerçants, les artisans, les industriels, les professions libérales, tout comme pour leurs conjoints et collaborateurs. Toutefois, sont exclues de ce processus de fiscalité avantageuse, les professions agricoles qui bénéficient déjà d'un régime complémentaire avec une fiscalité avantageuse.

2.3.2 La réforme des TNS

C'est dans un contexte de crise sanitaire que la réforme des TNS a vu le jour. En effet, cette période a été ponctuée par de nombreux arrêts de travail pour raisons multiples : infection Covid, cas contact, garde d'enfants pour écoles fermées et encore diverses autres raisons.

De plus, les TNS contrairement aux autres catégories de professionnels, ne bénéficiaient jusqu'au 30 juin 2021, d'une indemnité journalière versée par l'Assurance Maladie qu'à compter d'un délai de carence de 90 jours. De nombreux arrêts courts n'étaient donc pas couverts dans ce délai de 90 jours.

Dans le but d'égaliser le versement des prestations, la réforme des TNS est entrée en vigueur à compter du 1er juillet 2021.
Elle permet aux professions libérales affiliées à la Caisse Nationale d'Assurance Vieillesse des Professions Libérales (CNAVPL) de se voir verser des indemnités journalières. Les agents généraux, les experts comptables ainsi que les professions médicales, sont des professions qui pourront bénéficier de cette réforme.

A noter que les artisans commerçants ne sont pas concernés par cette réforme car déjà affiliés au Régime Général de l'Assurance Maladie.

Figure 14 - Répartition des prestations avant la réforme des TNS

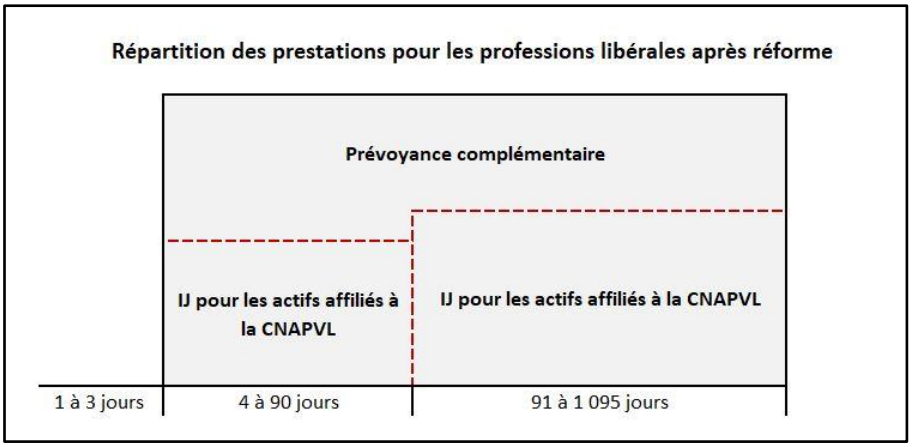


Figure 15 - Répartition des prestations après la réforme des TNS

II. Le portefeuille SwissLife et ses caractéristiques

Afin de mener à bien ce mémoire, toute modélisation réalisée dans celui-ci sera appliquée à un portefeuille SwissLife choisi, à savoir le portefeuille SLPI.

1 Le produit SLPI

1.1 Ses caractéristiques

Etabli en 2010, le produit SwissLife Prévoyance Indépendants plus communément appelé SLPI, est le produit de la marque SwissLife conçu pour répondre aux attentes du marché en matière de prévoyance individuelle des indépendants.

Ce dernier cible en particulier les travailleurs non-salariés non agricoles et exploitants agricoles qui depuis la loi Madelin ont un grand avantage à se protéger.

En quelques mots, SLPI est en premier lieu un contrat collectif à adhésion individuelle de prévoyance. Il doit être souscrit par l'Association Générale Interprofessionnelle de Solidarité (AGIS) auprès des deux entités SwissLife suivantes : SwissLife Prévoyance et Santé et SwissLife Assurance et Patrimoine.

En termes de plan de prévoyance, le produit propose au client selon ses besoins trois garanties classiques en prévoyance individuelle. La première garantie proposée étant la garantie en cas de décès, la seconde étant la garantie de maintien des revenus et la dernière couvrant le paiement des frais généraux.

Concernant la deuxième garantie relative au maintien des revenus, celle-ci s'applique également à la famille de l'adhérent. Le conjoint et enfants potentiels de l'assuré seront couverts en cas d'incapacité, d'invalidité ou de décès.

De surcroit, trois conditions sont nécessaires à la souscription du produit SLPI.
Premièrement, comme énoncé plus haut, il faut obligatoirement être membre de l'AGIS.

Dans un deuxième temps, il faut être âgé de moins de 63 ans. Dans certains cas, une souscription jusqu'à 67 ans peut être autorisée mais cela relève de l'exception et d'une demande spécifique auprès de la compagnie d'assurance.
Par ailleurs, s'il s'avérait que l'âge de départ à la retraite était dépassé, une souscription aux garanties couvrant les arrêts de travail serait impossible.

Enfin, afin de satisfaire la dernière condition de souscription au produit SLPI, il convient d'être à jour en termes de paiement des cotisations auprès des régimes obligatoires d'assurance maladie et d'assurance retraite.

Dans ce qui suit, un détail des différentes garanties proposées va être exposé. Ces garanties étant relativement communes sur le marché de la prévoyance, et un zoom de ces dernières ayant déjà été proposé lors de la première partie sur les grands principes de la prévoyance, seules les spécificités propres au produit SLPI seront détaillées par la suite.

1.1.1 Maintien des revenus

Cette garantie proposée au client permet de couvrir celui-ci en cas d'incapacité ou d'invalidité. Comme évoqué, il existe plusieurs particularités relatives à cette garantie.
Premièrement, les indemnités journalières sont prises en charge en cas de mi-temps thérapeutique. De plus, une option quant à la réduction de la franchise à 3 jours est possible en cas d'hospitalisation de plus de 24 heures ou d'hospitalisation de moins de 24 heures ayant nécessité une intervention chirurgicale et une anesthésie qu'elle soit locale ou générale.
Enfin, un système de bonus de franchise est permis. En effet, si aucun sinistre ne se produit sur une année, un jour de franchise en moins sera gagné en cas de sinistre.

1.1.2 Décès

Sans équivoque, la garantie décès vise à couvrir les conséquences en cas de décès d'une personne, mais aussi en cas de perte totale et irréversible d'autonomie.
L'un des points forts du produit SLPI est le suivant : le capital décès ou le capital versé en cas de PTIA est maintenu jusqu'à 75 ans et ce sans formalité particulière.
Par ailleurs, en complément du capital décès, un capital supplémentaire sera versé en cas de décès accidentel de l'assuré ou en cas de décès simultané ou ultérieur du conjoint de l'assuré.
Pour compléter ceci, le capital versé aux bénéficiaires de l'adhérent sera exonéré en prestation.

1.1.3 Frais généraux

Pour ce qui est de la garantie concernant les frais généraux, aucune spécificité n'est à mentionner. Il s'agit d'une garantie somme toute classique.

Enfin, une des spécificités, valable pour l'ensemble des garanties, est que le contrat SLPI dans son ensemble est éligible à la fiscalité Madelin.

1.2 Quelques évolutions

Lancé en 2010, il est facile d'imaginer que le produit SLPI n'a cessé d'évoluer au cours de ses années de vie.
Adaptations des garanties, ajouts de garanties supplémentaires, remises à niveau règlementaires sont autant d'évolutions qui ont ponctué la vie du produit.

Parmi elles, l'évolution datant de 2018 concernant la grille de professions utilisée par les équipes de tarification mérite d'être mentionnée et mise en avant.

En effet, il s'agit d'une mesure qui orientera l'étude de ce mémoire comme cela pourra être observé par la suite à plusieurs reprises.

Cette évolution n'est rien d'autre qu'une refonte de la grille tarifaire différenciée par profession. Cette dernière permet notamment au produit SLPI d'être parfaitement positionné quant aux professions ciblées souhaitées.

La liste exhaustive des professions cibles ne sera toutefois pas exposée ici. Cependant, huit catégories de professions sont à relever. Pour information, ce seront très souvent ces regroupements qui seront utilisés pour mener les études à venir.

Voici le détail de ces huit catégories cibles, complété d'une liste non exhaustive d'exemples de professions s'intégrant dans chacune de celles-ci.

La première catégorie mentionnée concerne les **professions libérales réglementées**. Parmi elles, sont à retrouver les métiers suivants : avocat, architecte et notaire.

La deuxième catégorie regroupe quant à elle les **professions médicales**. Il s'agit sans nul doute de la catégorie comprenant le plus de professions. Peuvent être citées à titre d'exemples les professions suivantes : anesthésiste, dermatologue, vétérinaire, oncologue, et bien d'autres professions encore.

En troisième lieu, sont à retrouver les **professions paramédicales**. Ces dernières recouvrent par exemple les orthophonistes et les infirmières.

Dans un registre non médical cette fois, les **commerçants** peuvent être mentionnés.
Ils sont à distinguer de leurs comparses **artisans**. Par ailleurs, parmi les artisans, la distinction entre **artisans du BTP** (Bâtiments et Travaux Publics) et **artisans hors BTP** sera faite.
Enfin, les **exploitants agricoles** complètent la liste des professions cibles du produit SLPI en formant la dernière catégorie.

Par ailleurs, dans la suite de ce mémoire, chaque catégorie sera associée à un « code » commençant par la lettre C. De plus, concernant les professions libérales une distinction sera ajoutée entre les **professions libérales médicales** et les **autres professions libérales**.

Voici la cartographie qui sera utilisée :

C0	C1	C2	C3	C4	C5	C6	C7
Professions libérales réglementées	Professions médicales	Autres professions	Professions paramédicales	Commerçants	Artisans hors BTP	Exploitants agricoles	Artisans BTP

Figure 16 - Tableau récapitulatif des professions présentes dans le portefeuille SLPI

2 Le produit SLPI en chiffres

A présent, cette sous-partie a pour objectif de présenter la base utilisée lors de ce mémoire. Afin de mener à bien cet objectif, plusieurs analyses descriptives seront exposées, permettant au lecteur de mieux cerner les caractéristiques et les évolutions des données recueillies.

Avant de commencer, il est manifeste que la base de données ne regroupe que des assurés ayant choisi de souscrire un contrat SLPI.

2.1 Quelques généralités

Premièrement, le produit SLPI ayant été lancé en **2010**, la période d'observation se déroulera donc de cette année de lancement aux dernières observations les plus récentes, à savoir celles datant de l'année **2022**. La fenêtre d'observation est donc de douze ans, permettant ainsi un certain historique.

La base de données est composée de **528 959** lignes pour **107 644** assurés distincts. Il est possible qu'un même assuré apparaisse plusieurs fois dans la base de données si son contrat n'est pas résilié au bout d'un an. En effet, la base de données a été construite de façon à avoir une ligne pour chaque assuré par année comptable. Un assuré entrant dans le portefeuille en 2010 et quittant ce dernier en 2013 aura donc quatre lignes qui lui seront dédiées, pour les années comptables 2010, 2011, 2012 et 2013.

Par ailleurs, afin d'illustrer les caractéristiques de la base SLPI, plusieurs indicateurs ont été étudiés.

Le premier indicateur mentionné ici est l'âge moyen à la souscription. En moyenne, les assurés du portefeuille SLPI souscrivent leur contrat d'assurance à **41 ans**.
A partir de la quarantaine, les personnes ont généralement une situation plus stable et se sentent par ailleurs plus concernées par des questions de prévoyance et d'arrêt de travail, là où un jeune actif de 25 ans n'aura probablement pas la même réflexion. Cette remarque, bien que très brève permet de justifier la cohérence de ce 41 ans obtenu.

De plus, il a été observé qu'en moyenne un assuré restait **quatre années** au sein du portefeuille. Cette durée semble légèrement plus faible par rapport à d'autres études démontrant qu'en général un assuré reste en moyenne entre 6 et 7 ans. Toutefois, cela reste une caractéristique propre au portefeuille SLPI, qui n'aura aucun impact sur la suite de l'étude menée.

2.2 La répartition hommes / femmes

Concernant la répartition des hommes et des femmes au sein du portefeuille, aucune disproportionnalité majeure n'est à observer. La base de données compte **47%** de femmes pour **53%** d'hommes. Une répartition parfaite n'existant pas, la base de données peut alors être considérée comme équilibrée quant à la répartition hommes / femmes.

Bien que cette répartition puisse paraître futile au premier abord, elle est toutefois importante à analyser. En effet, une part disproportionnée de femmes pourrait potentiellement entrainer davantage d'arrêts de travail liés à des grossesses par exemple.

Le violon plot ci-dessous permet de synthétiser les deux informations évoquées précédemment en représentant conjointement l'âge à la souscription suivant le sexe.
Le second violon plot à droite, représente quant à lui la répartition de l'âge suivant le sexe au sein du portefeuille SLPI.

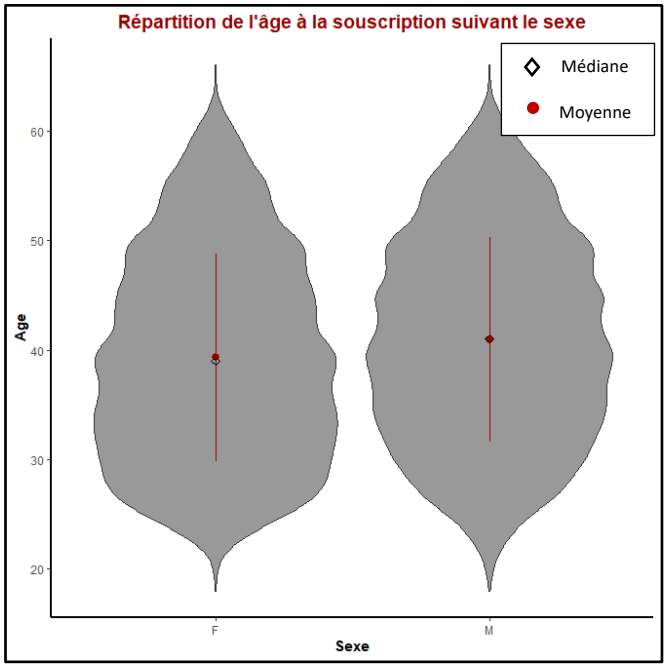


Figure 17 - Violon plot - Répartition de l'âge à la souscription suivant le sexe

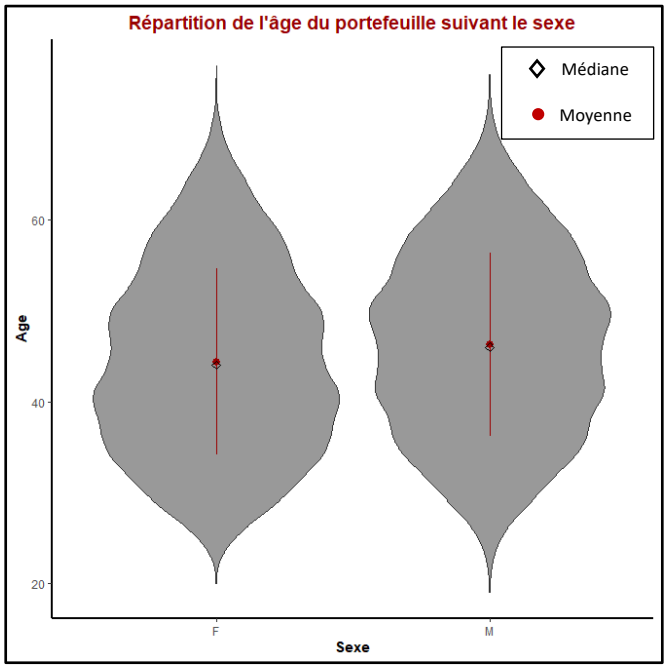


Figure 18 - Violon plot - Répartition de l'âge au sein du portefeuille SLPI en fonction du sexe

Grâce au violon plot de gauche, l'âge moyen de souscription autour de 41 ans est ainsi reflété. Sur celui de droite, il est possible d'observer que l'âge des assurés du portefeuille SLPI s'étend de 19 ans à plus de 60 ans.

2.3 Les catégories socio professionnelles

A présent, un focus sur la répartition des catégories socio professionnelles (CSP) va être proposé à l'appui du pie plot suivant.

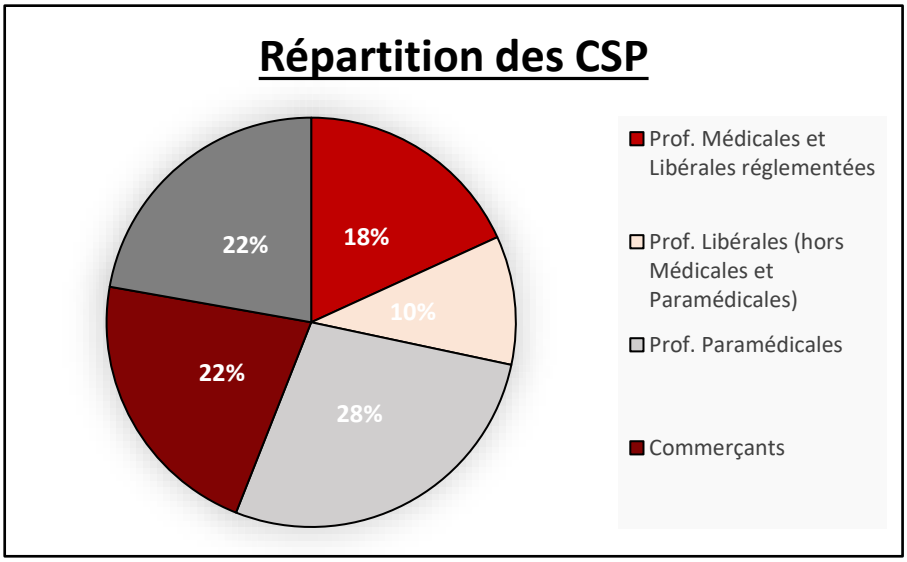


Figure 19 - Pie plot - Répartition des CSP au sein du portefeuille SLPI

A noter que ce graphique a été construit à l'aide des classes de professions qui avaient été détaillées lors de la présentation du produit SLPI. Par choix, et dans un souci de simplification du pie plot, certaines professions ont été regroupées comme suit :

- Professions médicales et libérales réglementées : classes **C0** et **C1**
- Professions libérales hors médicales et paramédicales : classe **C2**
- Professions paramédicales : classe **C3**
- Commerçants : classe **C4**
- Artisans ou exploitants agricoles : classes **C5**, **C6** et **C7**

Une telle association des professions met en lumière une prédominance des professions paramédicales, suivies de près par les commerçants et les professions libérales médicales et réglementées. Cette prédominance des professions paramédicales est à garder en tête pour la suite de l'étude, car elle permettra de comprendre certaines observations et décisions prises dans les modèles notamment.

2.4 Primes et sinistralité

Afin de mieux connaitre le portefeuille, il a semblé important de présenter également quelques éléments reflétant les volumes de production et la sinistralité propres au produit SLPI.

Grâce au graphique ci-dessous, présentant l'évolution du volume de primes depuis le lancement du produit en 2010, la première conclusion qui peut être apportée est que le portefeuille est en fort développement.
Pour information, les primes relatives à l'exercice comptable 2022 n'étant pas encore entièrement collectées au moment de l'étude, il a été décidé de ne pas les intégrer au graphique suivant. De plus, dans les Figures 20, 21 et 22, pour des raisons de confidentialité, les valeurs des axes des ordonnées ont été masquées.

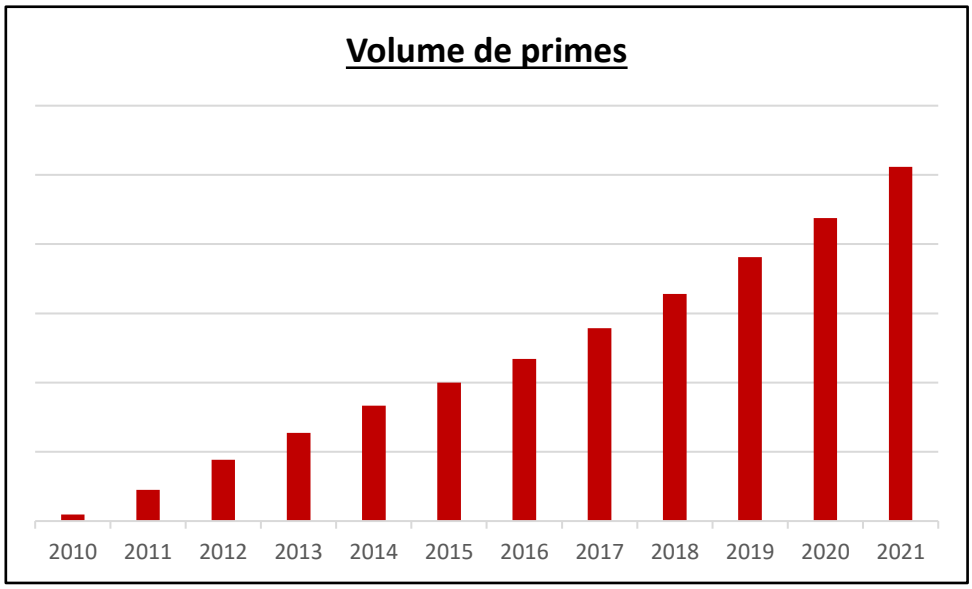


Figure 20 - Evolution du volume de primes pour le produit SLPI en fonction des années comptables

Dans un second temps, il est important d'avoir une idée précise au sujet de la sinistralité du portefeuille étudié dans le but de mieux le connaitre. C'est la vocation du graphique suivant.

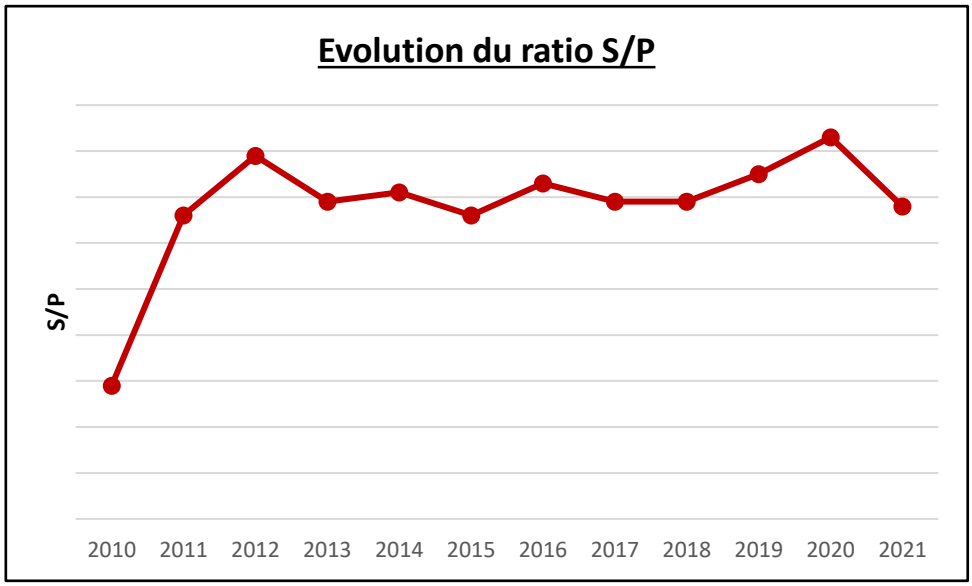


Figure 21 - Evolution du ratio S/P pour le portefeuille SLPI

Le pic de sinistralité est observable en 2020. Il est assez facile de voir que cette forte évolution est directement en lien avec la crise sanitaire de la Covid qui a eu pour conséquence de démultiplier les arrêts de travail au cours de cette période.

Il est important de souligner également qu'en prévoyance le ratio S/P peut rapidement fluctuer. En effet, il suffit d'un seul sinistre pour dégrader fortement ce dernier. Par conséquent, il est donc difficile de statuer une valeur cible de S/P.

Enfin, avant de refermer ce volet consacré à la description du portefeuille SLPI chez SwissLife, une analyse croisée entre catégories socio professionnelles et sinistralité a été menée.

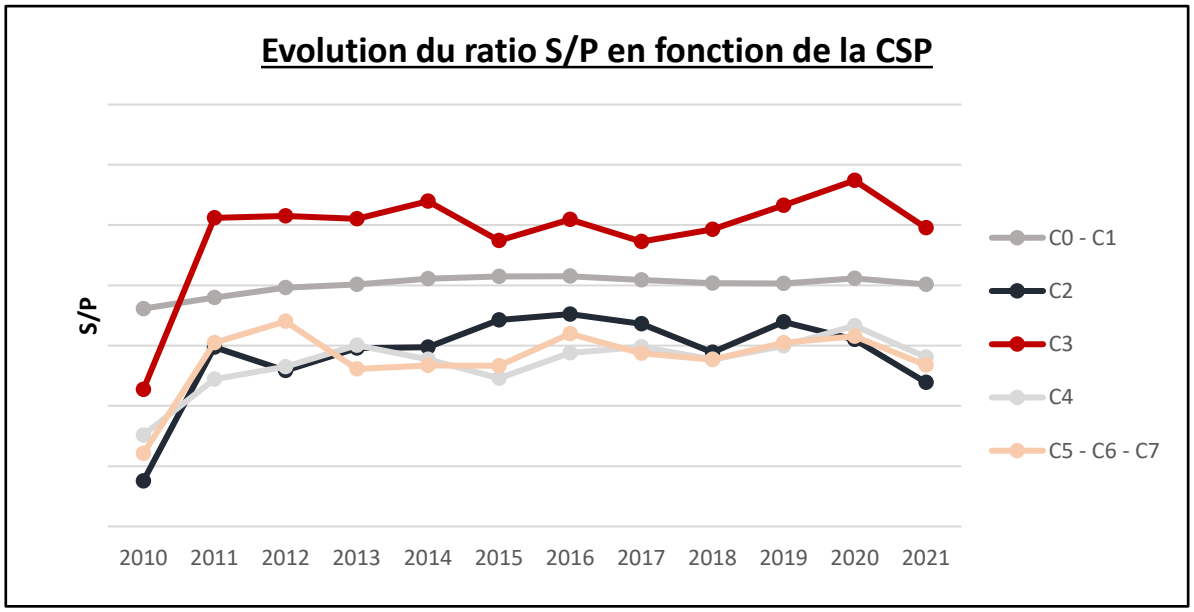


Figure 22 - Evolution du ratio de S/P en fonction des classes de professions

Grâce à ce dernier graphique, présentant l'évolution du ratio de sinistralité au cours des années comptables en fonction des classes professionnelles choisies, deux catégories se distinguent des autres. Il s'agit de la classe **C0** associée à la classe **C1**, et de la classe **C3**. Pour rappel, les classes C0 et C1 correspondent aux professions libérales médicales, et la classe C3 aux professions paramédicales.
Concernant les professions libérales médicales, telles que des médecins ou des chirurgiens, une explication possible afin de justifier une forte sinistralité pourrait résider dans le fait que de tels assurés ont un salaire de base élevé. Le montant d'une indemnisation en cas d'arrêt de travail étant dépendant du salaire, la prestation serait donc élevée.
Il est important de souligner que contrairement aux professions paramédicales, la hausse de sinistralité est relativement stable dans le temps. Un tel ratio est donc prévisible d'année en année.

En revanche, la hausse de sinistralité concernant les professions de type C3 s'est accentuée au cours des années avec pour point d'orgue l'année 2020, année de la pandémie. Ce pic de sinistralité est également présent pour les autres classes de professions.

Cette fois-ci, l'augmentation de la sinistralité pourrait être expliquée par la pénibilité des métiers paramédicaux. Depuis plusieurs années maintenant, une réelle crise des infirmières persiste. Des arrêts ayant des raisons émotionnelles ou psychiques pourraient être des facteurs explicatifs de cette hausse continuelle de la sinistralité au cours du temps.

Par ailleurs, il est intéressant de garder en tête que la fréquence des arrêts de travail pour la population TNS est inférieure à celle des travailleurs salariés. En effet, une entreprise constituée entièrement de salariés peut continuer à fonctionner correctement en cas d'un ou de plusieurs arrêts de travail, là où cela est bien plus délicat dans le cas d'une société gérée par des indépendants.

Pour résumer, le portefeuille SLPI développé en 2010 est un portefeuille en forte croissance. Bien que correctement équilibré quant à la répartition hommes / femmes, des disparités au niveau de la sinistralité entre les différentes professions semblent se dessiner.

Ces premières observations ne peuvent que conforter le rôle et l'importance de la sélection médicale au moment de la souscription. Une question reste alors entière : comment l'optimiser ?

III. Des outils théoriques au service de la modélisation

L'objectif de cette partie est de présenter les différents outils théoriques qui interviendront tout au long de ce mémoire.

Dans toute celle-ci, une indication relative à la source des informations utilisées sera mentionnée. Le lecteur pourra également être invité dans certains cas à consulter certaines références présentes à la fin du mémoire pour davantage d'informations sur l'aspect théorique exposé.

1 Une adaptation du ratio SMR

Généralement utilisé pour comparer la mortalité de deux populations distinctes, le ratio SMR ou Standard Mortality Ratio, peut être adapté afin d'effectuer une comparaison non plus relative à la mortalité mais à la morbidité.

Pour davantage d'informations sur la théorie exposée ci-dessous, le lecteur pourra consulter la référence présentée ici [1].

Soient les variables suivantes :

$\boldsymbol{S_x}$ **:** le nombre d'arrêts de travail concernant des assurés d'âge $\boldsymbol{x}$ de la population d'étude

$\boldsymbol{i_x}$ **:** la probabilité d'incidence en arrêt de travail pour la population de référence selon l'âge $\boldsymbol{x}$

$\boldsymbol{E_x}$ **:** le nombre d'assurés d'âge $\boldsymbol{x}$ de la population d'étude exposés à un arrêt de travail

L'idée du ratio standardisé est de construire un rapport entre deux quantités à comparer. Dans le cas présent, les deux quantités à comparer sont les suivantes :

- $\sum_x \boldsymbol{S_x}$: le nombre d'arrêts de travail observés dans la population d'étude
- $\sum_x \boldsymbol{i_x} \times \boldsymbol{E_x}$: le nombre d'arrêts de travail attendus pour les assurés d'âge $\boldsymbol{x}$ si l'incidence était identique à celle de la population de référence

Ces deux quantités ainsi construites permettent alors d'obtenir la statistique de test suivante :

$$t = \frac{\sum_x S_x}{\sum_x i_x \times E_x}$$

Concernant l'interprétation de ce ratio standardisé, trois conclusions sont possibles en fonction des valeurs prises par ce dernier.

- Une valeur de la statistique de test t **inférieure à 1** indiquera que la sinistralité de la population d'étude est inférieure à la sinistralité de la population de référence.
  Autrement dit, que la sélection médicale permet de se ramener à une sinistralité plus faible, et qu'elle a donc un effet.

- Une valeur de la statistique de test t **supérieure à 1** indiquera quant à elle que la sinistralité de la population d'étude est supérieure à la sinistralité de la population de référence.
  Autrement dit, que la sélection médicale n'est plus efficace.

- Une valeur de la statistique de test t **égale 1** signifiera que le nombre de sinistres observés dans la population d'étude est égal à celui observé dans la population de référence.
  Ce cas de figure est très rarement observable dans la réalité.

De plus, un intervalle de confiance relatif à ce ratio de morbidité peut être construit comme suit :

$$t \in \left[\frac{S}{I} \times \left(1 - \frac{1}{9 \times S} - \frac{q_{1-\frac{\alpha}{2}}}{3\sqrt{S}}\right)^3 ; \frac{S+1}{I} \times \left(1 - \frac{1}{9 \times (S+1)} - \frac{q_{1-\frac{\alpha}{2}}}{3\sqrt{S+1}}\right)^3\right]$$

où :

- S est la quantité définie par $S = \sum_x S_x$
- I est la quantité définie par $I = \sum_x i_x \times E_x$
- $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi normale centrée réduite

Pour la suite des différents travaux menés, le niveau de confiance sera fixé à 95%.

2 Les modèles linéaires généralisés ou GLM

La partie théorique à venir vise à présenter les principes généraux des modèles linéaires généralisés, encore connus sous l'acronyme de GLM.

Celle-ci s'inspire largement de cours dispensés lors du cursus de formation à l'ISUP, ainsi que d'autres références pour lesquelles le lecteur est invité à se référer ici si besoin [2].

D'un point de vue statistique, déterminer une prime d'assurance revient à effectuer une régression et par la suite de modéliser la quantité suivante :

$$g(X) = \mathbb{E}[Y|\, X]$$

avec

$$Y = g(X) + \varepsilon$$

Et où ε représente l'écart commis lorsque la variable Y est remplacée par son espérance conditionnelle. En d'autres termes, il s'agit d'un bruit blanc.

Dans le cas où g est linéaire et que ε est gaussien, le modèle de régression précédent sera défini comme un modèle linéaire gaussien.

Bien que ce dernier présente bien de nombreux avantages, il ne reste pas moins limité par deux problèmes.
D'une part, la forme linéaire de g qui est très restrictive, et d'autre part, le caractère gaussien du bruit blanc.

Dans le but de remédier à cela, ont été introduits en 1972 les modèles linéaires généralisés (Generalized Linear Models – GLM).
Ils permettent ainsi une forme plus générale en relâchant l'hypothèse de linéarité tout en gardant une certaine simplicité. Par ailleurs, les coefficients du modèle sont estimés via un maximum de vraisemblance provenant d'une famille de lois exponentielles.
La loi conditionnelle devra donc nécessairement appartenir à cette famille exponentielle pour que l'estimation soit probante.

Définition [Famille exponentielle] : Un modèle statistique $\left(\Omega, \mathcal{F}, (\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi>0}\right)$ est appelé famille exponentielle si les probabilités $\mathbb{P}_{\theta,\phi}$ admettent une densité f par rapport à une mesure dominante avec

$$f_{\theta,\phi} = c_\phi(y)\exp\left(\frac{y\theta - a(\theta)}{\phi}\right)$$

où :

- θ est le paramètre canonique
- ϕ est le paramètre de dispersion
- $a(\theta)$ est une fonction convexe de classe C^2
- $c_\phi(y)$ est une quantité indépendante de θ

Une proposition découle alors de cette définition.

Proposition : Si Y est distribuée selon une loi appartenant à une famille exponentielle, alors

$$\begin{aligned} \mathbb{E}[Y] &= a'(\theta) \\ \mathbb{V}[Y] &= \phi a''(\theta) \end{aligned}$$

A présent, voici une illustration de quelques distributions appartenant à une famille exponentielle.

- La loi normale : $\mathcal{N}(\mu, \sigma^2)$
  Dans ce cas, $\theta = \mu$, $a(\theta) = \frac{\theta^2}{2}$ et $\phi = \sigma^2$
- La loi exponentielle ou plus généralement la loi gamma de paramètres k et λ
  Dans ce cas, $\theta = \frac{-1}{\lambda}$, $a(\theta) = -k \times log(-\theta)$ et $\phi = 1$
- La loi de Poisson de paramètre λ
  Dans ce cas, $\theta = log(\lambda)$, $a(\theta) = e^{\theta}$ et $\phi = 1$

A noter que la loi Log-Normale ne fait pas partie des distributions appartenant à une famille exponentielle.

Définition [Modèle linéaire généralisé] : Un modèle est défini comme un modèle linéaire généralisé s'il satisfait les hypothèses suivantes :

a) $Y|X = x \sim \mathbb{P}_{\theta,\phi}$ appartient à une famille exponentielle

b) $g\big(\mu(X)\big) = g(\mathbb{E}[Y|X]) = X\beta$ pour une fonction g appelée fonction de lien bijective et $\mu(X) = \mathbb{E}[Y|X]$

Il est possible de remarquer, qu'en choisissant pour fonction de lien la fonction identité, et qu'en supposant que $Y|X = x$ suit une loi gaussienne, une bascule dans le cas des modèles linéaires gaussiens se fait. Ces derniers ne sont en réalité rien d'autre qu'un cas particulier des modèles linéaires généralisés.

A présent, voici un tableau récapitulatif présentant plusieurs distributions ainsi que leur fonction de lien canonique associée.

Loi	Fonction de lien canonique
Normale	$g(\mu) = \mu$
Poisson	$g(\mu) = \log \mu$
Gamma	$g(\mu) = \frac{1}{\mu}$
Bernouilli	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

Figure 23 - Exemples de fonctions de lien canoniques

Une fois le modèle défini, vient alors l'estimation des paramètres. Celle-ci concerne deux paramètres : le paramètre β et le paramètre $\emptyset$ de nuisance.

Estimation du paramètre $\boldsymbol{\beta}$

L'estimation du paramètre β est réalisée par maximum de vraisemblance. L'équation à résoudre est la suivante :

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}}{\operatorname{argmax}}\, \ell(\beta)$$

La vraisemblance du modèle, sans oublier le cadre d'une famille exponentielle est alors :

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} f(Y_i|\beta_i, \phi\,) = \exp\left(\sum_{i=1}^{n} \frac{Y_i\theta_i - a(\theta_i)}{\phi} + \sum_{i=1}^{n} c_\phi(Y_i)\right)$$

En supposant que les Y_i sont indépendants, la log-vraisemblance se réécrit alors de la façon suivante :

$$\ell(\beta) = \sum_{i=1}^{n} \log\left(f(Y_i|\beta_i, \phi)\right) = \sum_{i=1}^{n} \underbrace{\left\{\log\left(c_\phi(Y_i) + \frac{Y_i\theta_i - a(\theta_i)}{\phi}\right)\right\}}_{:=\ell_i(\theta_i)}$$

L'équation du 1er ordre à résoudre devient alors :

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_{\theta_i,\phi}(y_i)}{\partial \beta_j} = 0$$

Il est possible de démontrer après plusieurs étapes de calcul que :

$$\Longrightarrow \frac{\partial \ell}{\partial \beta_j} = 0 \iff X'D(Y - g^{-1}(X\beta)) = 0 \;\forall\, j \in [1, p]$$

où D représente la matrice diagonale dont les coefficients sont $\frac{1}{g'(\mathbb{E}[Y_i|\,X_i])a''(\theta_i)}$.

L'estimateur de $\beta, \widehat{\beta}$ est donc solution de :

$$X'D\big(Y - g^{-1}(X\beta)\big) = 0$$

Estimation du paramètre ϕ

Tout comme le paramètre β, le paramètre ϕ peut être estimé par maximum de vraisemblance. ϕ n'étant qu'un paramètre secondaire, la démonstration ne sera pas mentionnée ici.

Sélection du modèle

Une fois le modèle établi, il est important de vérifier la cohérence de ce dernier et notamment de sélectionner les variables à intégrer.

Contrairement aux modèles linéaires simples, il n'existe pas de caractéristiques à vérifier concernant les résidus. En effet, ces derniers n'ont aucune raison d'être gaussiens ou encore de posséder la même variance.

Une façon de vérifier la pertinence du modèle est alors de comparer celui-ci avec le modèle le plus général et le plus simple qu'il soit : le modèle saturé. Il s'agit du modèle avec autant de paramètres que de variables.

Soient :

- $\mathcal{M}$ le modèle étudié
- $\widetilde{\mathcal{M}}$ le modèle saturé associé
- $\tilde{\beta}$ l'estimateur du maximum de vraisemblance de β dans le modèle saturé
- $\tilde{\ell}$ la vraisemblance du modèle saturé

$\tilde{\beta}$ est alors solution de l'équation suivante :

$$X'D\big(Y - g^{-1}(X\beta)\big) = 0$$

Or, la matrice X est de rang plein dans le modèle saturé, elle est donc inversible. L'équation devient alors :

$$Y - g^{-1}\big(X\tilde{\beta}\big) = 0$$
$$\Leftrightarrow Y_i = g^{-1}\big(X_i'\tilde{\beta}\big) = \tilde{Y}_i \ \forall\, i$$

L'idée est donc de comparer la vraisemblance de ces deux modèles grâce au rapport de vraisemblance $\mathcal{D}$ défini par :

$$\mathcal{D} = -\log\left(\frac{\ell}{\tilde{\ell}}\right)$$

De grandes valeurs de $\mathcal{D}$ indiqueraient que le modèle choisi serait mal ajusté aux données par rapport au modèle saturé.

Définition [Déviance] : La déviance du modèle ajusté est définie par

$$\Delta = 2\mathcal{D} = 2(\tilde{\ell} - \ell)$$

Théorème de Wilks : Si les hypothèses du modèle linéaire généralisé sont satisfaites alors quand n tend vers l'infini,

$$\Delta \xrightarrow{\mathcal{L}} \chi^2_{n-p}$$

où p représente le nombre de paramètres à estimer et χ^2_{n-p} une loi du Khi-Deux à $n-p$ degrés de liberté

3 Des techniques de classification en Machine Learning

Tout comme pour les GLM, la théorie présentée ici s'appuiera en grande partie sur les cours dispensés dans le cadre de la formation universitaire à l'ISUP. Des informations complémentaires y seront également ajoutées. Pour plus de détails, le lecteur pourra se référer ici [3].

3.1 Les arbres CART

Membres des algorithmes de Machine Learning non paramétriques, les arbres CART (Classification And Regression Trees) sont des algorithmes d'apprentissage supervisé possédant une structure d'arborescence.
Ces derniers sont composés d'un nœud initial appelé racine, de branches, de nœuds intermédiaires et de nœuds finaux portant le nom de feuilles.

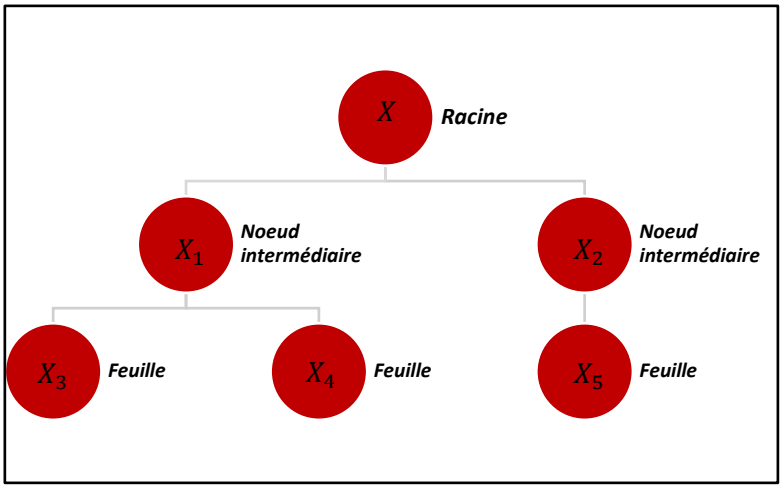


Figure 24 - Exemple d'arbre CART

Dans le cas de la classification, cas concerné par l'étude, le passage de nœud en nœud se fait en répondant à la question suivante :

$$X_j \geq \alpha \, ?$$

pour j et α bien définis

Tout l'objectif lorsqu'il s'agit d'arbres de classification est de réussir à trouver l'arbre optimal. Pour y parvenir, différentes mesures existent. Ces dernières permettent de refléter l'homogénéité au sein de chaque classe, celles-ci devant bien entendu être les plus homogènes possible.

Mesures d'impureté

Soit un arbre de classification ayant pour étiquettes les $y_i \in \{1, \dots, K\}$.

Un nœud N peut être séparé en deux nœuds « fils » : un nœud gauche N_L et un nœud droit N_R, définis comme suit :

$$N_L(j,t) := \{x \in N : x_j < t\} \text{ et } N_R(j,t) := \{x \in N : x_j \geq t\}$$

La distribution des classes est donnée par :

$$p_N = (p_{N,1}, \dots, p_{N,K}) \quad \text{où} \quad p_{N,k} = \frac{\#\{i : x_i \in N, y_i = k\}}{|N|}$$

avec $|N| = \#\{i : x_i \in N\}$ le cardinal de N

Les deux mesures d'impureté ou de non-homogénéité à considérer sont alors :

- L'indice de Gini
- L'indice d'entropie

Indice de Gini

L'indice de Gini peut être vu comme la fréquence à laquelle un individu est incorrectement classé dans une classe.

Voici la formule permettant de décrire ce premier indice :

$$\begin{aligned} G(N) &= G(p_N) \\ &= \sum_{k=1}^{K} p_{N,k} \times (1 - p_{N,k}) \end{aligned}$$

Une cellule pourra alors être considérée comme pure lorsque la valeur de l'indice de Gini associé sera égale à zéro.
En effet, $G(N) = 0$ car pour une unique valeur de $k, p_{N,k} = 1$ et pour toutes les autres valeurs $p_{N,k} = 0$.

Indice d'entropie

Les valeurs prises par l'indice d'entropie sont comprises entre 0 et 1. Dans le cas où tous les échantillons d'un jeu de données appartiendraient à la même classe, alors l'entropie serait nulle.

La formule associée à l'indice d'entropie est la suivante :

$$\begin{aligned} H(N) &= H(p_N) \\ &= -\sum_{k=1}^{K} p_{N,k} \times \log_2(p_{N,k}) \end{aligned}$$

Gain d'information

Que ce soit dans le cas de l'indice de Gini ou dans celui de l'indice d'entropie, le gain d'information noté IG est défini par :

$$IG(j,t) = I(N) - \frac{|N_L(j,t)|}{|N|} \times I\big(N_L(j,t)\big) - \frac{|N_R(j,t)|}{|N|} \times I(N_R(j,t))$$

où $I = G$ ou $I = H$

Les valeurs prises par ces deux indices peuvent être synthétisées à l'aide du graphique ci-dessous.

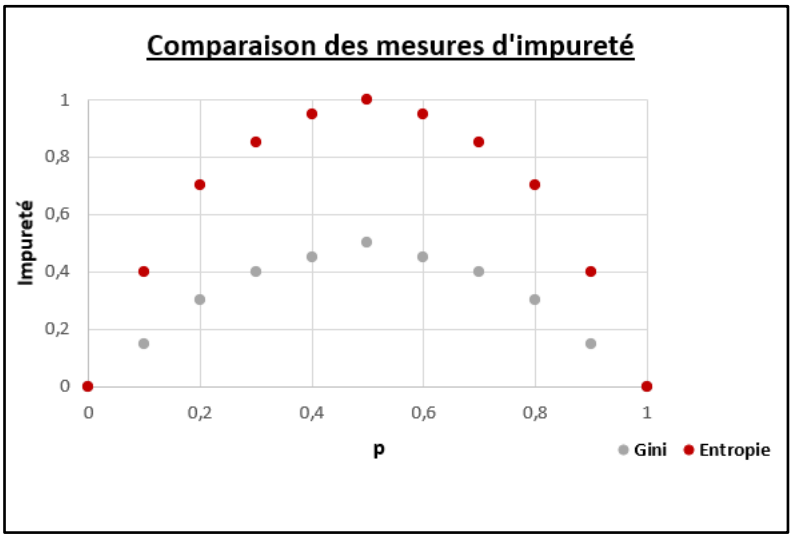


Figure 25 – Comparaison des mesures d'impureté

Bien qu'étant faciles à interpréter et très visuels, les arbres de décisions présentent différentes limites.

La première limite des arbres de décision concerne l'overfitting ou le surajustement. Bien souvent, plus l'arbre sera grand plus un surajustement sera présent. Une façon de combattre ce phénomène reste de mettre en place un élagage ou pruning permettant de limiter l'arborescence.

La seconde limite commune aux arbres de décision est qu'ils souffrent de forte variance. De fait, de légères variations dans l'échantillon utilisé peuvent mener à des arbres de décision très différents. La méthode dite de bagging permet alors de réduire cet inconvénient.

3.2 Une illustration du bagging : les forêts aléatoires

Bagging

Le bagging, aussi connu sous le nom de Bootstrap Aggregating, est une méthode couramment utilisée dans le but de réduire la variance d'un ensemble de données bruyant.

Le principe de cette technique est le suivant. De nouveaux échantillons bootstrappés sont créés par tirages aléatoires avec remise dans l'échantillon initial. Une fois les nouveaux échantillons obtenus, l'arbre de décision est entrainé sur ces derniers. Les estimateurs intermédiaires indépendants ainsi obtenus permettent de déterminer l'estimateur final en utilisant un vote à la majorité dans le cas de la classification par exemple.
Les forêts aléatoires en sont un parfait exemple.

Forêts aléatoires

Etablies en 2001 par Leo Breiman, les forêts aléatoires sont des algorithmes qui combinent arbres de décision et techniques de bagging.

L'algorithme effectue un apprentissage parallèle sur plusieurs classifieurs appelés classifieurs faibles.

Les forêts aléatoires sont aussi connues pour compter parmi les techniques de Machine Learning les plus efficaces.

3.3 Une illustration du boosting : l'algorithme du Gradient Boosting

Boosting

Le boosting est une technique qui comme le bagging permet l'agrégation de classifieurs faibles. Cependant, à la différence du premier, l'apprentissage n'est plus réalisé sur des échantillons indépendants. Les différents échantillons dépendent les uns des autres. En effet, une fois le premier arbre lancé sur le premier échantillon, le second échantillon sera obtenu en fonction des points incorrectement classés à l'issue de cette première étape. L'algorithme n'est donc plus réalisé de manière parallèle mais bien de façon séquentielle.

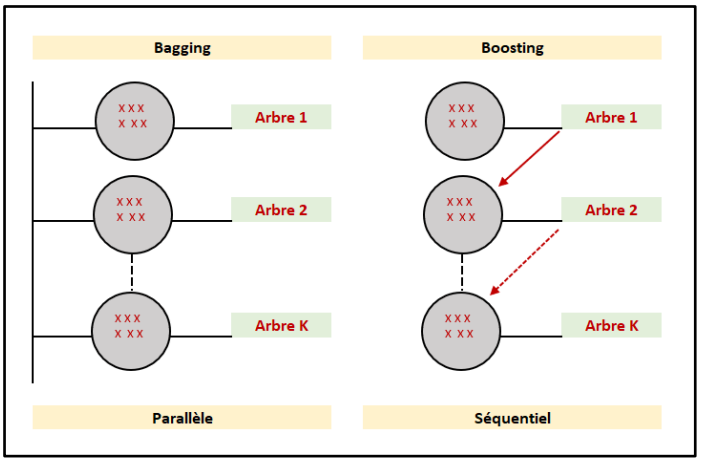


Figure 26 - Principe Bagging vs Boosting

Gradient Boosting

La méthode de Gradient Boosting est une méthode permettant de transformer les apprenants faibles en apprenants forts quelle que soit la fonction de perte à condition qu'elle reste convexe et différentiable. Par conséquent, cela offre un choix plus large en termes de fonctions de perte et de classifieurs faibles.

Soient $h_1, \dots, h_B \in \mathcal{H}$ où $\mathcal{H}$ désigne l'ensemble fini des classifieurs faibles et $\eta_1, \dots, \eta_B \in \mathbb{R}$ tels que

$$g_B(x) = \sum_{b=1}^{B} \eta_b h_b(x)$$

minimise le risque empirique $\frac{1}{n} \times \sum_{i=1}^{n} \ell(y_i, g_B(x_i))$ pour ℓ une fonction de perte comme par exemple celle associée aux moindres carrés, la fonction exponentielle ou encore la fonction de perte logistique.

Ainsi, de manière itérative à chaque pas $b+1$ seront estimés g_{b+1} et η_{b+1} de la façon suivante :

$$(\hat{\eta}_{b+1}, \hat{h}_{b+1}) = \underset{\eta \in \mathbb{R}, h \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(y_i, \hat{g}_b(x_i) + \eta h(x_i)) \quad (*)$$

Et $\hat{g}_{b+1} = \hat{g}_b + \hat{\eta}_{b+1} \hat{h}_{b+1}$.

Toutefois, un problème apparaît. La minimisation associée à ce problème est très couteuse à chaque itération.

Une solution est alors de remplacer chaque étape de minimisation « exacte » par une étape de gradient.
Le problème $(*)$ devient alors

$$\underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(y_i, \hat{g}_b(x_i) + \eta u)$$

avec

$$\hat{\delta}_b = \nabla_u \left(\sum_{i=1}^{n} \ell(y_i, \hat{g}_b(x_i) + u) \right)_{|u=0}$$

Le tableau ci-dessous permet de synthétiser les avantages et inconvénients des différentes méthodes de classification exposées plus haut.

	Arbres CART	Forêts aléatoires	Gradient Boosting
Avantages	- Facilité d'interprétation	- Efficacité - Réduction de la variance	- Large choix de fonctions de perte et de classifieurs - Apprentissage itératif
Inconvénients	- Overfitting - Forte variance	- Apprentissage parallèle - Boite noire	- Minimisation couteuse - Boite noire

Figure 27 – Tableau récapitulatif des avantages et inconvénients des méthodes de classification

La présentation des méthodes de classification vient clôturer cette partie théorique destinée à présenter les différentes techniques sur lesquelles s'appuieront les modélisations mises en place dans les parties suivantes.

IV. Modélisations autour de la sélection médicale : scénario central

Cette quatrième partie sera entièrement consacrée à la mise en place de modèles afin de répondre à la problématique exposée.
Chaque sous-partie de cette même partie sera alors toujours structurée comme suit :

- L'objectif du modèle proposé en vue de répondre à une sous-partie de la problématique
- Construction d'hypothèses complémentaires éventuelles au modèle choisi
- Le modèle d'un point de vue pratique
- Les résultats et réponses à la problématique

1 Le ratio de morbidité : un dérivé du ratio SMR

Bien qu'obligatoirement présente en prévoyance individuelle, il est légitime de se questionner sur le réel impact de la sélection médicale. Une population soumise à cette dernière aura-t-elle vraiment une sinistralité inférieure à une population exempte de sélection ?

Dans le but de répondre à cette problématique, il a été décidé de s'appuyer sur un ratio de morbidité.

Le Standard Mortality Ratio, en français le ratio standardisé de mortalité, ou encore le ratio SMR, est comme son nom l'indique un ratio communément utilisé dans la littérature afin de comparer la mortalité d'une population d'étude avec la mortalité d'une population de référence.

Le nombre de sinistres rattachés à un décès étant très peu nombreux dans la base d'étude du produit SLPI, il a été décidé d'appliquer ce ratio non plus à la mortalité mais à la morbidité au sens large. La morbidité au sens large sous-entendant les arrêts de travail liés à une incapacité ainsi que ceux liés à une invalidité.

Pour y parvenir une adaptation de la théorie du ratio SMR « classique » a été nécessaire, comme cela a été présenté lors de la partie théorie.

1.1 Les hypothèses complémentaires

Comme présenté dans la partie théorie, la comparaison de la sinistralité de la population d'étude avec la sinistralité de la population de référence repose notamment sur la probabilité d'incidence en arrêt de travail dans cette même population de référence. Cette donnée doit donc être supposée connue.

Une première étape des travaux a donc été de se procurer cette précieuse information.

Afin d'y parvenir, il a été décidé de s'inspirer d'une étude menée entre le 1er octobre 2002 et le 15 mai 2008 disponible sur le site de l'Assurance Maladie. Celle-ci avait été menée dans le but de présenter des résultats sur les taux d'incidence en arrêt de travail et les facteurs individuels associés à ces arrêts, et ce pour une population de travailleurs indépendants.

L'étude a donc été réalisée sur une population regroupant 48 654 personnes, toutes affiliées au Régime Social des Indépendants (RSI) à l'époque.
Plusieurs facteurs ont également conditionné l'étude, comme notamment : le sexe, l'âge, le lieu de naissance, le lieu de résidence, le statut marital, la présence ou non d'un conjoint et / ou d'enfants, la profession, l'ancienneté au sein de la profession et l'exonération de cotisation.

De plus, les informations ont été collectées à partir d'avis d'arrêts de travail et de données historiques issues de la base du RSI.

Concernant la modélisation des arrêts de travail et en particulier la censure, le modèle de Cox a été utilisé. En effet, les arrêts de travail ne débutant pas nécessairement à la date fixée du début d'observation, et ne se finissant pas à la date fixée de fin d'observation, il se peut que certains d'entre eux soient non-observés ou observés mais seulement partiellement, comme il est possible de le voir sur le schéma ci-dessous.

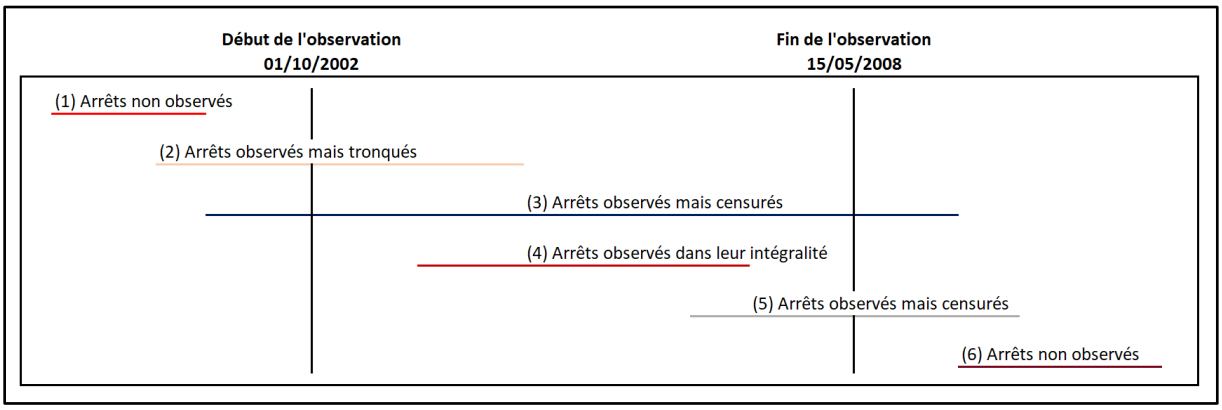


Figure 28 - Diagramme présentant les différentes censures

L'étude ainsi menée a permis d'aboutir aux résultats présentés dans le tableau suivant :

Age	i_x
18-39	7,59%
40-49	7,91%
50-60	9,45%

Figure 29 - Loi d'incidence en arrêt de travail pour la population de référence

A noter qu'une étude plus précise permettant de distinguer le taux d'incidence en fonction du secteur d'activité avait également été réalisée. Cependant, la répartition n'étant pas similaire et la quantité de données pour certaines classes d'activité n'étant pas suffisante, il a été décidé d'utiliser la loi d'incidence globale.

Pour toute information complémentaire au sujet de la loi d'incidence utilisée, le lecteur est invité à se référer à l'article suivant [4].

Comme cela a été rappelé, l'étude proposée plus haut s'arrête au cours de l'année 2008. Dans un souci de validité des informations recueillies, une analyse succincte de l'évolution de la structure de la population française suivant les classes d'âge relatives à la loi d'incidence a été menée. Celle-ci a été effectuée en s'aidant de données issues de l'INSEE et de l'INED.

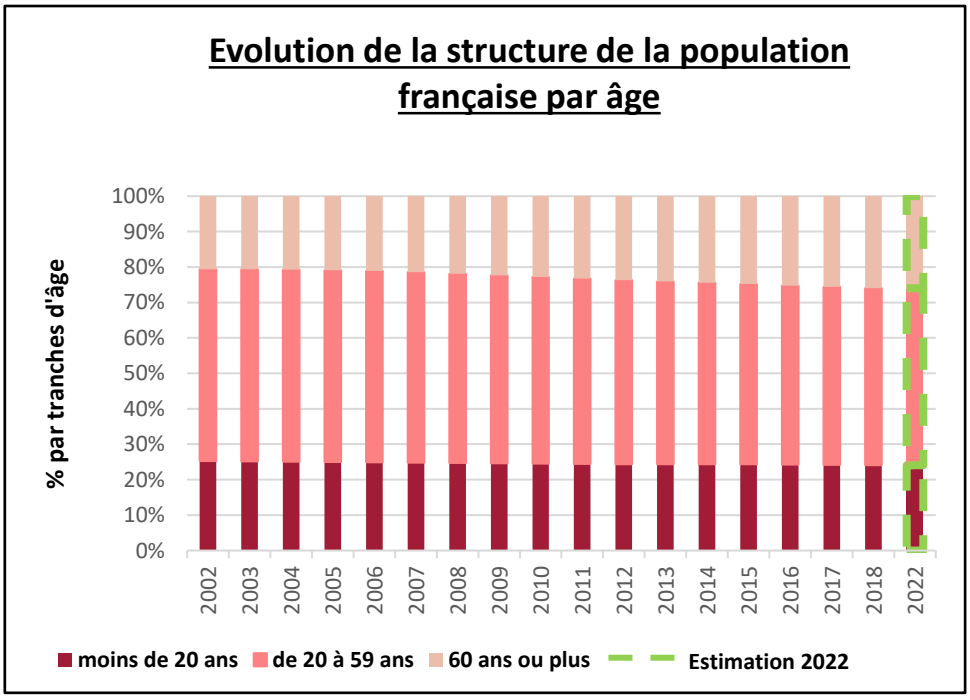


Figure 30 - Evolution de la structure de la population française par âge

A première vue, aucune déformation significative des cohortes d'âge n'est à signaler. La répartition par âge de la population est relativement stable depuis 2008, avec cependant une légère hausse de la population « des plus de 60 ans », en lien avec la hausse de l'espérance de vie et le poids démographique des baby-boomers.
Cela ne remet pas en cause l'étude et l'utilisation de la loi d'incidence obtenue.
Le ratio de morbidité peut à présent être appliqué.

1.2 Applications et résultats

L'objectif de cette partie est de vérifier l'efficacité de la sélection médicale au sein du portefeuille SLPI grâce au ratio de morbidité.

Dans ce cas, la base d'étude devient alors la base SLPI, qui, pour rappel, contient 107 644 assurés distincts.

La base de référence quant à elle est la population française, représentée par l'intermédiaire de la loi d'incidence obtenue précédemment. Il s'agit donc en réalité, dans un souci de cohérence avec la base SLPI, de la population française limitée aux indépendants.

Un raisonnement par tranches d'âge, à savoir [18 ans ; 39 ans], [40 ans ; 49 ans] et [50 ans ; 60 ans] a donc été mené afin d'obtenir le nombre d'arrêts de travail attendu si la sinistralité était similaire à celle de la population de référence selon ces trois classes d'âge.

Dans ce but, pour chacune de ces trois classes, le nombre d'assurés potentiellement capables de tomber en arrêt de travail a été relevé et la loi d'incidence lui a été appliquée.

Afin de mesurer l'efficacité de la sélection médicale, le ratio de sinistralité a également été ventilé par ancienneté. L'ancienneté maximale ayant été fixée à douze ans.

Voici les résultats obtenus :

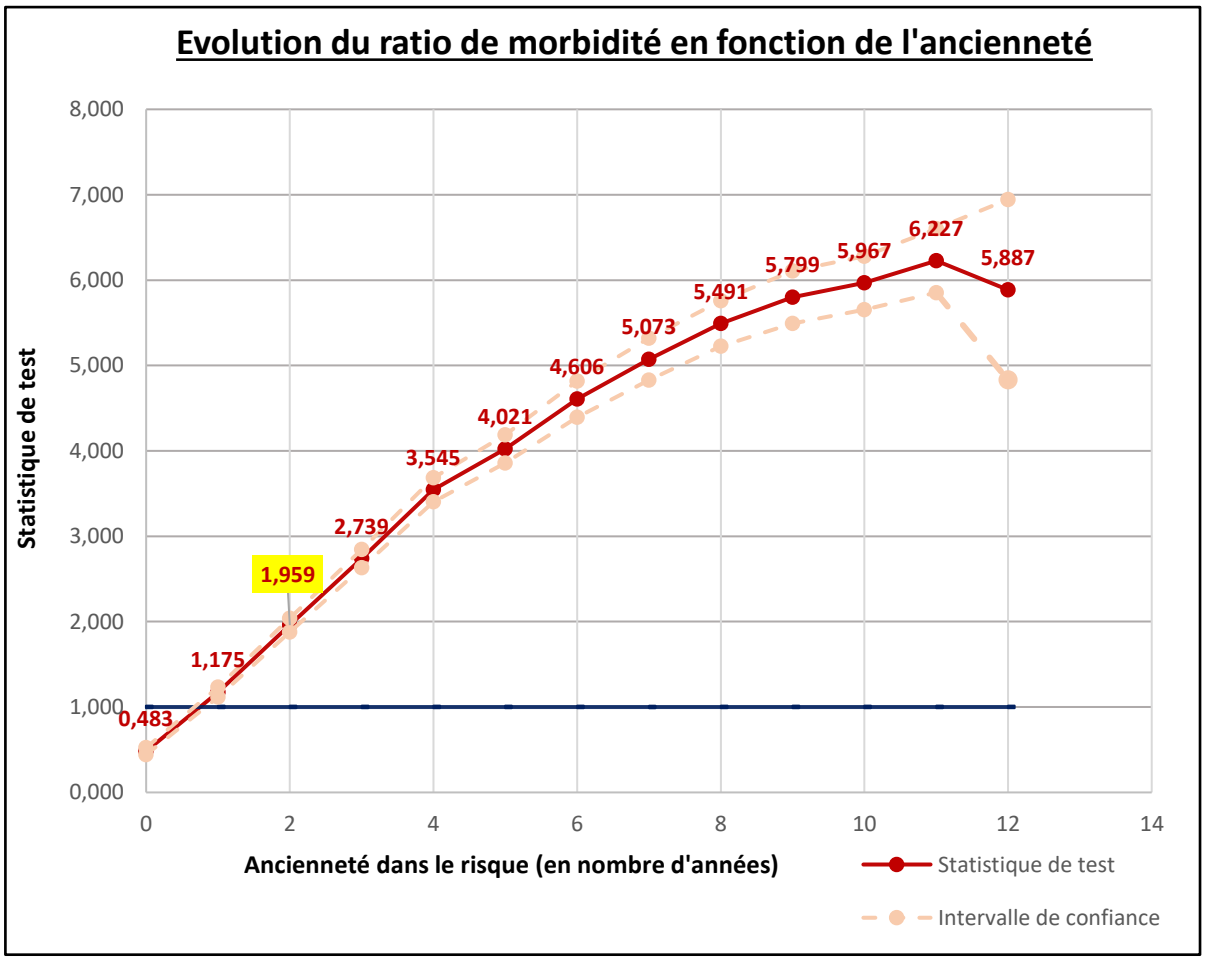


Figure 31 - Effet de la sélection médicale au cours du temps

Grâce à la représentation graphique de la statistique de test en fonction de l'ancienneté, il est possible de remarquer plusieurs éléments.

Premièrement, jusqu'à un peu moins d'un an d'ancienneté dans le portefeuille SLPI, la valeur de la statistique de test est inférieure à 1. Cela signifie que la sélection médicale a bien un effet sur la sinistralité. Elle permet donc de se ramener à une population présentant une sinistralité plus faible que la population générale des TNS.

Cependant, il est possible de souligner que dès lors que l'ancienneté dépasse une année, la sinistralité au sein du portefeuille SwissLife devient plus élevée que celle de la population de référence.

Non seulement la sinistralité est plus importante au bout d'un an d'ancienneté, mais il est aussi possible d'observer qu'à partir de deux ans la sinistralité de la population d'étude devient même deux fois plus élevée.
Ainsi, un « seuil critique » de deux années d'ancienneté pourrait être fixé. Cependant, estimer que la sélection médicale n'a plus d'effet dès lors que la sinistralité de la population d'étude dépasse la sinistralité de la population de référence, et donc choisir comme valeur de référence 1 pour la statistique de test est peut-être un peu osé et radical.
En effet, bien que la sinistralité soit plus importante au-delà de la valeur 1, cela ne signifie pas pour autant que la sélection médicale n'a plus aucun effet. Cet effet, tout en s'amoindrissant au cours de l'ancienneté dans le risque reste toujours perceptible.
D'autre part, il est important de garder en tête que les résultats obtenus ici dépendent fortement des bases de référence et d'étude utilisées. D'autres bases de données auraient probablement conduit à des résultats différents.

Sur ce même graphique, ont également été ajoutés en pointillés les intervalles de confiance. Il est notamment possible d'observer que plus l'ancienneté grandit, plus la volatilité autour de la statistique de test est grande. Cela s'explique tout simplement par le fait qu'un assuré reste en moyenne quatre années au sein du portefeuille SwissLife. Par conséquent, plus l'ancienneté est grande, moins le nombre d'assurés présents avec une telle ancienneté sera grand. Les statistiques sont alors réalisées sur un nombre plus faible de données rendant les résultats plus volatiles.

Toutefois, cela a peu d'impact sur les premiers résultats observés puisque ces derniers se réfèrent à une ancienneté dans le risque inférieure à l'ancienneté moyenne.

Cette première approche relativement globale a été complétée dans un second temps par une approche permettant de comparer le ratio de sinistralité en fonction du sexe.

Bien qu'une différenciation homme / femme sur le tarif ne puisse être appliquée, il est toujours intéressant d'avoir une idée du comportement du portefeuille notamment dans un but de provisionnement.

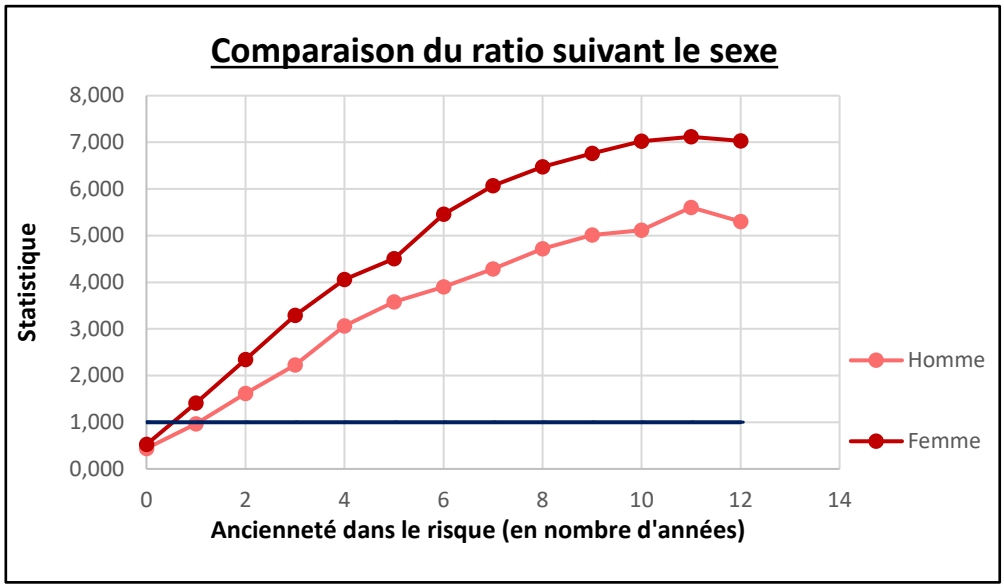


Figure 32 - Evolution du ratio de sinistralité - Zoom Homme / Femme

Tout comme dans le premier graphique qui ne faisait pas la distinction homme / femme, il est possible de remarquer les mêmes tendances en termes de sinistralité. Dans les deux cas, jusqu'à un peu moins d'un an la sinistralité du portefeuille SLPI est inférieure à la sinistralité de la population de référence.

Néanmoins, des dissimilitudes existent entre les assurés de sexe masculin et ceux de sexe féminin.

En effet, alors qu'au bout d'une année d'ancienneté la sinistralité relative aux hommes est encore égale voire inférieure à celle de la population de référence, la sinistralité relative aux femmes est quant à elle déjà supérieure à la population générale, ce phénomène se répétant à mesure que l'ancienneté croît.
Une explication possible à cette sur-sinistralité des femmes par rapport aux hommes pourrait peut-être s'expliquer par la prise en compte des congés maternité dans les arrêts de travail, venant ainsi augmenter la sinistralité.

Au sein de la deuxième partie de ce mémoire dans le cadre des statistiques descriptives, il avait été souligné que certaines catégories professionnelles présentaient une sinistralité plus importante que d'autres. Parmi elles, avaient été mises en lumière les professions libérales médicales et les professions paramédicales.

Dans le but d'étayer l'étude sur l'efficacité de la sélection médicale, le ratio de sinistralité a donc été appliqué en faisant la distinction entre les différentes classes d'activités, comme le révèle le graphique ci-après.

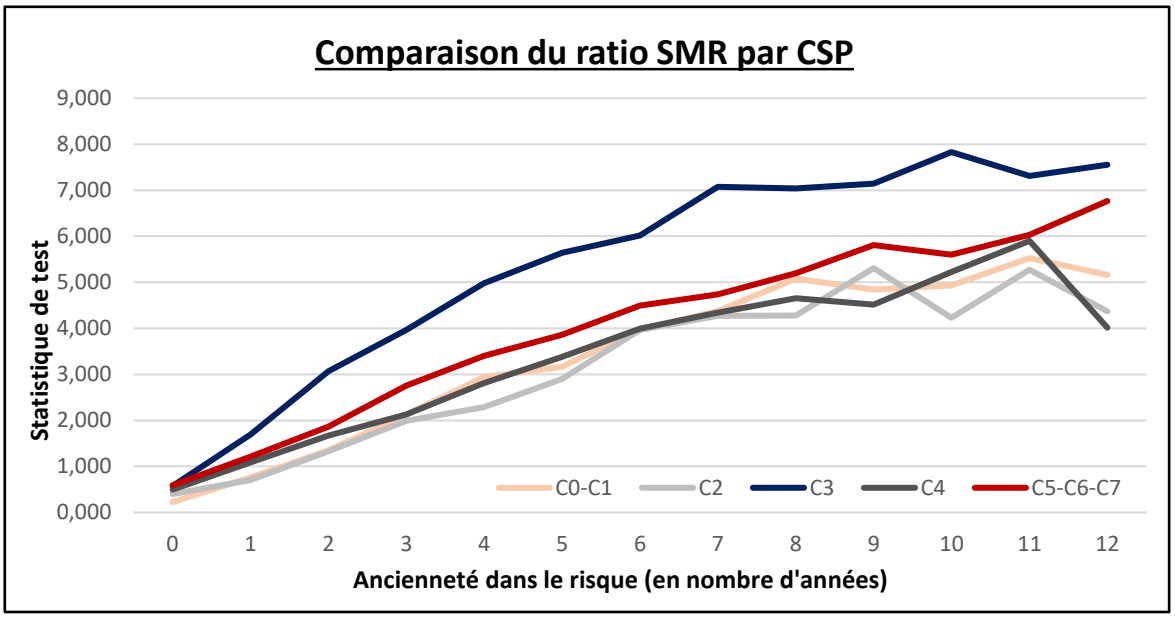


Figure 33 - Effet de la sélection médicale - Zoom par CSP

Les remarques mentionnées lors des deux autres approches restent toujours valables. Peu importe la classe professionnelle représentée, la statistique de test présente une valeur inférieure à 1 au cours de la première année d'ancienneté avant de croître plus ou moins fortement selon les professions.

En effet, il est possible de remarquer que certaines professions se détachent des autres. Parmi elles, et sans grande surprise, peuvent être mentionnées les professions paramédicales (classe C3) telles que les infirmières. Les autres professions semblent davantage converger vers les mêmes ordres de grandeur en termes de sinistralité, même s'il est possible de souligner que les artisans et exploitants agricoles présentent une sinistralité légèrement plus importante que les trois autres classes au sein de ce groupe.

D'un point de vue de la sélection médicale, ce graphique révèle que l'efficacité de cette dernière diffère en fonction des professions ciblées. Toutes les professions ne sont pas soumises équitablement au risque d'un arrêt de travail.

1.3 Conclusion

En conclusion de cette partie, plusieurs messages méritent d'être retenus.
Le premier, et celui qui était l'objectif de cette partie, est que la sélection médicale a bien un réel effet, et qui plus est, visible sur la sinistralité de la population du produit SLPI. Son impact est significatif et permet de ramener la population choisie à une sinistralité plus faible que celle de référence.

Le deuxième message, qui découle tout naturellement du premier, est que l'efficacité de cette dernière est limitée dans le temps. Plusieurs observations ont montré que ce phénomène s'atténuait au cours du temps avec l'ancienneté, un seuil autour de deux années pouvant être fixé.

Le troisième message à retenir est que l'impact de la sélection médicale n'est pas équitablement réparti entre les professions. Certaines professions et notamment les professions paramédicales présentent une sinistralité bien supérieure aux autres.

Ces conclusions étant tirées, une nouvelle question émerge alors : comment prendre en compte l'atténuation de l'effet de la sélection médicale dans le temps, et par conséquent une potentielle dérive de la sinistralité ?

2 Revalorisation grâce aux GLM

Comme évoqué précédemment, la sélection médicale a certes un effet sur la sinistralité, mais ce dernier semble disparaître au cours du temps avec l'ancienneté.

Par ailleurs, lorsque qu'une dérive de sinistralité est observée dans un portefeuille assurantiel, un des leviers possibles afin de contrer celle-ci est de mettre en place une revalorisation à la hausse des tarifs d'assurance. Toutefois, cette revalorisation ne peut pas être appliquée de but en blanc et arbitrairement. Pour être la plus juste possible elle se doit de dépendre de différentes variables.

L'objectif de cette partie est donc de présenter une solution à ce constat. Pour cela, il a été décidé de modéliser une revalorisation au travers d'un GLM selon différents critères qui seront exposés par la suite.

2.1 Modélisations et résultats

Cette sous-partie a pour but d'expliquer les résultats de modélisation de la revalorisation mise en place pour le produit SLPI.

Pour rappel, il avait été observé que la sélection médicale permettait bien de réduire la sinistralité du portefeuille SLPI, mais que cet effet s'atténuait au cours du temps. Une revalorisation a donc été mise en place selon différents critères.

Le premier critère, et le plus évident à la vue des résultats obtenus, est l'ancienneté dans le portefeuille.
Cette variable « ancienneté » a par la suite été découpée en quatre classes distinctes : **[0-3], [4-7], [8-9]** et **plus de 10 ans**, l'unité choisie étant bien le nombre d'années d'ancienneté.

Ce découpage s'appuie notamment sur le graphique suivant.

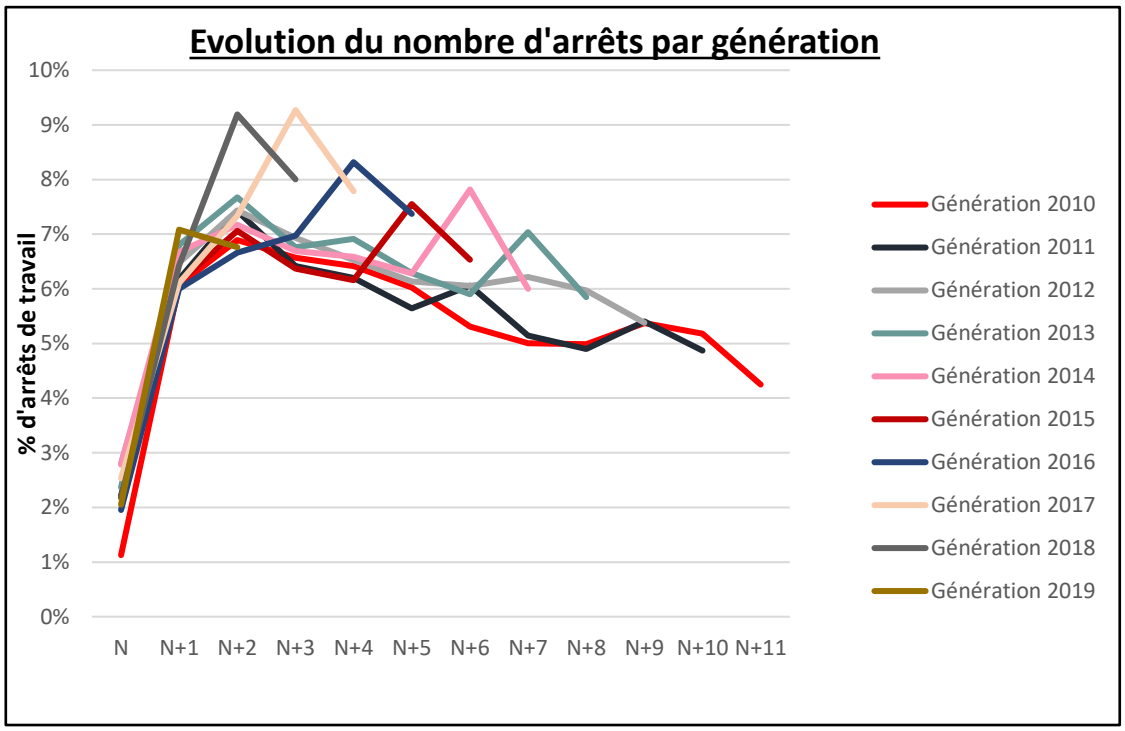


Figure 34 - Evolution du nombre d'arrêts de travail par génération depuis le lancement du produit SLPI

Ce dernier permet de mettre en lumière un pic de sinistralité au bout de la troisième année d'ancienneté, justifiant ainsi la première classe de découpage.

Pour information, les générations 2020, 2021 et 2022 ne s'étant pas encore suffisamment déroulées au cours du temps, il a été choisi de ne pas les ajouter au graphique.

De plus, le saut « atypique » qu'il est possible de repérer pour toutes les générations lors de la dernière année correspond bien aux conséquences de la Covid en 2020.

Comme cela avait été également souligné, il s'avère que toutes les catégories professionnelles ne sont pas soumises à la même sinistralité. Certaines professions telles que les professions paramédicales présentaient une sinistralité plus importante.
Par conséquent, la seconde variable explicative choisie est la « catégorie socio professionnelle », découpée en huit classes : **C0, C1, C2, C3, C4, C5, C6** et **C7.**

Afin d'appuyer ces découpages, un arbre CART (Classification And Regression Tree) a été lancé sur R.

Le choix du paramétrage ne sera pas détaillé ici, l'arbre obtenu permettant d'obtenir simplement une répartition en classes à titre indicatif.

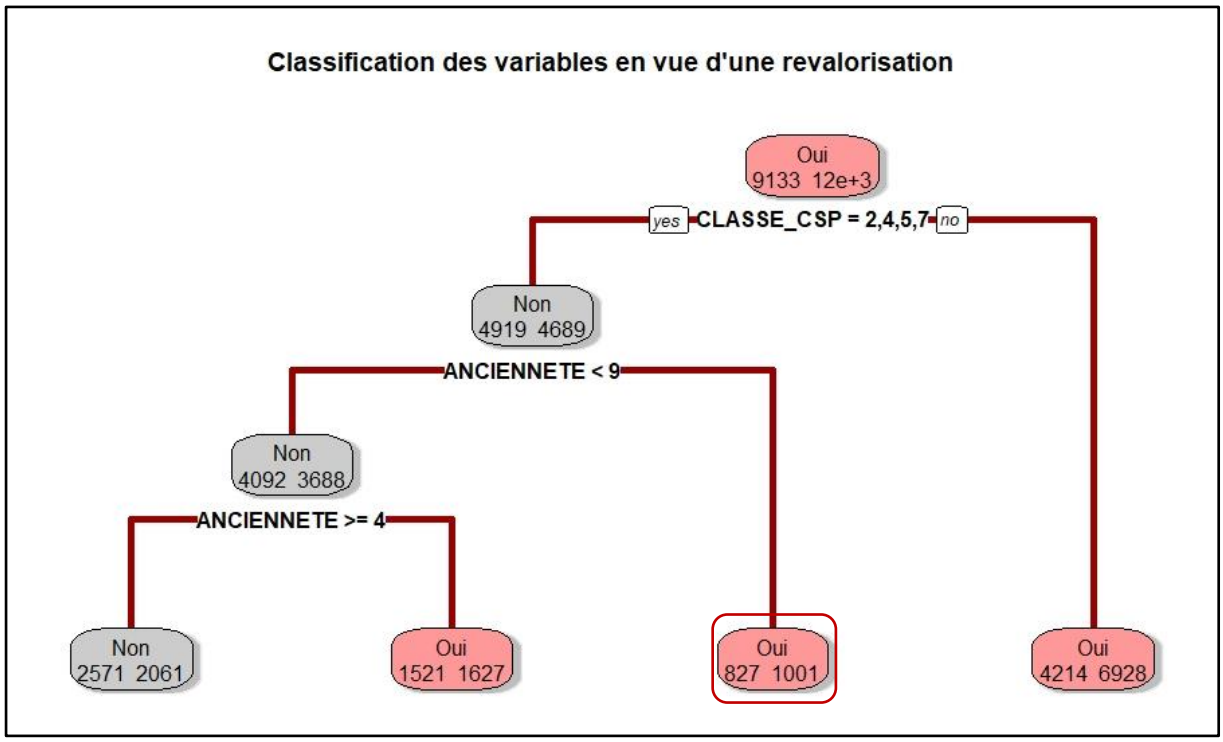


Figure 35 - Arbre CART - Classification des variables pour la revalorisation

Avant d'effectuer l'analyse de la sortie obtenue, il convient de faire un bref rappel sur la manière d'analyser les différents nœuds de l'arbre.
Au niveau de chaque nœud représenté par un rectangle, deux nombres se font face. Le nombre de gauche représente le nombre d'individus dans la classe n'ayant pas fait l'objet d'une revalorisation, et le nombre de droite représente le nombre d'individus dans la classe ayant bénéficié d'une revalorisation. Le nombre le plus grand associera par la suite la classe à sa caractéristique : revalorisation : oui ou non, caractéristique mentionnée en haut du rectangle.

Pour illustrer ceci, l'individu sélectionné en rouge servira d'exemple. Il s'agit d'un individu dont la classe CSP est parmi les classes 2,4,5 ou 7. Ce même individu présente une ancienneté dans le risque supérieure ou égale à 9 années. Enfin, il s'agit d'un individu pour lequel une revalorisation sera appliquée.

A première vue, bien que ne présentant pas exactement les mêmes classes que celles choisies manuellement, l'algorithme semble appuyer les choix effectués. En effet, celui-ci semble tout d'abord dissocier les classes C0, C1, C3 et C6.

Ce choix est quasiment trivial pour la classe C3. Concernant les classes C0 et C1 qui correspondent aux professions libérales médicales, il avait été montré lors de la partie sur les caractéristiques du produit SLPI que ces dernières présentaient une sinistralité plus élevée que les autres classes, notamment à cause des salaires importants des personnes assurées.

La classe C6 correspond quant à elle aux exploitants agricoles. De prime abord, aucune raison ne semble expliquer ce choix. Il avait cependant été démontré que l'effet de la sélection médicale était moindre sur le regroupement des classes C5, C6 et C7. Il se pourrait donc que ce soit la classe C6 qui entraine une sinistralité plus forte au sein de ce groupe.

Concernant la variable « ancienneté », l'algorithme semble renforcer le choix du découpage au-delà de trois années d'ancienneté et au-dessus de dix années. Les découpages intermédiaires ne sont quant à eux pas visibles. Cela n'est toutefois pas alertant puisque le découpage choisi est simplement un découpage plus fin.

La dernière variable venant compléter les autres variables explicatives est une variable représentant la sinistralité.
En effet, comme il est possible de le voir sur le graphique ci-dessous, à mesure qu'une génération se développe, le ratio de sinistralité associé ne cesse d'augmenter. Cette observation souligne le fait que la dérive de sinistralité est une donnée à ne pas négliger lors de la revalorisation.
Cette variable de sinistralité a quant à elle été découpée en cinq classes distinctes que voici :

- **[0%,50%]**
- **[51%,63%]**
- **[64%,73%]**
- **[74%,83%]**
- $>$ **83%**

Pour information, ce découpage s'appuie sur des conventions propres à SwissLife.

De plus, comme précédemment, pour des raisons de confidentialité, les valeurs de l'axe des ordonnées de la Figure 36 ont été masquées.

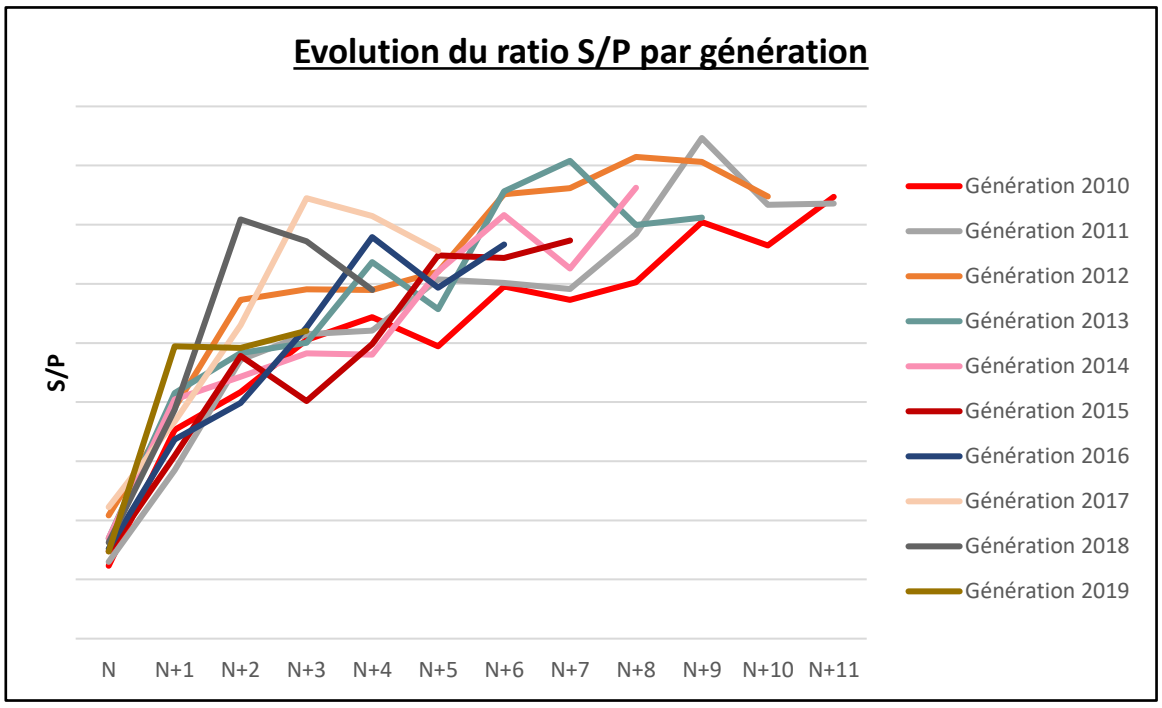


Figure 36 - Evolution du ratio S/P par génération

En résumé, la revalorisation proposée s'appuie donc sur les trois variables explicatives suivantes :

- La **catégorie socio professionnelle** découpée en huit classes : C0, C1, C2, C3, C4, C5, C6 et C7
- L'**ancienneté** découpée en quatre classes : [0-3], [4-7], [8-9] et +10 ans
- La **sinistralité** découpée en cinq classes : [0%-50%], [51%-63%], [64%-73%], [74%-83%], >83%

A présent, l'objectif de la suite de cette partie est de présenter les résultats des différents GLM appliqués à la revalorisation.

Découpage de la base de données

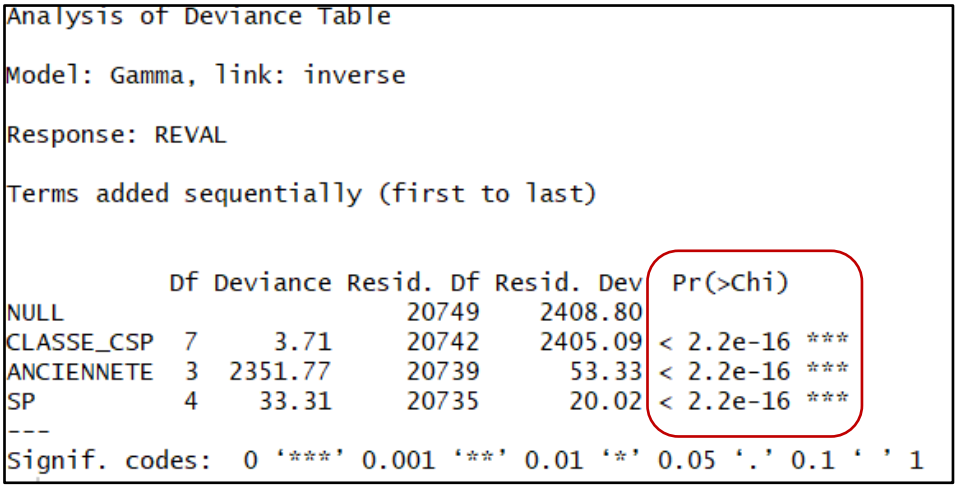
Avant de se lancer intégralement dans la mise en place de modèles, la base de données initiale a été séparée en deux bases distinctes : une base d'apprentissage regroupant 75% des données et une seconde base, appelée base de test regroupant quant à elle 25% des données.
La vocation de ce découpage est de permettre la validation des modèles.

Significativité des variables

Bien qu'à première vue toutes les variables semblent contribuer à l'attribution de la valeur de la revalorisation, il a tout de même été choisi de lancer un test de significativité concernant les variables pour s'en assurer.
Afin d'y parvenir, un premier GLM a donc dû être lancé. Le choix s'est alors porté sur le GLM Gamma.
En effet, le montant de revalorisation étant un montant positif ici, le choix de la loi s'est alors tourné vers la loi Gamma. Toutefois, comme cela sera démontré par ailleurs, ce choix n'est pas un choix définitif. D'autres GLM seront également testés sur les données.
D'autre part, il est vrai qu'une réduction tarifaire via une revalorisation négative aurait pu être envisagée, ce n'est cependant pas le choix retenu ici, d'où l'hypothèse d'un montant positif.

Concernant le choix de la fonction de lien, là encore cette dernière a été choisie de manière relativement classique, puisqu'il s'agit tout simplement de la fonction de lien canonique, à savoir la fonction inverse.

Les résultats de l'analyse ANOVA avec ce GLM permettent de confirmer la significativité des variables explicatives du modèle.



```
Analysis of Deviance Table

Model: Gamma, link: inverse

Response: REVAL

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       20749    2408.80
CLASSE_CSP  7     3.71      20742    2405.09 < 2.2e-16 ***
ANCIENNETE  3  2351.77      20739      53.33 < 2.2e-16 ***
SP          4    33.31      20735      20.02 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 37 - Sortie Anova - GLM Gamma

Toutes les variables explicatives présentent une p-value dont la valeur est inférieure à 2.2e-16. Il est donc possible de rejeter l'hypothèse nulle qui suppose que tous les coefficients associés à ces variables sont nuls. Toutes les variables explicatives seront donc conservées dans le modèle.

Choix des GLM

Dans un second temps, différents modèles ont été appliqués aux données. Afin de déterminer celui qui sera le plus adapté à ces dernières, un algorithme de cross-validation a été appliqué.

Bien qu'il ne s'agisse pas ici de redévelopper l'entière théorie de la cross-validation, voici quelques outils afin de comprendre son fonctionnement.

Tout d'abord, l'objectif est de permettre une estimation de l'erreur de généralisation d'un modèle, dans le but de comparer plusieurs modèles. L'idée est alors de découper la base d'apprentissage en K blocs. Pour chaque K bloc, l'ajustement est réalisé avec les $(K-1)$ blocs restants servant de base d'entrainement, alors que le dernier bloc restant servira quant à lui de test.

Voici le principe sous forme de schéma.

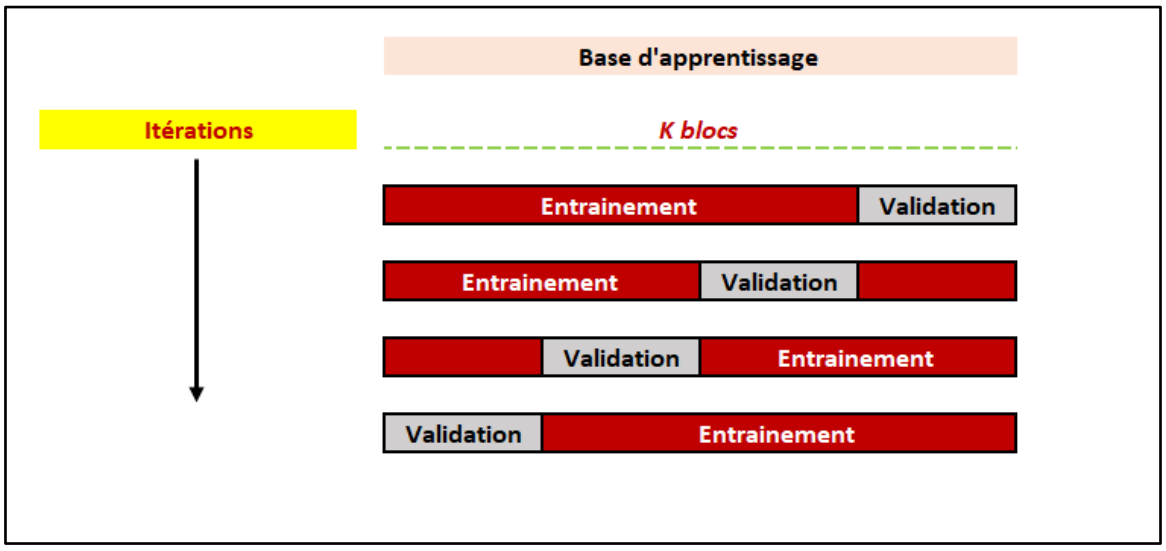


Figure 38 - Fonctionnement de la Cross-Validation

Dans le cas présenté ci-dessous, le choix du nombre de blocs sera porté à cinq. De manière générale, dans la littérature, il est commun de voir le paramètre K prendre une valeur égale à cinq ou à dix.

Pour information, si le paramètre K est choisi comme étant égal au nombre d'individus dans la base de données, alors ce type de validation est appelée « Leave-one-out cross-validation » (LOO), signifiant « validation croisée d'un contre tous ». L'entrainement est alors réalisé sur $n-1$ observations et la validation sur l'observation restante.
Toutefois, ce type de cross-validation est généralement très couteux opérationnellement.

Toujours dans le but de modéliser la revalorisation de la prime, elle-même mise en place dans l'objectif de prendre en compte les dérives de sinistralité potentielles liées à la diminution de l'effet de la sélection médicale, il a été décidé de regarder en détail cinq GLM différents.

Comme énoncé plus haut lors du choix des variables significatives, le premier GLM concerné est le **GLM Gamma** muni de la fonction de lien canonique à savoir la **fonction inverse**.

Le second choix de GLM s'est porté encore une fois vers le **GLM Gamma**, mais en modifiant la fonction de lien. Il s'agira cette fois-ci de la fonction de lien **logarithmique**.

Dans un troisième temps, un **GLM Gaussien** a aussi été sélectionné, associé à la fonction de lien **identité**. En effet, bien que la distribution d'une loi normale soit souvent imaginée centrée en zéro, il se pourrait qu'en jouant sur les paramètres de moyenne et d'écart-type, la distribution puisse être ramenée dans la partie positive du repère conduisant à une masse quasi nulle dans la partie négative.

La distribution **inverse gaussienne** fait également partie de la sélection pour affiner le choix du GLM. Distribution généralement utilisée dans le cas du mouvement brownien, cette dernière a la caractéristique d'être continue et à valeurs uniquement strictement positives. Pour information, la fonction de lien canonique associée est définie par :

$$g(\mu) = \frac{1}{\mu^2}$$

Le dernier GLM venant compléter la liste de ceux déjà mentionnés est le GLM Tweedie.
Les distributions Tweedie sont des distributions peu connues qui appartiennent à la classe des distributions de dispersion exponentielle qui sont caractérisées par une variance de la forme suivante :

$$Var(Y) = \phi\mu^k$$

où μ est la moyenne de la distribution, ϕ le paramètre de dispersion et k l'indice de la distribution.

Voici quelques exemples de distributions connues sous la forme d'une distribution Tweedie :

Distribution Tweedie	Distribution associée
$k = 0$	Loi Normale
$k = 1$	Loi de Poisson
$k = 2$	Loi Gamma
$k = 3$	Loi Inverse Gaussienne

Figure 39 - Exemples de distributions Tweedie

Par ailleurs, la distribution Tweedie a la particularité d'attribuer une masse importante en zéro aux observations.

Voici la sortie R des résultats obtenus par cross-validation.

	model	`mean(err)`
	<chr>	<dbl>
1	GLM Gamma - link = inverse	7.45
2	GLM Gamma - link = log	3.97
3	GLM Gaussien	0.00000558
4	GLM Inverse gaussienne	5.76
5	GLM Tweedie	0.00000560

Figure 40 - Résultats Cross-Validation - Choix GLM

La colonne droite de la sortie R représente l'estimation de l'erreur obtenue par cross-validation, associée au modèle correspondant mentionné quant à lui dans la colonne gauche.

A première vue, deux GLM semblent se différencier des autres. Il s'agit du **GLM Gaussien** et du **GLM Tweedie**. Le GLM Gamma qui avait pourtant été intuitivement choisi dès le début de l'étude ne semble donc pas être celui représentant au mieux les données.

Dans le but de compléter davantage le choix du GLM, un nouvel algorithme de cross-validation a été lancé permettant de comparer l'erreur obtenue pour le GLM Gaussien en fonction des fonctions de lien suivantes : identité, inverse et logarithmique.
Les résultats de la sortie R obtenue sont représentés dans le tableau ci-après.

	model	`mean(err)`
	<chr>	<dbl>
1	GLM Gaussien - link = identité	0.00000558
2	GLM Gaussien - link = inverse	7.45
3	GLM Gaussien - link = log	3.97
4	GLM Tweedie	0.00000560

Figure 41 - Sortie R Cross-Validation - Zoom GLM Gaussien & Tweedie

Malgré différents tests concernant les fonctions de lien, ce sont bien le GLM Gaussien muni de la fonction de lien identité et le GLM Tweedie qui se révèlent être les plus probants.
Un choix plus affiné entre ces deux derniers candidats seraient maintenant compliqué à justifier, c'est la raison pour laquelle les études menées par la suite seront appliquées aux deux cas : Gaussien et Tweedie.
La seule raison pouvant orienter le choix du lecteur vers l'un de ces GLM, est que le GLM Tweedie est couramment utilisé lors d'études actuarielles.
Ci-contre le lecteur pourra trouver les sorties Anova pour chacun des deux GLM mentionnés.

```
Analysis of Deviance Table

Model: gaussian, link: identity

Response: REVAL

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       15561    15297.2
CLASSE_CSP  7     28.7     15554    15268.5 < 2.2e-16 ***
ANCIENNETE  3  15051.4     15551      217.1 < 2.2e-16 ***
SP          4    217.0     15547        0.1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 42 - Sortie Anova pour le GLM Gaussien

```
Analysis of Deviance Table

Model: Tweedie, link: mu^1

Response: REVAL

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       15561     5153.4
CLASSE_CSP  7     10.0     15554     5143.4 < 2.2e-16 ***
ANCIENNETE  3   5055.5     15551       88.0 < 2.2e-16 ***
SP          4     87.9     15547        0.0 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 43 - Sortie Anova pour le GLM Tweedie

Analyse des résidus

L'analyse des résidus de la déviance amène aux graphiques suivants :

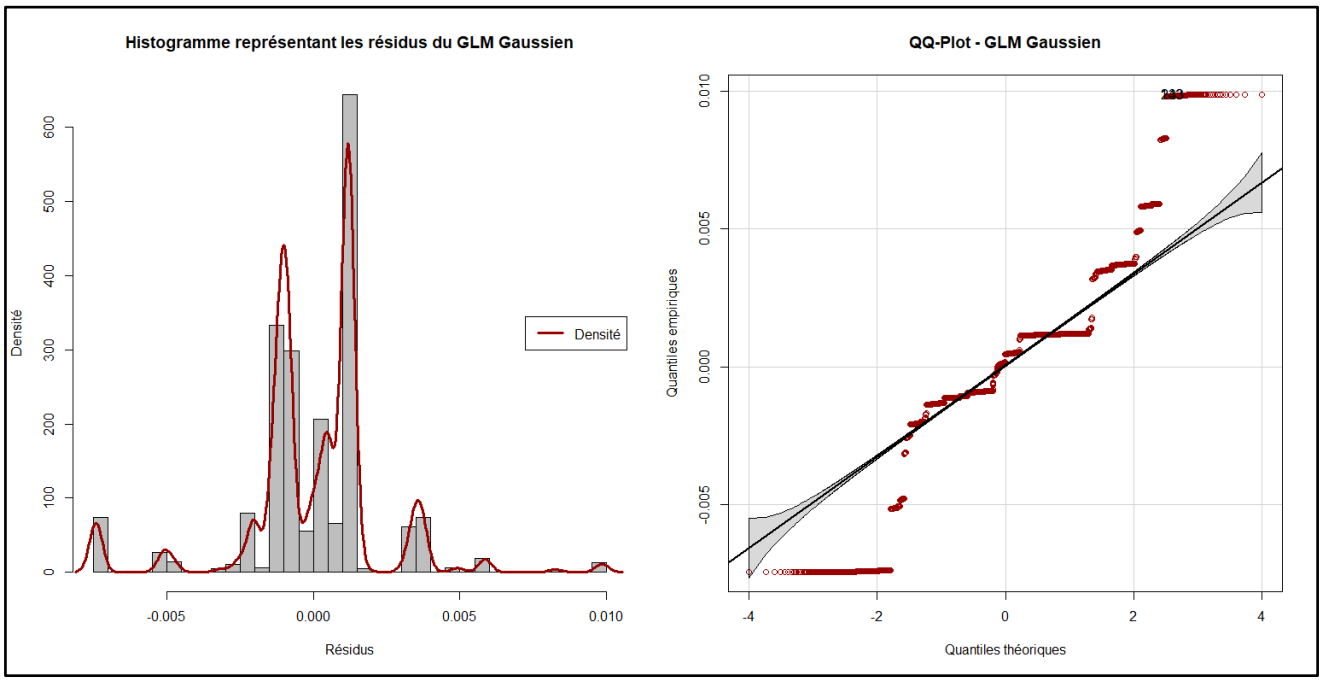


Figure 44 - Analyse des résidus - GLM Gaussien

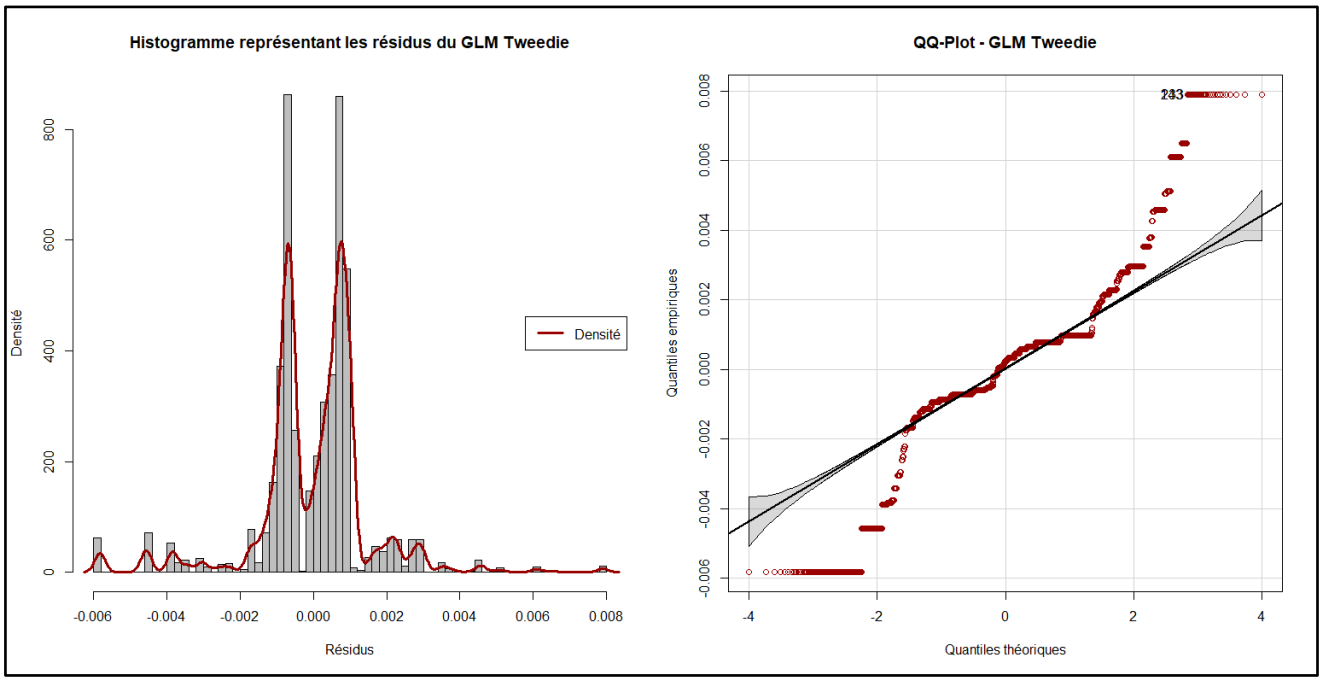


Figure 45 - Analyse des résidus - GLM Tweedie

Que ce soit pour le GLM Gaussien ou pour le GLM Tweedie, la première remarque à mentionner est la suivante : dans les deux cas, les résidus apparaissent relativement centrés autour de la moyenne nulle.

La seconde observation qui découle quant à elle des QQ-Plots est que les résidus ne sont pas bien alignés selon la première bissectrice, notamment en queue de distribution. Ces derniers ne semblent pas être gaussiens, ce qui ne remet toutefois pas en cause la pertinence du modèle, puisqu'il ne s'agit point d'une des hypothèses des GLM.

Dans le but de confirmer ou d'infirmer les observations décrites ci-dessus, un test a été lancé.

Ce test est connu sous le nom de test de Student. Il permet notamment de comparer la moyenne de deux groupes d'échantillons indépendants via l'hypothèse nulle suivante :

$$H_0 : \textit{La moyenne est égale à 0}$$

```
        One Sample t-test

data:  res1
t = 8.4726e-09, df = 15561, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -3.720139e-05  3.720139e-05
sample estimates:
   mean of x
1.608034e-13
```

Figure 46 - Test de Student - GLM Gaussien

```
        One Sample t-test

data:  res2
t = -0.028432, df = 15561, p-value = 0.9773
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.482862e-05  2.411864e-05
sample estimates:
    mean of x
-3.549934e-07
```

Figure 47 - Test de Student - GLM Tweedie

Dans les deux cas, la valeur de la p-value obtenue permet de ne pas rejeter l'hypothèse H_0 au seuil $\alpha = 5\%$. La moyenne des résidus peut alors être considérée comme nulle.

Pour compléter l'étude portant sur les résidus de la déviance, les résidus studentisés ont également été étudiés.

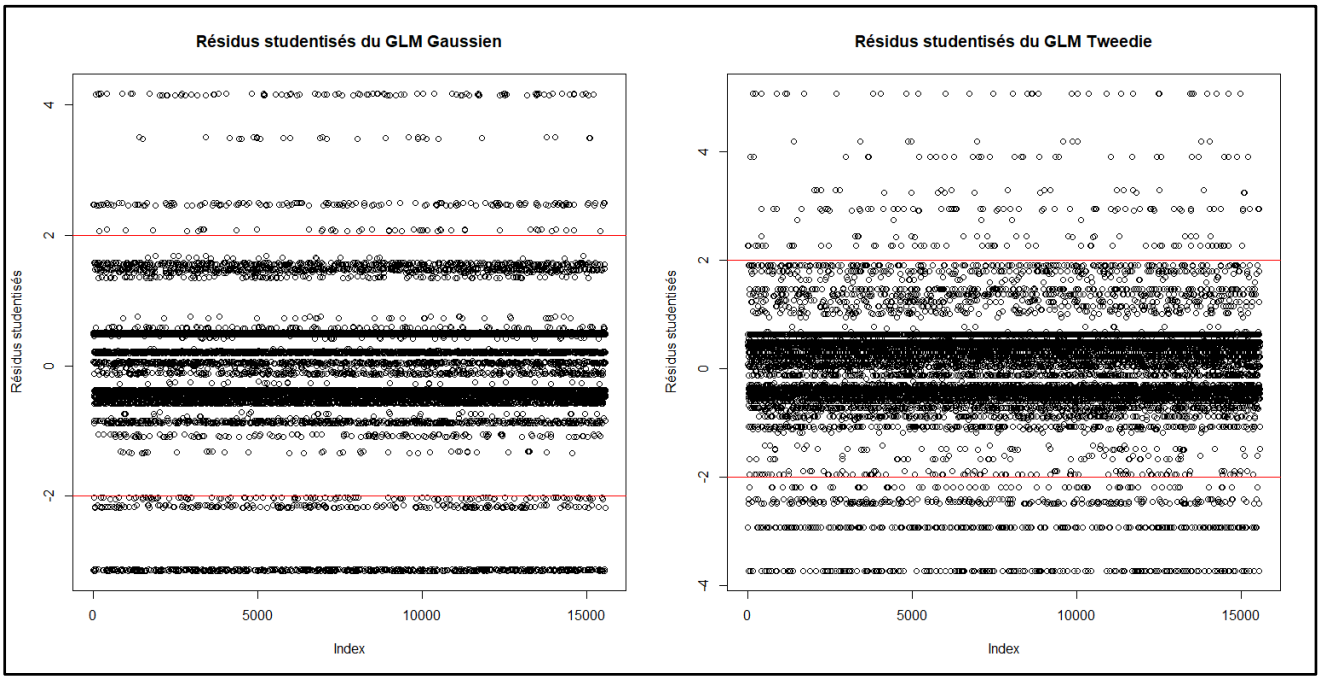


Figure 48 - Résidus Studentisés - GLM Gaussien & Tweedie

Les deux graphiques ci-dessus permettent de représenter les résidus studentisés associés aux deux GLM et de mettre en lumière une éventuelle alerte sur le comportement de ces derniers. L'examen des résidus studentisés ne doit révéler aucune forme particulière et les valeurs prises par ceux-là doivent être comprises entre -2 et 2.

A première vue, tous les résidus studentisés ne se situent pas dans l'intervalle $[-2;2]$. Cependant, cette proportion reste relativement faible dans le cas des deux GLM comme le montre le tableau suivant.

Type de GLM	Proportion exclue de [-2,2]
Gaussien	7,79%
Tweedie	6,14%

Figure 49 - Analyse des résidus studentisés

Concernant la présence ou non d'une éventuelle forme, il est possible de dire qu'aucune forme particulière n'est à relever. En effet, bien qu'il soit possible d'observer des lignes horizontales, celles-ci résultent davantage d'une concentration des points que d'une dépendance éventuelle. Par ailleurs, il suffit de regarder les valeurs des ordonnées selon chaque graduation des abscisses pour se rendre compte que les résidus sont uniformément répartis.

Coefficient de détermination

Dans le cadre des modèles linéaires généralisés, le coefficient de détermination peut être analysé via le $pseudo\ R^2$. Ce dernier permet de donner une information relative à la qualité du modèle. Dans les deux cas des GLM présentés ici, 99% de la variance des données est expliquée par le modèle, ce qui est une valeur plus que correcte.

2.2 Conclusion

En conclusion, pour remédier à l'effet de la sélection médicale s'atténuant au cours de l'ancienneté dans le risque, il est possible de mettre en place une revalorisation. Cette revalorisation permettra de contrer de potentielles dérives de sinistralité. Elle sera également conditionnée par différents facteurs tels que l'ancienneté, la profession ou encore la sinistralité comme cela a pu être démontré grâce aux modèles linéaires généralisés.
Bien que ces derniers ne soient pas les uniques modèles qui auraient pu expliquer la revalorisation mise en place pour le produit SLPI, ils ont l'avantage d'être simples à utiliser et de permettre une certaine flexibilité quant aux choix des distributions appartenant aux familles exponentielles et aux fonctions de lien.
Une limite de ces modèles étant cependant le fait qu'ils ne permettent pas de modéliser des valeurs extrêmes, et que, bien qu'un choix concernant la fonction de lien soit possible, celui-ci est généralement orienté vers la fonction de lien canonique pour des raisons de simplification de calcul.

De plus, une revalorisation trop élevée pourrait entrainer un risque de résiliation lui aussi important, ce qui n'est pas souhaité ici.

Maintenant que la question de la dérive de sinistralité a été traitée, une autre question émane elle aussi de la sélection médicale. En effet, si celle-ci veut avoir un réel impact sur la sinistralité, ne nécessite-t-elle pas d'évoluer avec le temps et notamment de s'adapter à l'essor de nouvelles maladies ?

V. Modélisations autour de la sélection médicale : le cas d'une pandémie

Si la sélection médicale est définie comme l'étape primordiale lors de la souscription d'un contrat en prévoyance individuelle permettant de connaître au mieux la sinistralité, que devient-elle lorsque la sinistralité elle-même varie au cours du temps ?

De plus, nul n'est sans ignorer qu'en 2020 une pandémie sans précédent a vu le jour portant le nom de Covid.

Par conséquent, il est légitime de se questionner sur le potentiel impact d'une telle pandémie concernant la sinistralité du portefeuille. De surcroît, il devient intéressant de s'interroger sur la possible intégration d'une question relative à la pandémie lors du fameux questionnaire médical.

Il est indéniable que l'apparition de nouvelles maladies bouscule la sélection médicale et amène de nombreuses questions sur le devant de la scène pour les assureurs. L'objectif de cette partie sera donc d'éclairer le lecteur sur ces différents points évoqués, et notamment de proposer une réponse à la prise en compte de l'impact de la Covid sur la sinistralité lors du questionnaire de sélection médicale.

Pour y arriver, deux étapes seront nécessaires.
La première concerne tout un travail de refonte du questionnaire de sélection médicale propre au produit SLPI. En effet, la mise à disposition de ce dernier étant impossible pour l'étude, il a été obligé de passer par une étape de reconstruction via la création d'une base de données.

La seconde étape, qui est une digression de la première, est l'intégration de l'information Covid dans le questionnaire ainsi reconstitué grâce aux arbres de classification notamment.

Enfin, en complément de ces deux étapes, une dernière sous-partie viendra compléter cette partie au sujet de la corrélation entre sélection médicale et nouvelle maladie. Il s'agira d'un scénario ORSA. L'objectif sera alors de répondre à la question suivante : quel serait l'impact sur la sinistralité de l'entreprise si la Covid ou toute autre nouvelle maladie venait à ne plus être prise en compte via la sélection médicale ?

1 Le ratio de sinistralité en cas de pandémie

Premier modèle établi dans ce mémoire, le ratio de sinistralité avait permis de mettre en lumière l'impact de la sélection médicale sur la sinistralité au cours des premières années d'ancienneté dans le risque.

Celui-ci avait été construit de manière à permettre la comparaison entre la sinistralité du portefeuille SLPI et une sinistralité de référence représentée par la population des TNS en France.
Par conséquent, il devient intéressant de se questionner sur les résultats de ce modèle dans le cas de la pandémie Covid. Il est important de préciser que l'exemple de la Covid est choisi ici, mais que l'objectif du raisonnement n'est pas de se limiter uniquement à la Covid mais bien à toute nouvelle pandémie ou maladie qui pourrait être amenée à se développer au cours des années à venir.

Tout comme lors de la mise en place du dérivé du ratio SMR, il a été question de la construction d'une loi d'incidence Covid. Cette dernière a été construite sur la base de données recueillies sur le site Covid Tracker. Pour plus d'informations au sujet des données utilisées, le lecteur pourra se référer au site mentionné ici [5].

Le taux d'incidence sera défini ici comme le nombre de cas positifs à la Covid sur sept jours rapporté à 100 000 habitants de chaque tranche d'âge. Les tranches d'âge utilisées seront les mêmes que précédemment, à savoir : [18 ans, 39 ans], [40 ans, 49 ans], [50 ans, 60 ans] et plus de 60 ans.

Par ailleurs, afin de permettre l'analyse la plus précise qu'il soit, différentes périodes depuis le début de la propagation du virus ont été retenues.

La première période s'étend du 6 au 12 octobre 2021, constituant une situation plutôt faible en termes d'incidence du virus. La période suivante couvre la semaine du 1er au 7 décembre 2021. Celle-ci précède une période de forte incidence qui s'est déroulée du 19 au 25 janvier 2022. Enfin, la dernière période choisie est une période de retour à la normale ou encore de dépression, couvrant le mois de mai 2022 du 4 au 10.

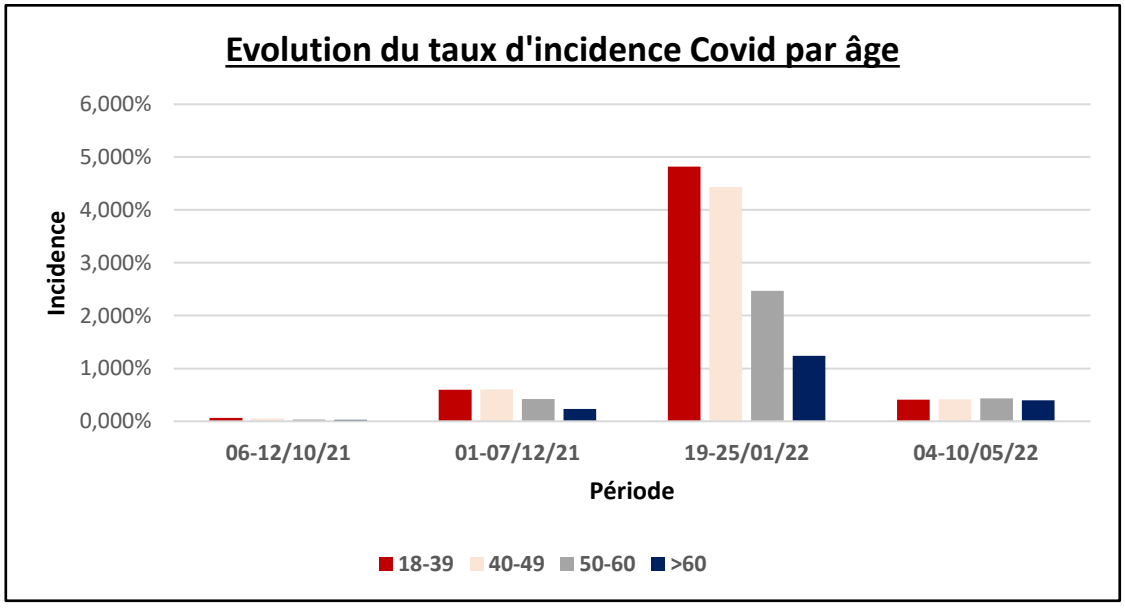


Figure 50 - Loi d'incidence Covid

Les taux d'incidence précis sont rappelés dans le tableau suivant.

	Incidence			
Age	06-12/10/21	01-07/12/21	19-25/01/22	04-10/05/22
18-39	0,06%	0,60%	4,82%	0,41%
40-49	0,05%	0,60%	4,43%	0,41%
50-60	0,04%	0,42%	2,47%	0,43%
>60	0,03%	0,23%	1,24%	0,40%

Figure 51 - Tableau - Loi d'incidence Covid

Par ailleurs, la population de référence reste elle aussi inchangée par rapport au modèle présenté dans la partie précédente. Il s'agit une nouvelle fois de la population TNS en France.

Le graphique suivant présente les résultats obtenus.

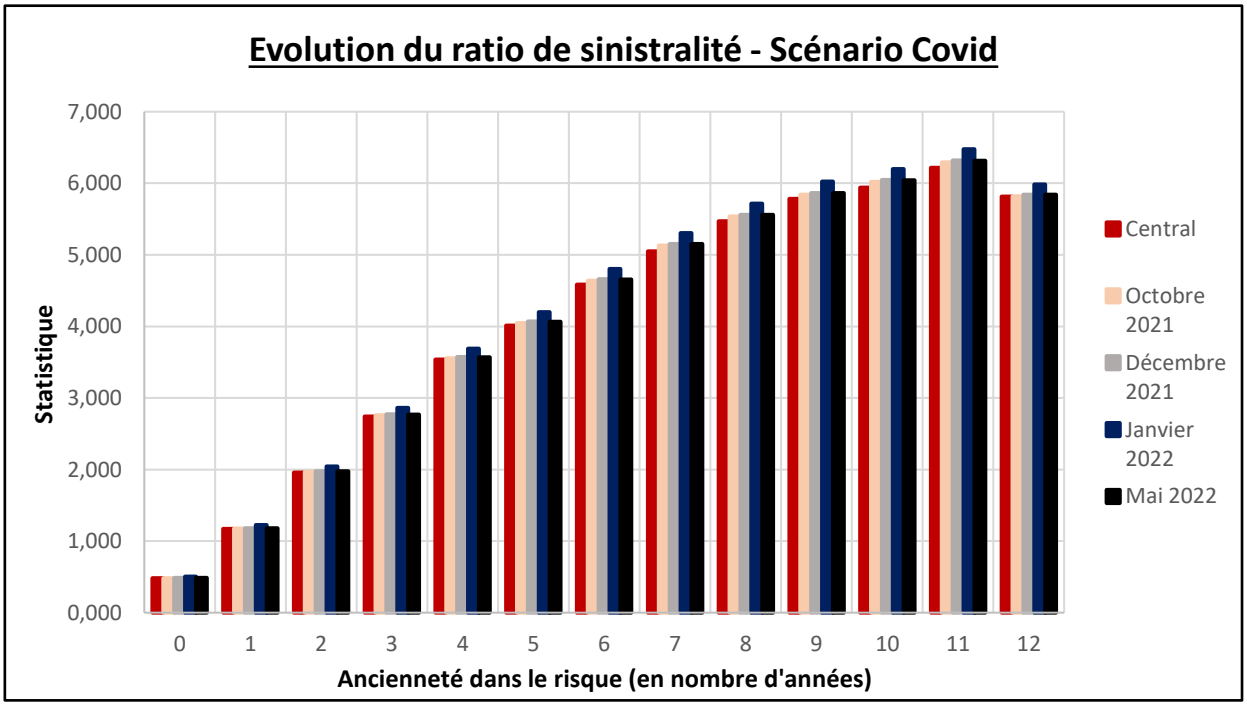


Figure 52 - Ratio de sinistralité - Prise en compte de l'incidence Covid

Le phénomène observé précédemment se répète. La sélection médicale permet bien de ramener la sinistralité propre au portefeuille SLPI à une sinistralité plus faible que celle de la population de référence TNS. Toutefois, comme mentionné à de nombreuses reprises maintenant, cet effet s'atténue avec l'ancienneté dans le risque. De plus, il est intéressant de souligner que cet effet est d'autant plus faible que le taux d'incidence de la pathologie Covid est grand. En effet, la statistique de test se voit augmentée de près de 4% entre le scénario central et la période dénommée « Janvier 2022 » par exemple, justifiant la nécessité d'ajuster le questionnaire médical.

Par conséquent, il est donc primordial de prendre en compte la Covid puisque cette dernière à un réel impact sur la sinistralité. Et cet impact est en réalité double. En effet, plus il y a de recul par rapport à la Covid, plus il est connu que cette maladie en plus d'impacter un individu

à un moment précis, peut aussi potentiellement laisser des séquelles plus longues. Or, de telles séquelles pourraient conduire à une hausse de la morbidité, d'où la nécessité d'intégrer l'information Covid au sein du questionnaire de sélection médicale.

Il est donc à présent manifeste que la sélection médicale ne peut être un processus figé dans le temps. Pour maximiser son efficacité dans le temps, elle se doit d'évoluer en fonction des nouvelles pathologies apparaissant au cours des années.

Cependant, avant de formaliser cette remarque et de l'intégrer au questionnaire de sélection médicale du produit SLPI, une étape intermédiaire mérite d'être présentée : la modélisation de la sélection médicale propre au produit SLPI.

2 Le questionnaire médical chez SwissLife

Composé de multiples questions allant de l'âge à la profession en passant par l'état de santé de la personne, le questionnaire médical est un outil essentiel voire stratégique, utilisé en amont de la souscription dans le but de déterminer l'acceptation ou non d'un futur assuré. C'est la raison pour laquelle, d'un point de vue de confidentialité, il a été impossible de se procurer un tel document pour le produit SLPI.

De manière contrainte mais non résignée, il a donc été nécessaire de passer par une étape dite de reconstruction du questionnaire de sélection médicale.

De plus, avant d'aboutir à cette modélisation appuyée sur la théorie des arbres de classification, il a également été indispensable de procéder à la création d'une base de données type.

En effet, comme cela a été plus amplement détaillé dans la partie théorie prévue à cet effet, les arbres de classification sont des méthodes d'apprentissage de Machine Learning. Ainsi, qui dit apprentissage dit nécessairement base de données pour apprendre. Or, la base de données utilisée jusqu'à maintenant est une base de données comprenant des individus ayant déjà passé l'étape de sélection médicale. Plus encore, ces individus l'ont passée avec succès, puisqu'ils ont été acceptés dans le risque.

Utiliser cette base, bien que propre au produit SLPI, engendrerait donc un biais. C'est la raison pour laquelle, la première étape de cette partie sera entièrement consacrée à la reconstruction d'une base de données.

2.1 Construction d'une base de données pour l'apprentissage

Choix des variables à modéliser

L'idée générale qui a motivé la mise en place de la base de données tout au long de sa construction était de coller au mieux à une population type de personnes souhaitant souscrire un contrat d'assurance.

Cette première idée a permis de définir trois premières variables nécessaires à la base de données :

- **L'âge**
- **Le poids**
- **La taille**

De plus, la variable **fumeur** a été ajoutée à cette première trilogie.

Afin de compléter ces quatre premières variables, il a été nécessaire de se référer à des études qui avaient été menées antérieurement par SwissLife permettant de relier certaines pathologies au fait d'accepter ou non le risque à assurer.

Il faut également garder en tête que la reconstruction d'une base de données est un travail long et fastidieux. C'est la raison pour laquelle, en plus d'une question de confidentialité, toutes les questions présentes dans le questionnaire de sélection médicale ne seront pas toutes modélisées par des variables. Un choix qualitatif et exhaustif a donc dû être fait.

Les variables dites pathologiques choisies sont donc les suivantes :

- **L'asthme**
- **La thyroïde**
- **La thyroïde cancéreuse**
- **L'hypertension artérielle**
- **Le cholestérol**

Bien que la variable Covid n'intervienne pas encore ici, il est important de souligner que ces variables ont également été choisies pour leur potentielle corrélation avec cette dernière.

La dernière variable modélisée et non des moindres est la variable portant le nom de **risque**. Celle-ci permet de modéliser le fait d'accepter ou non le risque à assurer, ou dans le cas intermédiaire, de proposer une étude plus approfondie du dossier relatif à la personne.

Modélisation des variables choisies

Une fois le choix des variables effectué, la deuxième étape consiste en la modélisation de celles-ci.
Chaque variable étant différente et n'ayant pas la même information à révéler, la modélisation ne peut donc pas être commune pour chacune d'elles. Toutefois, une base de raisonnement commune a été utilisée.
Pour illustrer cela, il est temps de se pencher sur la modélisation de la première variable choisie, à savoir l'âge.

- Variable AGE

Afin de modéliser cette première variable, deux hypothèses ont été retenues. La première est que les personnes souhaitant souscrire un contrat d'assurance ont un âge uniformément réparti entre 18 et 65 ans. Cette première hypothèse pourrait être contestée puisqu'il est fort probable que passé un certain âge, certainement supérieur à 18 ans, un nombre plus important de

personnes prennent conscience de l'importance de se couvrir. Toutefois, dans un souci de simplicité cette hypothèse a été retenue.
D'un point de vue mathématique, cette modélisation a été réalisée grâce à l'utilisation d'une loi uniforme conditionnée entre deux bornes : 18 et 65 ans.

La seconde hypothèse est que cette variable **âge**, a été par la suite découpée en trois classes que voici : [18 ans, 39 ans], [40 ans, 49 ans] et [50 ans, 62 ans], les individus avec un âge supérieur à 62 ans étant automatiquement non acceptés dans le risque.

- Variable TAILLE

Concernant la modélisation de la variable **taille**, l'hypothèse choisie est la suivante : la taille des individus est supposée prendre des valeurs comprises entre 145 et 195 cm.
Là encore, cette hypothèse est discutable puisqu'il existe dans la population des personnes mesurant plus d'1m95. Toutefois, ce n'est pas la majorité et ces valeurs sont en général extrêmes.

Une fois ce critère fixé, la simulation a été réalisée en conditionnant une loi uniforme par les deux bornes choisies.

- Variable POIDS

Une première idée pour simuler la variable **poids** avait été de procéder comme précédemment en fixant deux bornes dans le but de conditionner une loi uniforme. Cependant, procéder ainsi aurait pu conduire à des poids et des tailles complètement décorrélés, comme un individu mesurant 1m50 et pesant 100 kg. Ce cas peut certes exister mais ce n'est qu'une exception.
Par conséquent, dans le but de garder toutefois le caractère aléatoire, il a été décidé de borner les poids possibles en fonction de la taille.

La règle construite est la suivante :

- Si $T \in [145\text{ cm}, 165\text{ cm}]$, alors $P \in [50\text{ kg}, 80\text{ kg}]$
- Si $T \in [166\text{ cm}, 176\text{ cm}]$, alors $P \in [52\text{ kg}, 95\text{ kg}]$
- Si $T \in [177\text{ cm}, 195\text{ cm}]$, alors $P \in [60\text{ kg}, 105\text{ kg}]$

où P représente le poids et T la taille

- La variable RPT

Directement liée aux variables poids et taille, la variable **rapport poids taille (RPT)** est similaire à l'indice de masse corporelle (IMC).

Voici la formule utilisée permettant de le calculer :

$$RPT = 1 + \frac{P}{T^2} \times 100$$

Cette variable a également été segmentée en quatre classes selon le découpage suivant :

- RPT $\leq 1{,}25$
- RPT $\in [1{,}26; 1{,}30]$
- RPT $\in [1{,}31; 1{,}35]$
- RPT $> 1{,}35$

- Les variables pathologiques : FUMEUR, ASTHME, THYROIDE, THYROIDE CANCEREUSE, HTA et CHOLESTEROL

Dans le but de déterminer le conditionnement à ajouter à la loi uniforme, il a été nécessaire de collecter différentes informations sur les pathologies choisies.

Pour aboutir à cela, différents sites de référence ont été utilisés tels que Santé Publique France, l'Inserm, la Haute Autorité de Santé ou encore le site Cancer.org.
Pour toute information complémentaire sur les renseignements collectés, le lecteur est invité à se reporter aux références mentionnées ici [6].

Fumeur

L'information retenue pour modéliser la variable **fumeur** est celle-ci. En 2017, dans la population française il y avait 26,9% de fumeurs.

Asthme

La France compte aujourd'hui près de 4 millions d'asthmatiques. Rapporté aux 67 millions de français cela représente 5,94% des Français.

Thyroïde

La thyroïde est une pathologie qui touche 10% de la population française. De plus, parmi ces 10%, 1% des cas sont des thyroïdes cancéreuses.

Hypertension artérielle

En France, 15 millions de personnes souffrent d'hypertension artérielle. Rapporté à la population totale française, ce chiffre ne représente alors pas loin de 22% de cas.
Pour information, il s'agit de la maladie chronique la plus fréquente en France, l'âge n'étant pas l'unique facteur de risque.

Cholestérol

Enfin, concernant le cholestérol, il s'agit d'une pathologie présente chez plus de 20% des Français.

Toutes ces variables dites pathologiques seront regroupées en deux classes : Oui et Non. Le Oui, indiquant que la personne souffre de la pathologie, et le Non indiquant au contraire que la personne n'en souffre pas. Par ailleurs, un même individu peut souffrir d'aucune, d'une ou de plusieurs pathologies à la fois.

- Variable RISQUE

Objet central de l'étude, la variable dénommée sous le nom de **risque**, est une variable dont le but est de signifier l'acceptation ou non d'un assuré dans le portefeuille en fonction de ses caractéristiques, caractéristiques représentées par les variables énoncées plus haut.

Cette variable factorielle se décompose en trois classes :

- L'acceptation mentionnée par le facteur « Oui »
- Le refus mentionné par le facteur « Refus »
- L'étude approfondie du dossier symbolisée par la mention « Study »

Avant de refermer cette page spécialement dédiée à la reconstruction d'une base de données dans le but de construire un arbre de décision pour la sélection médicale, voici un aperçu de quelques statistiques descriptives propres à la base construite.

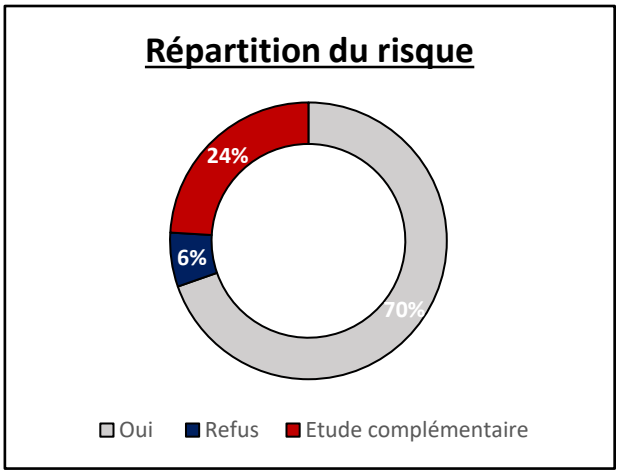


Figure 53 - Répartition du risque

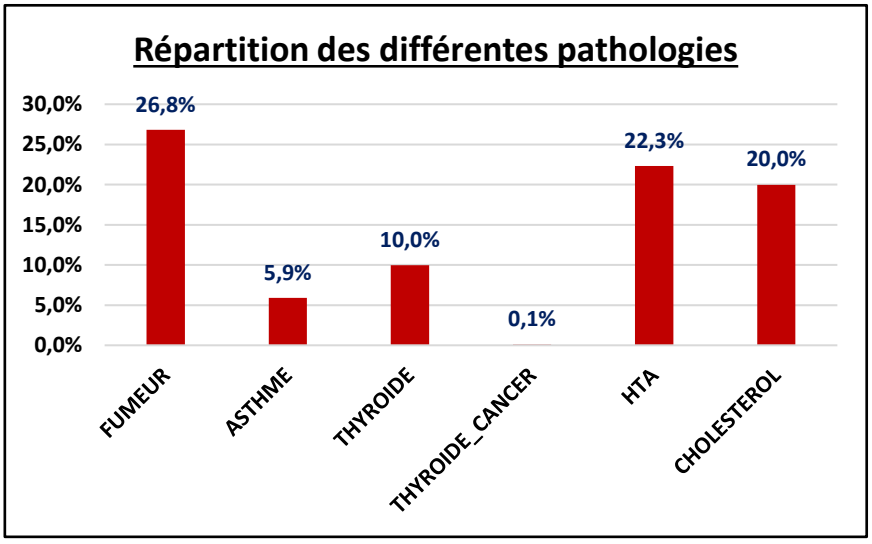


Figure 54 - Répartition des pathologies présentes dans la base

Le premier graphique permet de mettre en lumière la répartition du risque au sein de la base de données créée.

En majorité, les personnes souhaitant souscrire au produit SLPI sont acceptées. Dans un peu moins d'un quart des cas une étude complémentaire et plus approfondie sera demandée à la personne. Enfin, le pourcentage de personnes refusées à la souscription reste minoritaire avec seulement 6% des cas.

Le deuxième diagramme reflète quant à lui la répartition des différentes pathologies au sein de la base de données. Il est intéressant de remarquer que les pourcentages évoqués au sein de la population nationale sont correctement représentés dans l'échantillon proposé ici. Pour information, cet échantillon compte au total 318 060 têtes.

Bien qu'étant une étape indispensable pour la bonne conduite de l'étude, la reconstruction de la base de données n'est en rien une finalité.
A présent, l'idée est de modéliser la manière dont le risque est accepté ou non. Pour y parvenir, la modélisation s'appuiera sur différentes méthodes de Machine Learning telles que les arbres de classification, les Random Forests ainsi que les techniques de Gradient Boosting.

2.2 Modélisations et résultats

L'objectif de cette sous-partie est de présenter les résultats obtenus sur la base de données reconstruite grâce aux différentes méthodes détaillées dans la partie théorie.

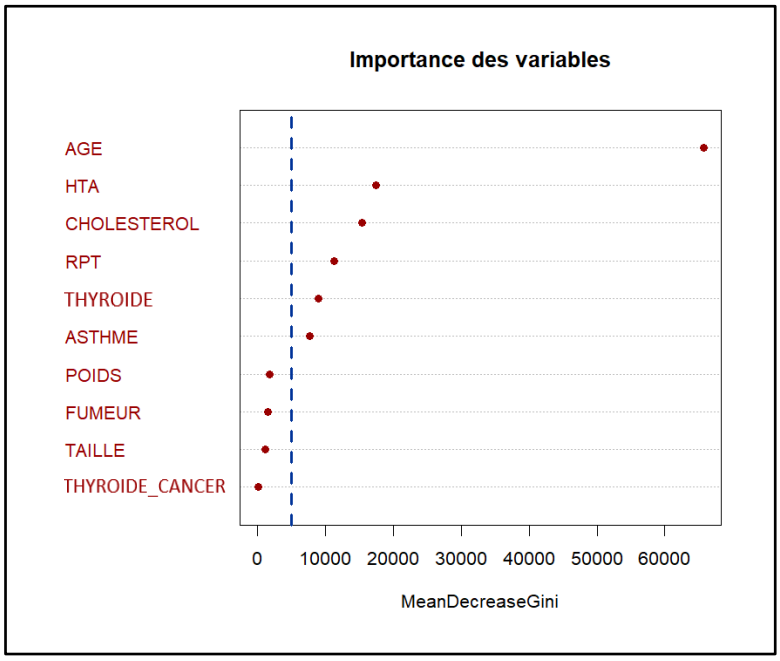
Pour rappel également, l'idée est ici d'exposer la manière dont l'acceptation du risque propre au portefeuille SLPI est faite par l'intermédiaire du questionnaire de sélection médicale.

Découpage de la base de données

Comme dans le cadre des GLM, avant de se lancer pleinement dans une quelconque modélisation, la base initiale a été séparée en deux bases : une base d'apprentissage représentant 75% des données et une base de test regroupant quant à elle 25% des données.

Importance des variables

En parallèle du découpage de la base de données en deux échantillons, un test permettant de mettre en lumière l'importance des variables a été effectué. Pour y parvenir, il a été décidé de lancer une forêt aléatoire et d'étudier le graphique présentant l'importance de chaque variable. La sortie R obtenue est la suivante :



AGE	HTA	CHOLESTEROL	RPT	THYROIDE	ASTHME	POIDS	FUMEUR	TAILLE	THYROIDE_CANCER
65 679	17 373	15 428	11 198	8 893	7 634	1 735	1 542	1 155	22

Figure 55 - Importance des variables - Sans Covid

Une variable se démarque des autres. Il s'agit de la variable AGE. Cela peut s'expliquer par le fait que l'exclusion du risque est entièrement dépendante de l'âge. Pour rappel, dans la modélisation choisie ici, au-delà de 62 ans, la personne souhaitant souscrire un contrat SLPI se verra automatiquement refusée.

En second plan, viennent cinq variables : les variables HTA, CHOLESTEROL, THYROIDE, ASTHME pour les variables dites pathologiques et la variable RPT.

Concernant les variables POIDS, FUMEUR, TAILLE et THYROIDE_CANCER présentant une importance bien plus faible que les autres, presque au minimum cinq fois plus faible, il a été décidé de les exclure.

La faible importance de la variable THYROIDE_CANCER s'explique par le fait que cette pathologie est rare et peu représentée dans la base d'étude.

Par ailleurs, au sujet des variables TAILLE et POIDS, l'explication de cette faible importance est que les informations apportées par ces deux variables sont déjà reflétées dans le ratio RPT.

2.2.1 Arbres CART

Lorsqu'un arbre de décision est modélisé sur R, il est primordial de réaliser différents tests dans le but de fixer un certain nombre de paramètres. Parmi eux, sont à retrouver les critères d'arrêt et d'élagage.

Règles d'arrêt

Plusieurs règles d'arrêt peuvent être fixées via différentes commandes sur R.

La première règle d'arrêt est la profondeur maximale de l'arbre grâce à la commande *maxdepth.* Par défaut, sa valeur est fixée à 30.

Le second paramètre pouvant faire office de règle d'arrêt est le nombre minimal d'individus présents au niveau d'un nœud pour envisager une coupure. Autrement dit, le nombre minimal d'observations requis pour couper chaque nœud. Ce paramètre permet notamment d'éviter le surapprentissage. Il est représenté par la variable *minsplit* dans R. Sa valeur par défaut est égale à 20.

Enfin, la troisième commande en lien avec la précédente porte le nom de *minbucket.* Il s'agit du nombre minimal d'individus qui engendrerait la coupure d'un nœud parent. Sa valeur par défaut dans R est reliée de manière proportionnelle à la valeur de la variable *minsplit* par la formule décrite ci-dessous :

$$minbucket = \frac{minsplit}{3}$$

Pour répondre à ces règles d'arrêt optimales, plusieurs tests ont donc été effectués, dont voici une illustration :

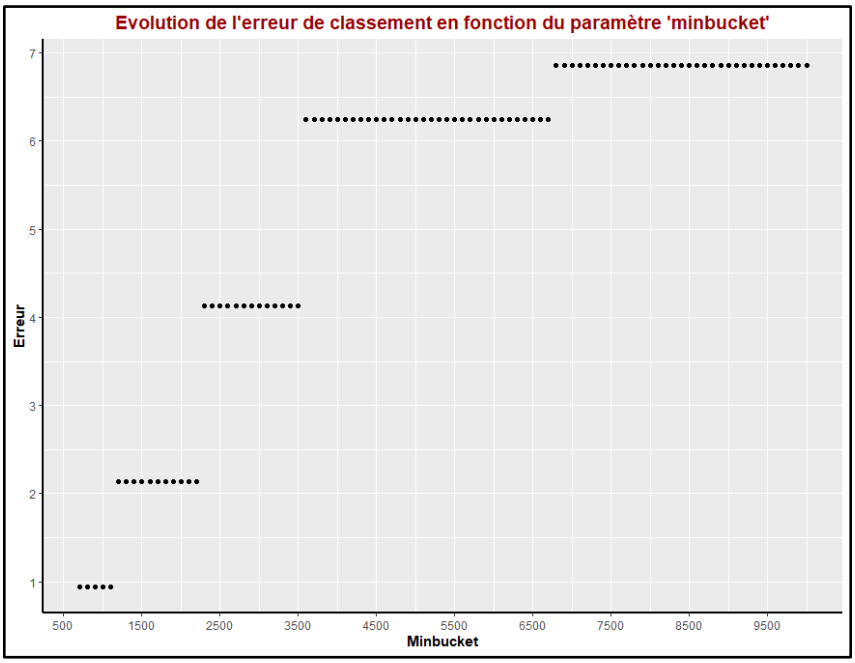


Figure 56 - Evolution de l'erreur de classement en fonction du paramètre « minbucket »

L'erreur de classement restant constante et égale à 0,94 jusqu'à une valeur de *minbucket* égale à 1100, il a donc été décidé de se focaliser sur l'impact des valeurs de *minbucket* au-delà de 1100.

Par ailleurs, il est possible d'observer deux choses. La première est qu'à partir de cette valeur de 1100 exclue, l'erreur de classement ne cesse d'augmenter.
Deuxièmement, cette dite erreur augmente par paliers.

Par conséquent, il a été décidé de conserver une valeur quasi moyenne pour le paramètre de *minbucket,* à savoir 500. Ainsi obtenue, cette valeur permet de déterminer le paramètre de *minsplit* par proportionnalité.

Concernant le paramètre permettant de fixer la profondeur maximale de l'arbre de décision, aucune modification n'y a été apportée. En effet, plusieurs tests réalisés ont démontré que la profondeur maximale par défaut n'était jamais atteinte.

Elagage

L'élagage ou le pruning est une étape cruciale qui doit être réalisée de manière judicieuse. En effet, il s'agit de trouver le bon compromis entre précision et pouvoir prédictif. Pour y parvenir, il convient alors de déterminer le coefficient de complexité optimal aussi noté *cp*, qui minimise l'erreur relative. Cette étape est généralement réalisée par cross-validation.

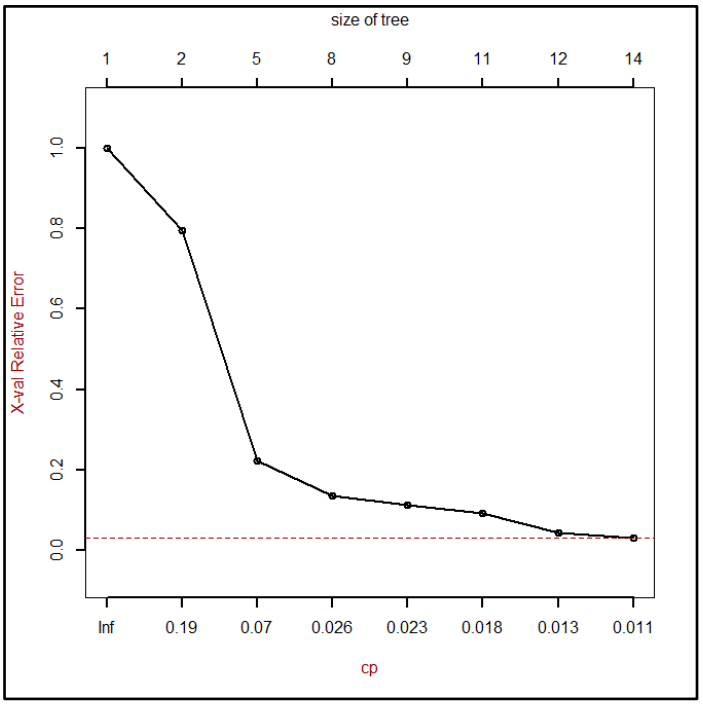


Figure 57 - Evolution de l'erreur en fonction du paramètre de complexité optimal

Dans le cas de la base d'étude, il est possible d'observer que l'erreur relative diminue fortement avec les premiers paramètres de complexité. Elle vient ensuite se stabiliser autour d'un coefficient égal à 0,011.
Un tel élagage conduit alors à un arbre de décision composé de 14 feuilles.

Pureté de l'arbre

Concernant la pureté de l'arbre, les résultats obtenus en considérant l'indice de Gini sont identiques à ceux obtenus en considérant l'indice d'entropie. Par conséquent, il a été décidé de conserver l'indice de Gini comme mesure de pureté.

Une fois tous les paramètres fixés et optimisés une représentation de l'arbre optimal est obtenue. Pour une meilleure lisibilité, le lecteur est invité à se référer à la partie **Annexes**.

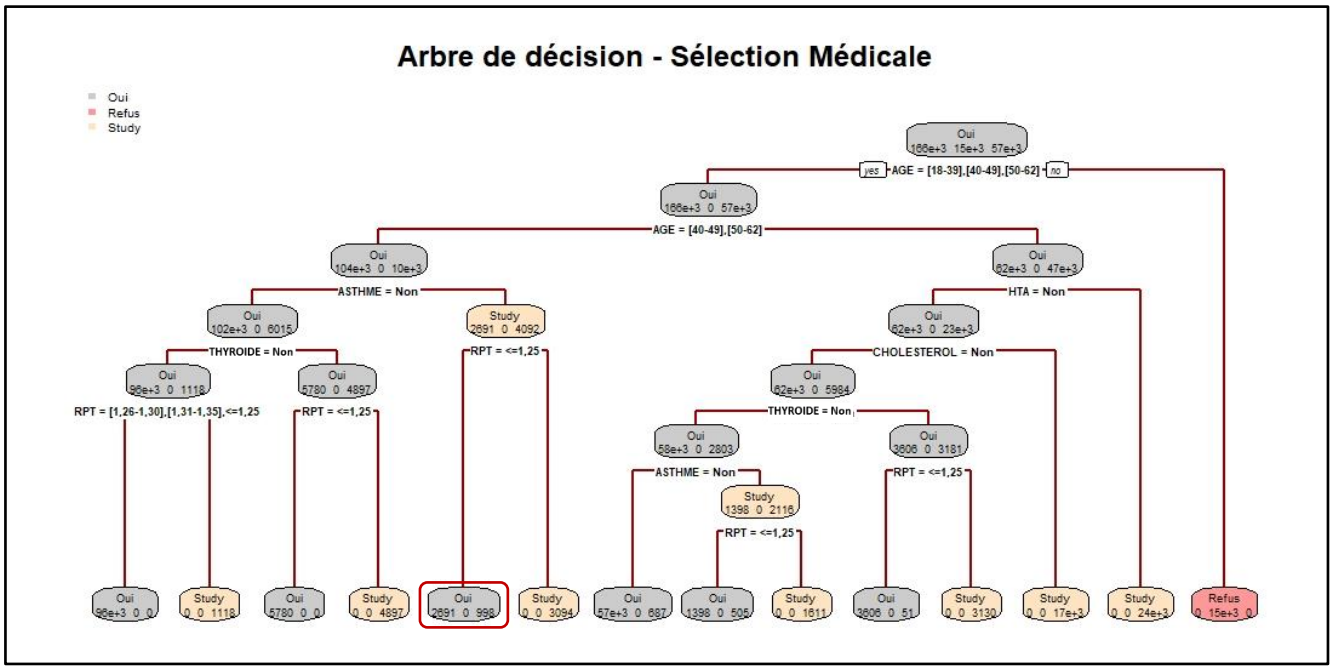


Figure 58 - Arbre optimal de décision - Sélection Médicale sans Covid

Chaque nœud obtenu contient une information de type « texte » et une information de type « nombre » sous la forme d'un triplet.

L'information de type « texte », à savoir « Oui », « Refus » et « Study » concerne les valeurs possibles prises par la variable RISQUE.

L'information de type « nombre » indique quant à elle le nombre d'individus contenus dans chaque nœud et selon ses spécificités. Elle permet de justifier le choix de l'information de type « texte ».
Le nombre de gauche représente le nombre d'individus associés à la mention « Oui », donc acceptés dans le risque. A l'inverse, le nombre du milieu représente le nombre d'individus non acceptés dans le risque, donc associés à la catégorie « Refus ». Enfin, le nombre de droite, représente le nombre d'individus reliés à la classe d'étude nommée « Study ».
Le nombre le plus grand dans le nœud permet par la suite l'association de ce dernier à la classe correspondante.

Pour illustrer ceci, l'individu entouré en rouge sur l'arbre ci-dessus servira d'exemple. Il s'agit d'un individu dont l'âge est compris entre 40 et 62 ans, présentant de l'asthme et ayant un rapport poids taille inférieur à 1,25.

De plus, comme annoncé lors de la partie relative à l'importance des variables, la variable AGE joue un rôle important dans l'attribution des classes, en intervenant en premier.
Une fois les différents filtres sur les variables pathologiques passés, la variable RPT représentant le ratio poids taille permet d'affiner la classification.
Une fois le modèle obtenu, une bonne approche pour juger de la qualité de ce dernier est d'étudier la matrice de confusion associée.

```
      data.test_predict
        Oui Refus Study
Oui   55196     0     0
Refus     0  5204     0
Study   749     0 18366
```

	Oui Prédit	Refus Prédit	Study Prédit
Oui Réel	55 196	0	0
Refus Réel	0	5 204	0
Study Réel	749	0	18 366

Figure 59 - Matrice de confusion - Arbre CART sans Covid

Les termes diagonaux de la matrice de confusion correspondent aux individus correctement classés. Les autres coefficients de la matrice représentent quant à eux les individus dont la valeur de prédiction est erronée.

L'erreur globale est obtenue en faisant la somme des termes diagonaux rapportée à la somme de tous les termes de la matrice de confusion. La valeur de l'erreur de prédiction est alors proche de 0,94%.

Comme cela a déjà été évoqué à maintes reprises, un des risques des modèles de Machine Learning est le surapprentissage. L'erreur de prédiction peut sembler correcte à première vue mais cacher de l'overfitting. Une autre manière de tester le pouvoir prédictif d'un algorithme est alors de procéder à une cross-validation.

```
CART

238545 samples
     6 predictor
     3 classes: 'Oui', 'Refus', 'Study'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 214691, 214690, 214691, 214690, 214691, 214690, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.07258634  0.8970088  0.7419642
  0.10576510  0.8102125  0.4700142
  0.20609039  0.7408326  0.2085392

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.07258634.
```

Figure 60 - Résultats de la cross-validation - Arbre CART sans Covid

Dans le cas d'un arbre de classification et non de régression, deux métriques sont pertinentes à regarder : l'accuracy et la métrique dite Kappa de Cohen.

Très communément utilisée, l'accuracy permet de mesurer la précision entre les valeurs prédites et les valeurs réelles.

La métrique Kappa de Cohen, ou tout simplement le k de Cohen, est quant à elle utilisée pour quantifier la performance du classifieur mis en place en la comparant à la performance d'un classifieur aléatoire.

Une valeur proche de 100% de ces deux métriques est signe du bon pouvoir prédictif de l'algorithme, et donc d'un bon modèle.

Dans le cas précis de l'étude, il est possible d'observer une valeur d'environ 89,7% pour l'accuracy et 74,2% pour la métrique Kappa, confirmant ainsi un pouvoir prédictif correct.

Les arbres sont des algorithmes simples et facilement interprétables. Cependant, leur pouvoir prédictif est souvent limité. De plus, ils conduisent très souvent à du surapprentissage. Par conséquent, d'autres techniques plus robustes comme les forêts aléatoires peuvent être mises en place.

2.2.2 Forêts aléatoires

Bien que la technique de Machine Learning évolue, l'idée reste la même : en fonction des variables choisies, déterminer l'acceptation ou non d'un individu dans le risque.

Les variables sélectionnées précédemment lors de l'application des arbres de décision sont conservées. Pour rappel, il s'agissait des variables : AGE, HTA, CHOLESTEROL, THYROIDE, RPT et ASTHME.

Par ailleurs, tout comme dans le cadre des arbres CART, une étape de choix des hyperparamètres est nécessaire.

Parmi eux, peut être cité le nombre d'arbres composant la forêt aléatoire. Ce dernier représentera alors le nombre d'apprenants faibles.

Une manière de procéder afin de le déterminer est de regarder l'erreur appelée Out-Of-Bag (OOB).
Cette mesure permet de quantifier l'erreur des algorithmes d'agrégation. En effet, la forêt aléatoire est construite sur un échantillon de données dit « In Bag ». L'échantillon restant est alors dit « Out-Of-Bag » et permet de tester l'erreur de prédiction du modèle. Le but est alors de minimiser cette erreur de prédiction et donc l'erreur Out-Of-Bag.

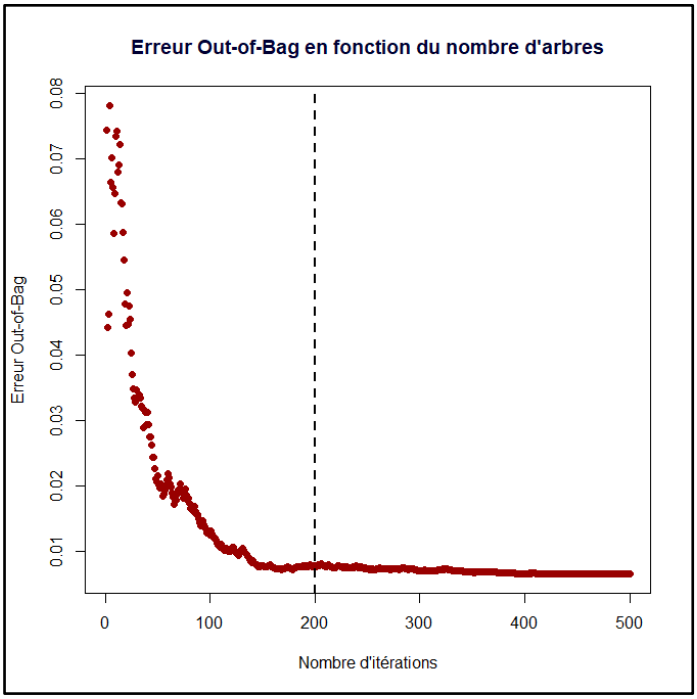


Figure 61 - Evolution de l'erreur OOB en fonction du nombre d'itérations - Sans Covid

Le graphique ci-dessus représente l'évolution de l'erreur « Out-Of-Bag » en fonction du nombre d'apprenants faibles du modèle. Il est possible de remarquer que cette erreur se stabilise à partir de 200 itérations. C'est la raison pour laquelle le paramètre *ntree* a été fixé à 200 dans le modèle.

Une fois les paramètres fixés et le modèle entrainé vient l'étape d'analyse de la matrice de confusion que voici :

```
       forest_predict_2
          Oui Refus Study
Oui     55194     0     2
Refus       0  5204     0
Study     509     0 18606
```

	Oui Prédit	Refus Prédit	Study Prédit
Oui Réel	55 194	0	0
Refus Réel	0	5 204	0
Study Réel	509	0	18 606

Figure 62 - Matrice de confusion - Forêt aléatoire sans Covid

Une fois encore, le pouvoir prédictif du modèle est satisfaisant. En sommant les coefficients diagonaux de la matrice, puis en les rapportant à la somme de tous les coefficients, l'erreur obtenue est de 0,64%.

Toutefois, tout comme dans le cas des arbres, de l'overfitting peut apparaître. Une étape de cross-validation est donc requise. Pour information, l'étape de cross-validation pour les forêts aléatoires a été réalisée sur cinq blocs.

K	Erreur
1	0,64%
2	0,61%
3	0,61%
4	0,73%
5	0,68%

Figure 63 - Résultats Cross-Validation - Forêts aléatoires sans Covid

2.2.3 Gradient Boosting Machine

Dernière méthode exposée ici, le Gradient Boosting Machine (GBM) permet de venir compléter les deux premières approches présentées.

Tout comme les forêts aléatoires, cet algorithme ne possède pas de représentation visuelle similaire aux arbres CART. Toutefois, l'algorithme du Gradient Boosting permet l'agrégation d'apprenants faibles et ce, quelle que soit la fonction de perte utilisée.

Comme il est facile de le prévoir maintenant, une série de paramètres est à déterminer avant de lancer un quelconque modèle.

Il a été choisi de fixer les paramètres suivants du logiciel R :

- ***nrounds :*** il s'agit du nombre maximal d'itérations. Dans le cas de la classification, ce nombre est équivalent au nombre maximal d'arbres
- ***maxdepth :*** il s'agit de la profondeur maximale d'un arbre, ce paramètre permet notamment d'éviter l'overfitting. Par défaut, il est fixé à 6.
- ***colsample_bytree :*** il s'agit du rapport de sous-échantillons des colonnes lors de la construction de chaque arbre. Par défaut, il est fixé à 1.
- ***eta :*** ce paramètre permet de contrôler le taux d'apprentissage en adaptant la contribution de chaque arbre par un facteur $\eta \in]0;1[$ lorsqu'il est agrégé au classifieur central. Plus la valeur de ce paramètre est élevée, plus le modèle est robuste. Sa valeur par défaut est de 0,3.
- ***gamma :*** il s'agit de la valeur minimale requise pour créer un nouveau nœud. Plus la valeur de ce paramètre est grande plus l'algorithme est conservateur.
- ***min_child_weight :*** il s'agit de la somme minimale des poids requise pour engendrer un partitionnement supplémentaire. Sa valeur par défaut est fixée à 1.
- ***subsample :*** il s'agit du rapport entre les données utilisées comme échantillonnage d'entrainement et celles utilisées comme base de test. Par défaut, il est fixé à 1.

Afin de déterminer les paramètres optimaux du modèle, une étape de cross-validation dont voici les résultats a été lancée.

```
  nrounds max_depth eta gamma colsample_bytree min_child_weight subsample
3     100         3 0.1     0              0.6                1         1
```

Figure 64 - Paramètres optimaux obtenus par cross-validation - GBM sans Covid

Une fois ces paramètres déterminés, le modèle a pu être entrainé.

```
Confusion Matrix and Statistics

          Reference
Prediction   Oui Refus Study
     Oui   66389    59     0
     Refus     1  6015    11
     Study     0   380 22561

Overall Statistics

               Accuracy : 0.9953
                 95% CI : (0.9948, 0.9957)
    No Information Rate : 0.6958
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9896

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Oui Class: Refus Class: Study
Sensitivity              1.0000      0.93198       0.9995
Specificity              0.9980      0.99987       0.9948
Pos Pred Value           0.9991      0.99801       0.9834
Neg Pred Value           1.0000      0.99509       0.9998
Prevalence               0.6958      0.06764       0.2366
Detection Rate           0.6958      0.06304       0.2364
Detection Prevalence     0.6964      0.06317       0.2404
Balanced Accuracy        0.9990      0.96592       0.9971
```

	Oui Prédit	Refus Prédit	Study Prédit
Oui Réel	66 389	59	0
Refus Réel	1	6 015	11
Study Réel	0	380	22 561

Figure 65 - Matrice de confusion & Kappa et Accuracy - GBM sans Covid

L'algorithme de Gradient Boosting conduit lui aussi à des résultats corrects en termes de prédiction. Les métriques de précision et Kappa sont proches de 1. L'erreur de prédiction quant à elle est légèrement plus faible qu'avec les autres méthodes déjà expérimentées comme le souligne le tableau récapitulatif ci-dessous.

Méthode	Erreur
Arbre CART	0,94%
Forêts Aléatoires	0,64%
GBM	0,47%

Figure 66 - Tableau récapitulatif des erreurs en fonction de la méthode choisie - Cas sans Covid

2.3 Conclusion

Cette partie avait pour vocation de modéliser la sélection d'un individu dans le risque pour le produit SLPI en se calquant au mieux avec ce qui est fait dans le questionnaire de sélection médicale.

Après un premier travail de reconstitution d'une base de données pour l'étude, plusieurs algorithmes de Machine Learning choisis, paramétrés et entrainés, rendent ainsi possible la prédiction du risque de chaque nouvel assuré potentiel en fonction de ses caractéristiques. Ces différentes méthodes, à savoir arbres CART, Random Forest et Gradient Boosting Machine, permettent de manière graduelle d'améliorer la qualité de prédiction tout en essayant de ne pas franchir la limite ténue de surapprentissage.
L'autre point de limite des algorithmes de Machine Learning est l'interprétabilité des résultats. En effet, plus les algorithmes se complexifient moins les résultats sont interprétables de manière visuelle et plus ces derniers deviennent semblables à des boites noires. D'un point de vue commercial, un arbre de décision sera certainement plus facilement assimilé par un très grand nombre que les résultats d'un Gradient Boosting Machine.

En complément de ces remarques, il est important de garder en tête que les résultats obtenus dépendent de la base initiale construite artificiellement et du bon paramétrage sur celle-ci. Le même travail appliqué à une base différente amènerait probablement à des résultats différents.

Enfin, bien que cette étape de mise en place de la modélisation du questionnaire de sélection médicale soit primordiale, elle ne constitue pas la finalité des travaux. Pour rappel, en amont de cette partie, il avait été mis en lumière que la sélection médicale ne pouvait être figée dans le temps et qu'elle se devait d'être adaptée en permanence en fonction des nouvelles pathologies voyant le jour. Parmi elles, peut être citée la Covid.

3 Un questionnaire médical amélioré : prise en compte de la pathologie Covid

Tant que le risque évoluera, la sélection médicale sera elle aussi amenée à évoluer pour ne pas sous-estimer la sinistralité.

Il devient alors intéressant de se questionner sur la façon dont prévenir une hausse de sinistralité liée à une nouvelle pathologie. Une première manière de raisonner est de se placer au niveau de l'assuré et d'intégrer l'information relative à cette nouvelle pathologie au sein du questionnaire médical, les individus présentant un risque trop peu connu et donc trop important étant ainsi exclus du portefeuille.

L'étude qui va suivre a pour objectif de présenter l'application de cette première manière de faire dans le cas précis de la Covid. Toutefois, ce processus pourrait se généraliser à toute autre maladie ayant un impact sur les pathologies déjà intégrées au sein du questionnaire de sélection médicale.

3.1 Construction d'une base de données pour l'apprentissage

Choix des variables à modéliser

Tout comme dans le scénario central sans Covid, il a été nécessaire de passer par une étape de reconstruction d'une base de données pour l'apprentissage. L'idée n'étant pas de refaire un travail en doublon, mais bien d'ajouter quelques variables à modéliser.

Par conséquent, les variables présentes initialement ont été conservées. Pour rappel, il s'agissait des variables suivantes :

- **Age**
- **Poids**
- **Taille**
- **RPT**
- **HTA**
- **Cholestérol**
- **Thyroïde**
- **Thyroïde cancéreuse**
- **Asthme**
- **Fumeur**
- **Risque**

En complément de ces variables et pour satisfaire l'objectif de l'étude, il a été choisi de modéliser trois variables supplémentaires relatives à la Covid.

La première variable portant le nom de **covid** permet simplement de relever si une personne a eu ou non la Covid.

La seconde variable sera nommée **hospi**. Cette dernière aura pour but de savoir si l'individu concerné a été hospitalisé en réanimation ou non alors qu'il était atteint de la Covid.

Enfin, la troisième et dernière variable portera le nom de **covid_long**. Bien qu'il soit encore trop tôt pour émettre une définition universelle d'un covid long, il a été décidé pour l'étude de considérer un covid comme long pour tout individu présentant encore des symptômes un mois après son infection.

Croiser des informations relatives à la Covid avec les pathologies préexistantes dans le questionnaire de sélection médicale est également justifié par la nature des symptômes de cette pandémie.
En effet, une étude de l'Inserm, l'Institut national de la santé et de la recherche médicale a révélé que parmi les symptômes liés à la Covid, il était possible de retrouver des douleurs thoraciques, de la dyspnée (essoufflement soudain), des troubles digestifs, de l'agueusie (perte de goût) ou encore de l'asthénie (fatigue). Il s'agit d'autant de symptômes pouvant être corrélés aux pathologies mentionnées dans le questionnaire de sélection médicale, telles que l'asthme ou encore le cholestérol.

Modélisation des variables choisies

Le raisonnement évoqué lors de la partie précédente quant à la modélisation des variables reste identique. Pour chacune d'elles, différentes études inspirées notamment de sites gouvernementaux ont permis leur caractérisation. Pour davantage d'informations, le lecteur pourra se référer ici [7].

Voici de manière plus détaillée les informations ayant permis de conditionner la loi uniforme afin d'obtenir la répartition de chaque variable.

- COVID

Entre le 1er mars 2020 correspondant au début de l'épidémie et le 30 juin 2021, environ 30 276 632 personnes ont été touchées par la Covid. Rapporté à la population française, ce nombre représente environ 44,9% des Français. Attention, ici il s'agit d'une simplification de modèle que de considérer qu'une personne équivaut à un cas de contamination. En effet, une même personne peut avoir eu plusieurs fois la Covid.

- HOSPI

Bien que quasiment un Français sur deux ait été touché par la Covid, la proportion de personnes hospitalisées en soins critiques et nécessitant une surveillance continue est bien plus faible. En effet, cette situation concerne 0,16% de la population

- COVID_LONG

Même si tous les spécialistes ne convergent pas encore vers la même définition d'un « Covid long », 25% des personnes infectées par la Covid présentent encore des symptômes un mois après selon l'OMS. C'est sur cette information que la variable COVID_LONG a été construite.

- RISQUE

Anciennement décomposée en trois classes, la variable RISQUE se voit désormais augmentée d'une classe, la classe « Study_covid ». Cette quatrième classe a pour but de prendre en compte les personnes initialement dans la classe « Oui », donc acceptées dans le risque sans étude complémentaire mais dont les informations relatives à la Covid les feraient changer de catégorie. C'est notamment le cas des personnes présentant toujours des symptômes un mois après le dépistage, ainsi que des personnes ayant été hospitalisées. En effet, il est fort probable qu'une personne ayant été hospitalisée conserve des séquelles un mois après son séjour à l'hôpital.

3.2 Modélisations et résultats

L'objectif de cette partie est de montrer l'impact de la considération de l'information Covid lors de la sélection médicale.

En premier lieu, tout comme dans la partie où la Covid était volontairement mise de côté, différentes méthodes de Machine Learning ont été appliquées sur la base de données construite afin de prédire l'acceptation du risque ou non.

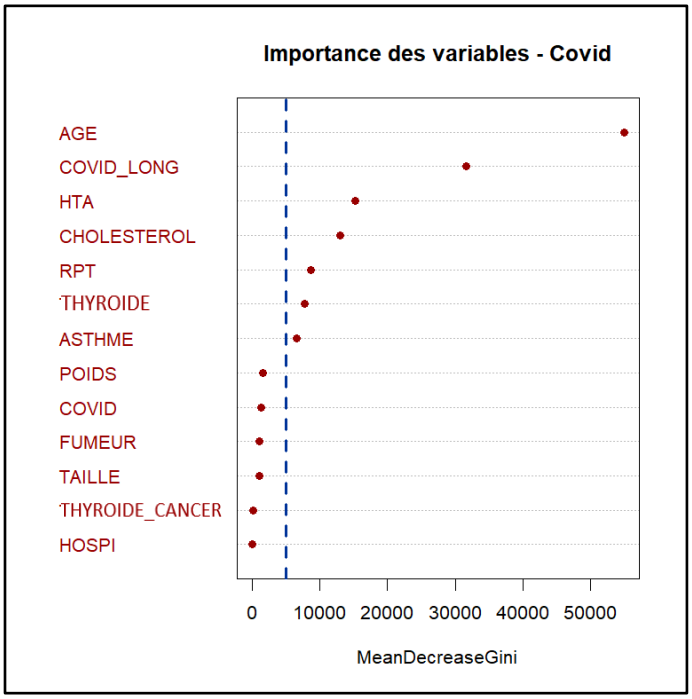
Dans un second temps, un exemple illustré de l'impact de l'intégration de l'information Covid dans la sélection médicale sera proposé.

Découpage de la base de données

Comme dans le cadre des GLM et dans le cas du questionnaire médical sans Covid, un découpage de la base de données a été effectué. Les proportions choisies précédemment ont été conservées. Pour rappel, il s'agissait de répartir 75% des données pour l'apprentissage et les 25% restants pour la base de test.

Importance des variables

Par ailleurs, l'importance des variables a également fait l'objet d'une étude via un algorithme de forêt aléatoire. Une importance particulière a été portée sur la contribution des nouvelles variables créées pour le cas Covid. La sortie R obtenue est présentée ci-dessous.



AGE	COVID_LONG	HTA	CHOLESTEROL	RPT	THYROIDE	ASTHME	POIDS	COVID	FUMEUR	TAILLE	THYROIDE_CANCER	HOSPI
41 971	23 845	11 444	10 346	6 806	5 901	4 988	1 244	1 088	847	867	129	26

Figure 67 - Importance des variables - Avec Covid

Comme dans le cas sans Covid, la variable AGE est celle influant le plus sur le risque. Cependant, à la différence du premier cas, elle est suivie par la variable COVID_LONG. Il est possible de remarquer que l'information Covid est quasiment intégralement portée par cette dernière, les variables COVID et HOSPI étant bien derrière en termes d'importance. Une explication à cela peut être trouvée dans le fait qu'une personne ayant été hospitalisée a de fortes chances de présenter des séquelles un mois après son infection. Par ailleurs, un individu ayant eu la Covid mais ne présentant plus de symptôme au bout d'un mois est considéré comme

un cas standard et ne mérite généralement pas une étude supplémentaire pour être accepté dans le risque.

De plus, bien qu'ayant toujours une importance relativement faible, il est intéressant de souligner la contribution en hausse de la variable THYROIDE_CANCER contrairement à toutes les autres variables déjà présentes qui ont perdu en termes d'importance en raison du nombre de variables plus important.

Enfin, les variables POIDS, TAILLE et FUMEUR conservent leur contribution relativement faible dans la détermination du risque.

3.2.1 Arbres CART

De manière identique, un calibrage des paramètres d'élagage et de pureté a été effectué. L'arbre optimal ainsi obtenu est le suivant. Pour une meilleure lisibilité, le lecteur est invité à se référer à la partie **Annexes**.

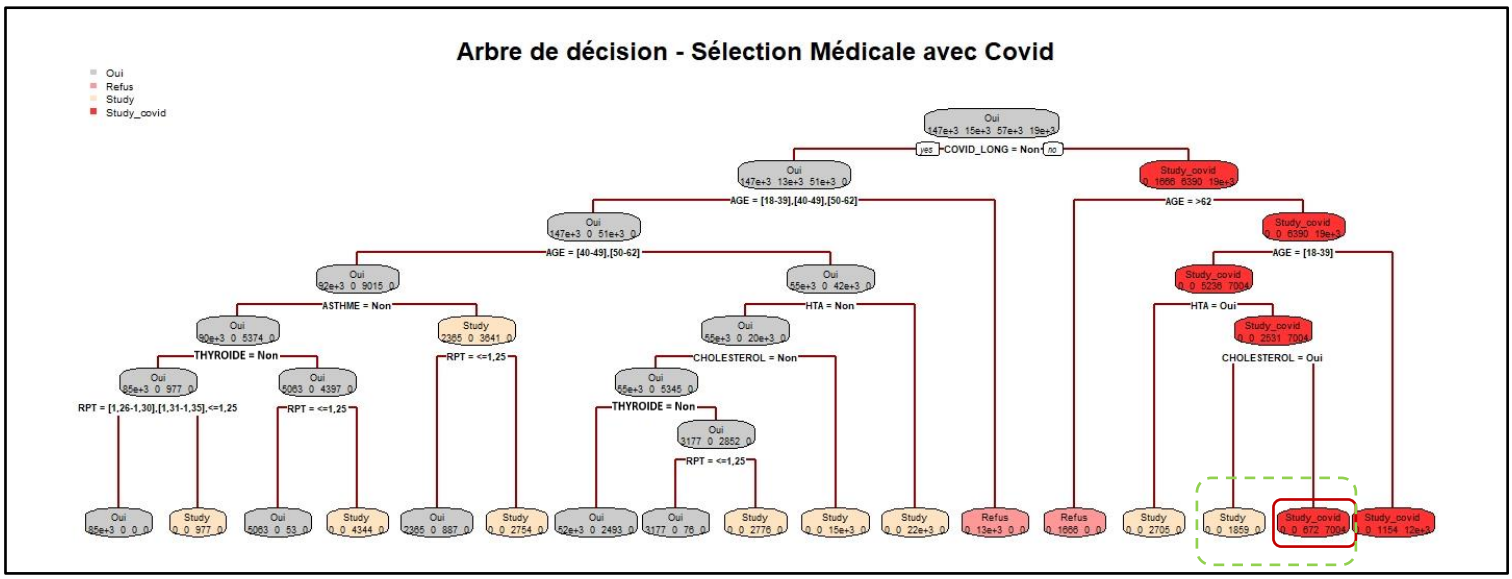


Figure 68 - Arbre optimal de décision - Sélection Médicale avec Covid

En intégrant une quatrième catégorie pour les valeurs possibles de la variable RISQUE, chaque nœud représenté par un rectangle est alors composé de quatre nombres. Les trois premiers correspondent toujours aux mêmes classes déjà mentionnées dans le cas sans Covid, à savoir, en partant de gauche à droite, « Oui », « Refus » et « Study ». Le quatrième nombre le plus à droite correspond quant à lui à la nouvelle classe introduite portant le nom de « Study_covid ».

Par ailleurs, l'attribution de la classe suit la même règle que précédemment. La classe ayant le plus de représentants, donc le nombre le plus élevé, sera attribuée au nœud correspondant.

Concernant l'interprétation visuelle de l'arbre la première chose marquante est la position de la variable COVID_LONG. Cette dernière intervient en premier lieu, telle un premier filtre.

Un individu ayant eu un Covid long aura de très fortes chances d'être classé dans la classe appelée « Study_covid ».

Afin de rendre plus parlante l'interprétation, un exemple a été sélectionné. Il s'agit d'un individu classé dans la catégorie « Study_covid », matérialisé par un rectangle rouge sur l'arbre de décision.

L'individu sélectionné est un individu présentant encore des symptômes de la Covid un mois après son infection. En effet, même si cela est trompeur à première vue, il s'agit d'une personne dont la variable COVID_LONG n'est pas égale à « non ». Autrement dit, une personne pour qui la réponse à la question « N'avez-vous plus de symptômes un mois après votre infection ? », traduction de la phrase « COVID_LONG = non », est non. Par ailleurs, cet individu possède un nombre d'années compris entre 18 et 39 ans, et il n'est atteint ni d'hypertension artérielle ni de cholestérol.

Un second point d'attention mérite d'être mentionné concernant l'interprétation visuelle de l'arbre de décision. Il s'agit du lien conjoint entre la classe « Study » et celle dénommée « Study_covid ». La frontière entre ces deux classes apparaît comme ténue. En effet, la classe « Study_covid » a été construite de telle sorte qu'un individu préalablement classé dans la classe « Study », bien qu'étant atteint de la Covid ne passera pas dans la catégorie « Study_covid ». Au contraire un individu classé dans la catégorie « Oui » et présentant des symptômes d'un covid long basculera dans la nouvelle classe.

Par conséquent, afin de refléter l'impact de l'intégration de l'information Covid dans le questionnaire de sélection médicale, un focus sur la répartition du risque avant et après prise en compte a été effectué.

Oui	Refus	Study	TOTAL
221 371	20 093	76 596	**318 060**

Figure 69 - Répartition du risque avant l'intégration de l'information Covid

Oui	Refus	Study	Study_covid	TOTAL
196 221	20 093	76 596	25 150	**318 060**

Figure 70 - Répartition du risque après l'intégration de l'information Covid

Sur les 318 060 têtes sur lesquelles les modèles ont été construits, avant l'intégration de l'information Covid dans le questionnaire de sélection médicale, 221 371 d'entre elles étaient classées dans la catégorie « Oui », donc acceptées dans le risque. Par ailleurs, 76 596 individus étaient classés dans la catégorie d'étude complémentaire dénommée « Study ».

En prenant en compte le fait d'avoir été atteint ou non de la Covid, il est possible d'observer un basculement de certains individus directement acceptés dans le risque dans la catégorie d'étude propre à la Covid. En effet, dans ce cas, la catégorie « Oui » ne recense plus que 196 221 individus. 25 150 individus ont donc changé de catégorie, soit près de 8% des individus de la base de données.

De plus, comme déjà évoqué précédemment, ce phénomène de bascule ne touche pas directement les personnes classées initialement dans les catégories « Refus » et « Study ».

Toutefois, cela met en lumière que l'adaptation de la sélection médicale permet de déceler des individus à risque plus élevé. Ne pas faire évoluer le système de sélection conduirait donc à augmenter le risque du point de vue de l'assureur.

En parallèle de cette analyse, une approche a été mise en place afin de juger de la qualité des modèles. Cette dernière suit la même logique que celle établie lors du scénario central sans la Covid.

Voici la matrice de confusion associée à l'arbre optimal de décision obtenu.

	data.test_predict			
	Oui	Refus	Study	Study_covid
Oui	49214	0	0	0
Refus	0	4947	0	0
Study	1212	0	17313	618
Study_covid	0	0	0	6211

	Oui Prédit	Refus Prédit	Study Prédit	Study_covid Prédit
Oui Réel	49 214	0	0	0
Refus Réel	0	4 947	0	0
Study Réel	1 212	0	17 313	618
Study_covid Réel	0	0	0	6 211

1

2

Figure 71 - Matrice de confusion - Arbre CART avec Covid

L'erreur de prédiction associée au modèle et calculée grâce aux coefficients de la matrice de confusion est de 2,30%.

Il est également important de garder en tête, que l'erreur au sens de 1 (confère Figure 71) sur la matrice de confusion est bien différente de l'erreur au sens de 2. En effet, l'erreur au sens de 1 pourrait avoir des conséquences bien plus importantes pour l'assureur, alors que dans le cas 2, que l'individu passe dans la catégorie étude pour Covid ou non, cela engendrera tout de même une étude approfondie.

Par ailleurs, afin de juger de la qualité du modèle sans se préoccuper des questions de surapprentissage, une étape de cross-validation a été réalisée.

```
CART

238545 samples
     7 predictor
     4 classes: 'Oui', 'Refus', 'Study', 'Study_covid'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 214690, 214690, 214690, 214691, 214691, 214690, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.07462507  0.7969442  0.58324628
  0.14475096  0.7353834  0.42913631
  0.20810071  0.6250220  0.03170205

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.07462507.
```

Figure 72 - Résultats de la cross-validation - Arbre CART avec Covid

Comme dans le cas sans Covid, les résultats de la sortie R affichent deux métriques : l'accuracy et le k de Kappa. Les valeurs obtenues sont légèrement plus faibles que précédemment. Alors que les deux métriques prenaient respectivement pour valeur 89,7% et 74,2%, elles sont ici aux alentours de 79,7% et 58,3%.
Une justification permettant d'expliquer cette baisse de précision au sein du modèle est peut-être le fait que le modèle est plus complexe que précédemment. Il s'agit non plus d'une variable comportant trois classes à prédire mais d'une variable comportant quatre classes. Par ailleurs,

d'autres variables ont été ajoutées permettant certes d'enrichir le modèle mais aussi de le complexifier.

Une bonne façon d'améliorer le pouvoir prédictif du modèle est alors de regarder les résultats obtenus lors de l'utilisation d'échantillons bootstrappés.

3.2.2 Forêts aléatoires

La première étape effectuée dans le cadre des forêts aléatoires concerne le calibrage des paramètres.
Bien que ce dernier ait déjà été effectué lors de la première partie sans l'intégration de l'information Covid, une actualisation est nécessaire. En effet, le modèle étant différent, il n'y a aucune raison pour que le paramétrage soit similaire. C'est effectivement ce qu'il est possible d'observer quant à l'évolution de l'erreur OOB en fonction du nombre d'itérations.

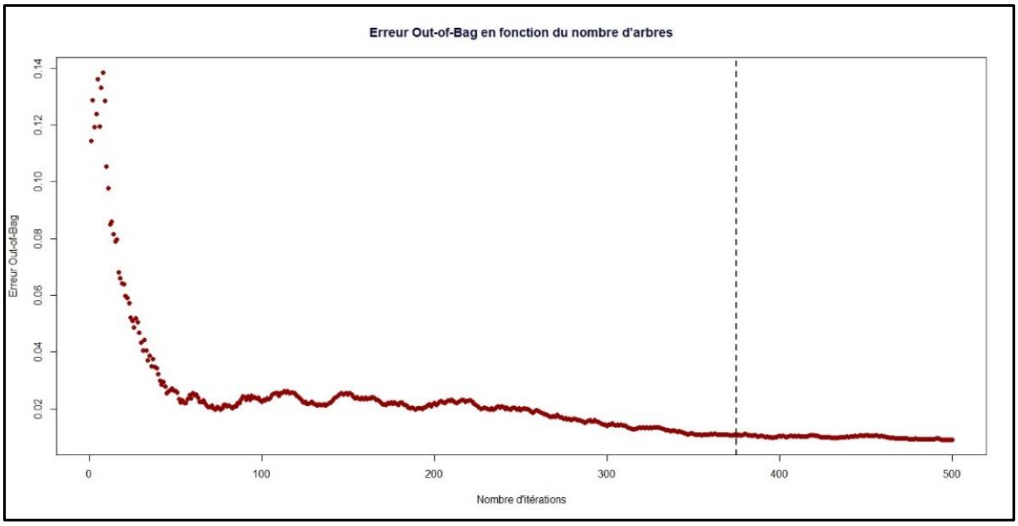


Figure 73 - Evolution de l'erreur OOB en fonction du nombre d'itérations - Avec Covid

Le graphique ci-dessus révèle que l'erreur se stabilise à partir de 375 itérations. Pour rappel, dans le cas précédent, il avait été décidé de choisir un paramètre d'itération ayant une valeur fixée à 200.

Une fois le paramétrage effectué, le modèle est alors prêt à être lancé et la matrice de confusion à être étudiée.

	forest_predict_2			
	Oui	Refus	Study	Study_covid
Oui	49008	0	18	0
Refus	0	5204	0	0
Study	685	0	18395	65
Study_covid	0	0	1	6139

	Oui Prédit	Refus Prédit	Study Prédit	Study_covid Prédit
Oui Réel	49 008	0	18	0
Refus Réel	0	5 204	0	0
Study Réel	685	0	18 395	65
Study_covid Réel	0	0	1	6 139

Figure 74 - Matrice de confusion - Forêt aléatoire avec Covid

Comme esperé, les résultats obtenus en termes de prédiction sont relativement meilleurs que dans le cas des arbres de décision. Le bagging par l'intermédiaire d'apprenants faibles a donc permis d'améliorer le pouvoir prédictif du modèle.

L'erreur obtenue grâce à l'algorithme des forêts aléatoires et selon les coefficients de la matrice de confusion est ici de 0,97%.

Les résultats de la cross-validation permettent eux aussi d'obtenir des erreurs de prédiction du même ordre de grandeur comme le montre le tableau ci-dessous.

K	Erreur
1	0,67%
2	0,93%
3	1,62%
4	0,96%
5	1,34%

Figure 75 - Résultats Cross-Validation - Forêts aléatoires - Avec Covid

Bien que les résultats obtenus se soient déjà nettement améliorés par rapport à ceux obtenus grâce aux arbres CART, l'étape de Gradient Boosting Machine n'a pour autant pas été négligée.

3.2.3 Gradient Boosting Machine

De nouveau, le paramétrage du modèle a été effectué en soutien d'une étape de cross-validation.

Tout comme précédemment, le modèle ayant été complété par de nouvelles variables, une étape de calibrage était donc nécessaire.

Voici la sortie R des résultats obtenus.

```
   nrounds max_depth eta gamma colsample_bytree min_child_weight subsample
11     100         5 0.1     0              0.5                1         1
```

Figure 76 - Paramètres optimaux obtenus par cross-validation - GBM avec Covid

Il est possible de remarquer que plusieurs paramètres ont changé de valeur. Parmi eux, la variable *max_ depth* qui prend ici une valeur égale à 5 contre 3 précédemment, ainsi que la variable *colsample_ bytree* qui prend la valeur 0,5 contre 0,6 dans le cas central.

Pour rappel, la variable *max_ depth* permet de fixer la profondeur de l'arbre, et la variable *colsample_ bytree* permet quant à elle de déterminer le rapport de sous-échantillons des colonnes lors de la construction de l'arbre.

Une fois calibré, le modèle optimal a pu être lancé.

```
Confusion Matrix and Statistics

          Reference
Prediction      Oui Refus Study Study_covid
  Oui         58843    23     0          0
  Refus           0  6004    23          0
  Study           0   599 20260       2119
  Study_covid     0     0    55       7490

Overall Statistics

               Accuracy : 0.9705
                 95% CI : (0.9694, 0.9715)
    No Information Rate : 0.6167
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9468

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Oui Class: Refus Class: Study Class: Study_covid
Sensitivity              1.0000       0.90613       0.9962            0.77948
Specificity              0.9994       0.99974       0.9638            0.99936
Pos Pred Value           0.9996       0.99618       0.8817            0.99271
Neg Pred Value           1.0000       0.99304       0.9989            0.97589
Prevalence               0.6167       0.06944       0.2132            0.10071
Detection Rate           0.6167       0.06292       0.2123            0.07850
Detection Prevalence     0.6169       0.06317       0.2408            0.07907
Balanced Accuracy        0.9997       0.95293       0.9800            0.88942
```

	Oui Prédit	Refus Prédit	Study Prédit	Study_covid Prédit
Oui Réel	58 843	23	0	0
Refus Réel	0	6 004	23	0
Study Réel	0	599	20 260	2 119
Study_covid Réel	0	0	55	7 490

Figure 77 - Matrice de confusion & Kappa et Accuracy - GBM avec Covid

Les valeurs des deux métriques obtenues permettent de valider le bon pouvoir prédictif du modèle. En effet, l'accuracy et le k de Kappa, encadrées en rouge sur la sortie R, ont des valeurs proches de 1.
Cependant, comme dans le cadre des arbres de décision et des forêts aléatoires, les valeurs obtenues sont légèrement plus faibles que dans le scénario central. Cela est également confirmé par l'erreur calculée grâce à la matrice de confusion qui est de 2,95%.

Un récapitulatif des erreurs obtenues en fonction des différents modèles est présenté ci-dessous.

Méthode	Erreur
Arbre CART	2,30%
Forêts Aléatoires	0,97%
GBM	2,95%

Figure 78 - Tableau récapitulatif des erreurs en fonction de la méthode choisie - Cas avec Covid

3.3 Etude de cas : un individu accepté basculé dans la classe d'étude Covid

Dans le but de compléter cette partie sur l'adaptation du questionnaire de sélection médicale en prévoyance individuelle, et de bien saisir les subtilités que peut apporter l'information Covid lors de la sélection, une étude de cas va être proposée. Cette dernière sous-partie, légèrement moins longue que les sous-parties précédentes, aura pour objectif de présenter les caractéristiques d'un invidu classé dans la catégorie « Oui » avec la méthode de sélection classique, et basculant dans la catégorie « Study_covid » avec la nouvelle méthode.

Afin de ne pas accentuer l'effet de la Covid sur la sélection médicale, un individu aux caractéristiques « standards » a été choisi.

Individu	
Age	52
Taille	162
Poids	58
RPT	1,22
Fumeur	Non
Asthme	Non
Tyroïde	Non
HTA	Oui
Cholestérol	Oui

Figure 79 - Caractéristiques d'un individu pour l'étude de cas

Il s'agit d'un individu âge de 52 ans donc appartenant à la classe $[50; 62]$ présentant un rapport poids taille inférieur à 1,25. Il s'agit donc d'un individu ni excessivement jeune ni excessivement âgé, et ni en obésité ou anorexie. Cette personne est également non fumeuse, mais atteinte d'hypertension artérielle et de cholestérol.
Toutefois, ces deux pathologies ne semblent pas être déraisonnables. L'âge de l'individu étant notamment un facteur naturel permettant d'expliquer le cholestérol.

Lorsque le premier algorithme de sélection médicale est lancé, le risque prédit pour cet individu est « Oui ». Ce dernier est donc accepté sans étude complémentaire.

```
  1
Oui
Levels: Oui Refus Study
```

Figure 80 - Etude de cas - Sélection médicale sans Covid

Toutefois, s'il est supposé que cette personne ait été atteinte de la Covid et présente encore quelques symptômes un mois après son infection, le risque prédit devient alors « Study_covid ». Et ce même si ladite personne n'a pas été hospitalisée.

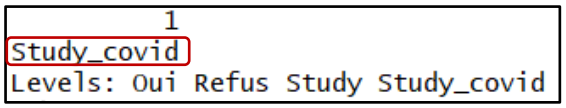


Figure 81 - Etude de cas - Sélection médicale avec Covid

Plusieurs explications à cette nouvelle catégorisation sont possibles.

La première, et sans doute la plus est évidente, est la présence de symptômes de la Covid un mois après alors même que la personne n'a pas été hospitalisée. Or aujourd'hui, de nombreuses études révèlent que des personnes présentent un « Covid long ». Parmi elles, une étude de l'Institut Pasteur souligne que le phénomène appelé « Covid long » toucherait près de 20% des patients cinq semaines après le dépistage de la Covid. Dans ces cas-là, les patients peuvent souffrir de fatigue sévère mais aussi de troubles cardiaques et respiratoires sévères, qui pourraient engendrer des arrêts de travail de longues durées et par conséquent, augmenter de manière imprévue la sinistralité de l'assureur. L'anticipation et la connaissance préalable de ces pathologies devient donc primordiale pour l'assureur, une façon d'y remédier restant la sélection médicale.

Pour davantage d'informations au sujet de l'article de l'Institut Pasteur, le lecteur pourra se référer au lien ici [8].

Une deuxième explication, propre à l'individu choisi ici, est le lien étroit qui existe entre hypertension artérielle et prévalence de la Covid.
Là encore, certaines études ont démontré que les personnes atteintes d'hypertension artérielle avaient un taux d'incidence relatif à la Covid plus important notamment dans le cas particulier des formes sévères de la maladie. Ces derniers doivent alors faire l'objet d'une surveillance accrue. En effet, des formes sévères de la maladie augmenteraient le risque d'arrêts de travail longs et donc pourraient entrainer des conséquences importantes en termes de sinistralité pour l'assureur.

Or, avoir de l'hypertension artérielle à 50 ans est plutôt répandu de nos jours. La preuve en est que l'individu sélectionné bien qu'en étant atteint, avait été accepté dans le risque sans étude complémentaire.
La pathologie Covid combinée à une autre pathologie classique, transforme donc un risque « standard » et connu en un risque méconnu et devant faire l'objet d'une étude supplémentaire.

Enfin, bien que n'étant plus des caractéristiques propres à l'individu de l'étude de cas, d'autres pathologies préexistantes peuvent conduire à prolonger les conséquences de la Covid. Parmi elles, il semblerait que le diabète ou encore des maladies liées à la thyroïde accentuent la persistance des symptômes.

3.4 Conclusion

L'objectif de cette partie était de proposer un nouveau modèle de prédiction du risque en intégrant notamment plusieurs variables liées à la Covid.
Grâce à des méthodes de Machine Learning lancées sur la base de données reconstituée, le risque a pu être prédit de manière satisfaisante, les résultats obtenus étant propres à la base de données artificielle.

Ne s'arrêtant pas à la prédiction seule d'un nouveau risque, cette partie a permis de se questionner sur les corrélations entre les différentes pathologies.
Il est ainsi possible de conclure que l'apparition de nouvelles maladies et qui plus est d'une pandémie ne peut être omise dans la détermination du risque pour l'assureur. Plusieurs possibilités pour anticiper ces conséquences sont alors possibles : la prise en compte en amont de la souscription grâce à l'appui de la sélection médicale comme cela a été détaillé jusqu'à présent ou bien le recours à un processus interne d'évaluation des risques : le scénario ORSA.

4 Scénario ORSA : un autre moyen d'anticiper les conséquences d'une pandémie sur la sinistralité

Si ces dernières années les assureurs ont été confrontés à l'apparition d'une pandémie portant le nom de Covid, l'évolution du questionnaire de sélection médicale n'est pas l'unique manière de prévenir d'éventuelles dérives de sinistralité.

En effet, il devient intéressant de réfléchir au cas où les conséquences d'un tel risque ne seraient plus directement portées par les assurés, via la sélection médicale ici, mais à présent par l'entreprise elle-même. Un moyen d'anticiper ces conséquences du point de vue de l'assureur est alors de mettre en place un scénario ORSA.

4.1 Quelques aspects théoriques

L'ORSA ou Own Risk and Solvency Assessment en anglais, est un processus d'évaluation interne des risques. Cette évaluation prospective des risques de l'entreprise s'inscrit dans le 2ème pilier de la directive européenne Solvabilité II dédiée à la gouvernance.

Le rapport ORSA est établi chaque année par l'organisme d'assurance. Il permet l'évaluation des risques en fonction de la tolérance aux risques de l'entreprise et ce de manière prospective.

De plus, ce dernier ne vise pas forcément à couvrir des scénarios extrêmes, avec des probabilités d'occurrence faibles, telles que 0,5% par an, mais plutôt des probabilités d'occurrence entre 5 et 15% par an.

En théorie, il s'agit d'un processus très libre dans sa mise en œuvre, mais trois évaluations doivent impérativement être respectées.

- Le Besoin Global de Solvabilité aussi appelé BGS
- L'évaluation du respect permanent des obligations règlementaires
- L'évaluation de l'écart entre le profil de risque et les hypothèses sous-jacentes à la formule standard

Le Besoin Global de Solvabilité est établi en fonction des risques majeurs y compris ceux qui ne sont pas pris en compte dans le SCR (Solvency Capital Requirement).

Concernant le second point et l'évaluation du respect permanent des obligations réglementaires, l'assureur projette de manière prospective les éléments prudentiels pour vérifier qu'il couvrira toujours le SCR même en cas de scénarios choqués.

Enfin, le dernier point, permet de valider l'utilisation de la formule standard ou du modèle interne le cas échéant.

Par ailleurs, l'ORSA n'est pas un simple outil de reporting prudentiel mais bien un réel outil de pilotage pour l'organisme d'assurance.

4.2 Modélisations et résultats

Comme cela a été évoqué lors de la partie introductive, l'apparition de la Covid n'est pas sans conséquences pour les sociétés d'assurance les poussant à mener de nombreuses réflexions autour de leur profil de risque et notamment de l'ORSA.

Parmi les scénarios envisageables, un semble répondre parfaitement à l'évaluation des risques liés à une pandémie. Il s'agit du **scénario Pandémie** de l'ORSA.

Au moyen de plusieurs années de projection, qui seront au nombre de quatre ici, le scénario choisi permet de proposer des simulations de chocs pandémiques, mettant ainsi en lumière la capacité ou non de l'entreprise à faire face aux conséquences d'une pandémie.

Dans le but d'aboutir à la construction d'un tel scénario, l'assureur se voit alors obligé de réfléchir à différentes composantes pouvant venir impacter son compte de résultat dans le scénario central.

Dans le cas de la Covid et du portefeuille SwissLife, différents points d'impact, dont la liste non exhaustive est présentée ci-après, ont été choisis dans le but de bâtir un scénario anticipant les conséquences d'une pandémie.

En premier lieu, il a été observé que dans le cas de la Covid, la gestion des sinistres, plus lourde qu'à l'accoutumée du fait des arrêts courts liés à la pandémie, a été favorisée au détriment de la production. Ceci a notamment eu un effet négatif sur la production et par conséquent, également sur le chiffre d'affaires de l'assureur.

Le second point observé lors de la pandémie fut que bien que le nombre d'arrêts de travail ait fortement augmenté, ce dernier concernait majoritairement des arrêts de courtes durées, durées souvent inférieures à la franchise définie dans le contrat. Bien évidemment, dans le cas de certains assurés touchés plus sévèrement, des arrêts de travail de longues durées ont été observés.

Ainsi, ce constat permet de souligner le fait que la Covid a finalement « peu » impacté le ratio S/P relatif à la prévoyance. Contre-intuitivement, bien que le fait générateur soit un arrêt de travail, les conséquences de celui-ci sont davantage portées sur la santé que sur la prévoyance. En effet, dans le cas le plus général, une personne sera amenée à voir sa consommation de frais de santé augmenter dans le but de subvenir aux dépenses de soins « standards » telles que des médicaments contre la toux, le rhume ou encore la fièvre par exemple. Ceci contribue alors à une augmentation des frais de santé à court terme.

Cependant, cette augmentation peut aussi être amenée à perdurer dans le temps. Comme exposé lors de la partie destinée à l'amélioration du questionnaire de sélection médicale, la Covid peut entrainer des dérives de sinistralité liées en partie à la dégénérescence de certaines pathologies « classiques » des suites de la pandémie comme l'asthme. Dans ce cas, des frais de santé à plus long terme seront alors nécessaires.
Bien évidemment, ce constat est à relativiser dans le cas de covid longs par exemple.

Pour information, les effets « bénéfiques », tels qu'un éventuel confinement et ses conséquences sur le S/P santé, n'ont pas été modélisés ici.

Ainsi listés, ces éléments ont permis la construction de deux sous-scénarios dits de pandémie pour le rapport ORSA.

Le premier d'entre eux, permet de mettre en lumière l'impact de la Covid sans modification de la méthode de sélection médicale.

A l'inverse, le second scénario Pandémie proposé ici, présente quant à lui l'impact de la Covid dans le cas où la méthode de sélection médicale est revue au bout de deux années de projection.

L'application proposée ici dans le cadre du portefeuille SwissLife se déroule sur quatre années de projection, communément appelées budget 2023, plan 2024, plan 2025 et plan 2026.

Par ailleurs, dans un souci de confidentialité, aucun chiffre ne sera directement exposé. Les chiffres obtenus dans les graphiques suivants ont été volontairement erronés de manière à garder des variations cohérentes.

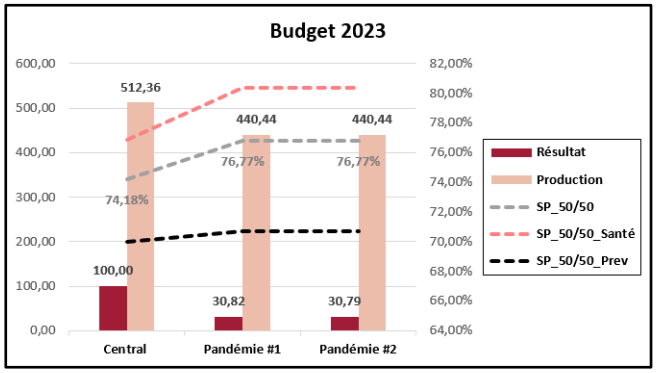


Figure 82 - Conséquences du scénario Pandémie – Horizon 2023

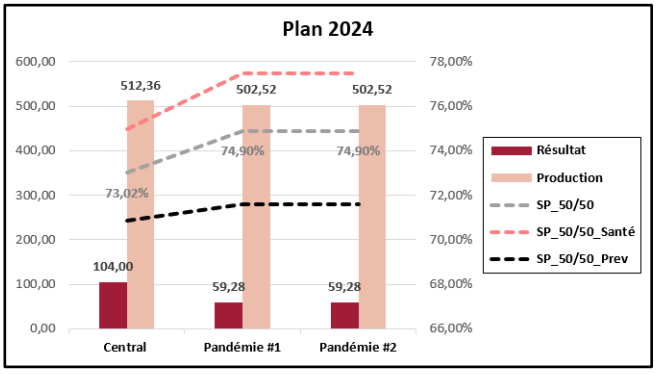


Figure 83 - Conséquences du scénario Pandémie – Horizon 2024

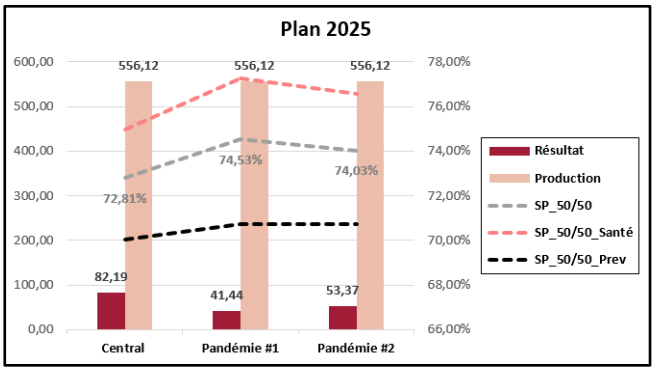


Figure 84 - Conséquences du scénario Pandémie – Horizon 2025

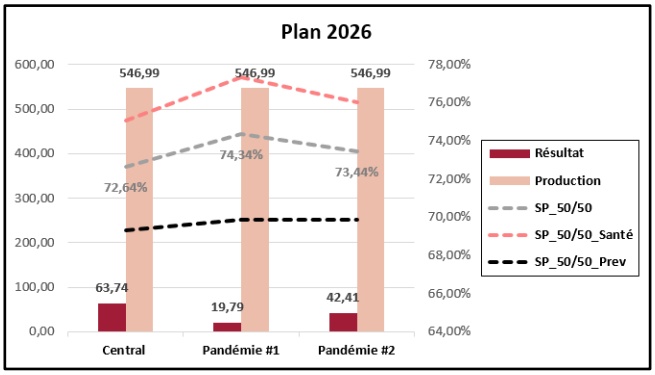


Figure 85 - Conséquences du scénario Pandémie – Horizon 2026

Le premier point d'observation et non des moindres, est que l'apparition d'une pandémie telle que la Covid par exemple, entraine une diminution significative du résultat de l'assureur au cours des quatre années de projection modélisées ici. La première année, une variation de près de 70% de ce résultat avant impôt est observable entre le scénario central et les scénarios choqués.

Cette baisse significative peut être expliquée en partie par la baisse, elle aussi significative, de la production et donc du chiffre d'affaires.
Le budget 2023 met ainsi en lumière une baisse de la production de l'ordre de 14% entre le scénario central et les scénarios Pandémie. Cette même baisse se stabilisant au cours de la deuxième année de projection autour de 6%, avant d'observer un retour à la normale lors du plan 2025.

La deuxième explication à la baisse significative du résultat de l'assureur est fortement liée à l'augmentation de la sinistralité. Cette augmentation est majoritairement portée par l'augmentation du ratio S/P lié à la santé. Comme évoqué précédemment, bien que le fait générateur soit un arrêt de travail, donc un évènement lié à la prévoyance, une forte hausse des dépenses de soins reste à prévoir.

A partir du plan 2025, deux options modélisées ici s'offrent alors à l'assureur.
Soit ce dernier décide de conserver son mode de sélection médicale, en prenant le risque de voir certains profils comportant des maladies classiques se détériorer au cours du temps avec l'apparition de la Covid. Ce scenario porte ici le nom de **« Pandémie #1 »**.
Soit, dans le cas contraire, l'assureur peut alors décider de repenser son mécanisme de sélection médicale en intégrant de nouvelles informations relatives à la Covid, et ainsi espérer réduire d'année en année son risque de dérives de sinistralité sur les affaires nouvelles. Ce scénario est ici nommé **« Pandémie #2 »**.

L'analyse des plans 2025 et 2026 permet de mettre en lumière l'atout du deuxième scénario. En effet, en revoyant son mécanisme de sélection médicale, l'assureur a ainsi la possibilité de progressivement se rapprocher de la trace du ratio de sinistralité relatif à la santé du scénario central. L'écart de S/P entre le scénario central et le scénario « Pandémie #2 » n'étant plus que 0,8% dans le plan 2026, laissant imaginer que cet écart disparaîtrait au fur et à mesure que le stock d'assurés serait renouvelé par le biais du nouveau système de sélection médicale.

Evidemment, repenser le mécanisme de sélection médicale n'est pas le seul levier à la disposition de l'assureur. En effet, dans l'objectif de redresser son résultat, l'assureur pourrait aussi avoir recours à une indexation tarifaire. Cependant, cette option amènerait la question de l'équilibre tarifaire entre la couverture santé et la couverture prévoyance de l'assuré. En effet, un assuré aura tendance à devoir payer plus cher sa cotisation santé du fait du tarif « libre », à l'inverse de la couverture prévoyance qui est souvent un pourcentage du salaire.

Toutefois, l'analyse du scénario Pandémie proposée ici permet de mettre en lumière le fait que bien que les confinements successifs aient permis d'endiguer la propagation de la Covid-19 au sein de la population française, ces derniers ne se révèlent pas sans conséquences pour les assureurs, notamment en termes de chiffre d'affaires.
Ces scénarios sont d'autant plus importants qu'ils peuvent être déterminants dans le choix des traités de réassurance et notamment des traités de couverture Pandémie.

Par ailleurs, du fait de sa définition et de ses divers scénarios, l'ORSA a l'avantage d'offrir à l'assureur la faculté de démultiplier ces derniers à d'autres pandémies que la Covid, permettant ainsi des discussions plus élargies.

Enfin, l'impact de la Covid aurait également pu être vu sous le prisme du ratio de solvabilité et notamment au travers du SCR Catastrophe. Cependant, une faiblesse du modèle Solvabilité II est la faible prise en compte du choc Pandémie en santé. Une maladie telle que la Covid, touchant certes les actifs mais sur une période relativement courte et avec peu de séquelles ne constitue pas un scénario suffisamment craint par les assureurs.

Conclusion

Mise en place pour l'ensemble des contrats en prévoyance individuelle, la sélection médicale permet à l'assureur de prévenir l'antisélection. Jouant ainsi un rôle primordial dans la connaissance du risque, elle se place alors sur le devant de la scène au cœur de multiples discussions actuarielles. Parmi elles, des questionnements autour de son incidence sur la sinistralité de l'entreprise.

Au moyen détourné de diverses questions, le questionnaire médical permet à l'assureur d'établir pour chaque assuré un profil de risque propre à son état de santé et ce au moment de la souscription.
Les travaux menés dans ce mémoire ont ainsi permis de confirmer l'impact de cette étape de sélection sur la sinistralité. En effet, ce mécanisme, en filtrant les risques choisis, offre la possibilité de rapprocher la sinistralité du portefeuille étudié à une sinistralité bien inférieure à la normale.
Toutefois, au bout de deux à trois années d'ancienneté, ce phénomène s'essouffle. Les ratios de morbidité alors mis en place dans les modèles traduisent des valeurs supérieures à un.

Sous l'hypothèse d'une sinistralité accrue au bout de quelques années, l'assureur peut alors décider de mettre en place une revalorisation. Par l'intermédiaire de celle-ci, ce dernier peut ainsi réajuster son tarif tout en gardant un mécanisme de levier quant à la sinistralité.
Dans les travaux menés, cette étape de revalorisation a été agencée grâce aux modèles linéaires généralisés. Bien que ne prenant pas en compte les valeurs extrêmes et se limitant à l'utilisation d'une famille exponentielle, les GLM permettent flexibilité et simplicité d'utilisation.
En définitive, cette étape de revalorisation offre la possibilité de soumettre de nouveau chaque assuré à une seconde sélection médicale détournée.

Cependant, dans un monde en perpétuelle évolution, et bien qu'ayant déjà une incidence matérielle sur la sinistralité, la sélection médicale ne peut se permettre d'être figée dans le temps. Elle doit sans cesse être réévaluée dans le but de maintenir à flot un niveau de sélection compétitif.
Un des derniers enjeux en date venu bousculer l'ordre établi en matière de sélection médicale est l'apparition de la Covid-19.
Avec cette dernière, le monde de l'assurance a vu son nombre d'arrêts de travail liés à cette pathologie augmenter. En parallèle de ces arrêts de travail, de nombreuses incertitudes persistent quant aux conséquences à long terme de la maladie.
De ce fait, les organismes d'assurance sont poussés à se questionner sur de nouvelles problématiques, les obligeant à revoir leurs mécanismes établis de sélection médicale.

Les travaux menés ont permis de souligner différentes conséquences qu'engendrerait la mise sous silence de l'information Covid au sein du questionnaire médical.
En faisant appel à diverses méthodes de Machine Learning, il a notamment été possible de mettre en lumière l'effet néfaste d'une telle maladie sur des pathologies initiales somme toute classiques et connues. Négliger les informations lors de la sélection médicale, ici relatives à la Covid, conduirait l'assureur à amplifier l'effet d'antisélection, effet qui a justement pour but d'être amoindri par celle-ci.

Les méthodes de Machine Learning utilisées dans ce mémoire ont permis d'obtenir des résultats probants. Néanmoins, il faut souligner que ces résultats restent propres à la base de données utilisée. Une autre base de données aurait probablement conduit à des conclusions légèrement différentes. De plus, bien qu'étant des méthodes non paramétriques, celles-ci sont très régulièrement sujettes à des problèmes de surapprentissage et d'interprétabilité. Qui plus est, tandis que les arbres de décision permettent une représentation visuelle des résultats, les méthodes GBM et de forêts aléatoires s'apparentent davantage à des boîtes noires, pouvant laisser le lecteur dubitatif s'il est peu initié au sujet.

Dans un second temps, dans le cas où la sélection médicale ne permettrait pas d'anticiper les dérives liées à une nouvelle pathologie, ou qu'elle ne serait pas le moyen choisi par l'organisme pour répondre à ce nouvel enjeu, un scénario ORSA pourra être utilisé.

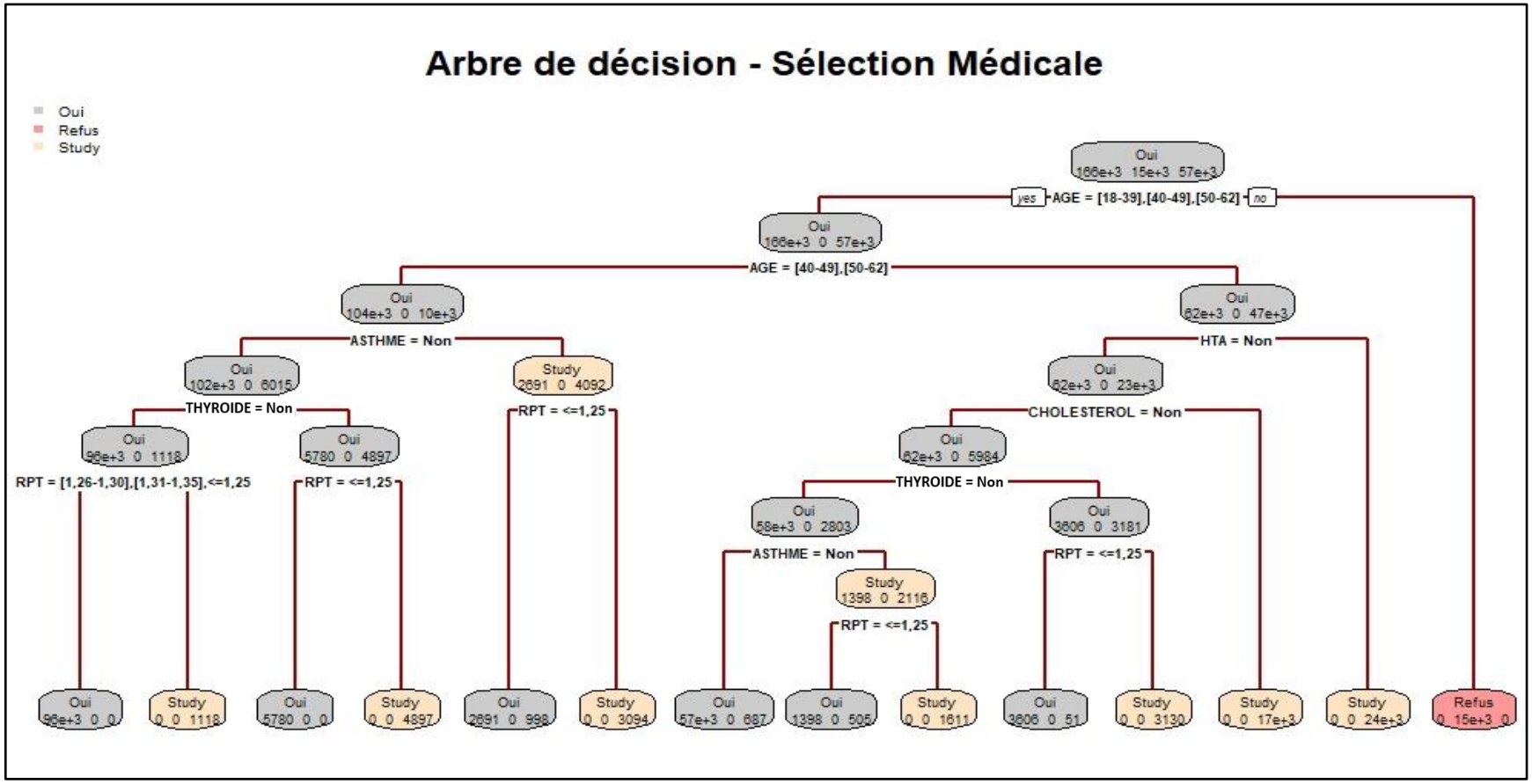
L'ORSA, de par sa faculté à dupliquer divers scénarios offre ainsi la possibilité à l'assureur de généraliser les impacts d'un nouveau risque à différents enjeux. En effet, la pandémie n'est pas le seul risque émergent. Risques cyber, risques climatiques mais aussi enjeux géopolitiques et internationaux devraient être autant de défis pour les assureurs dans les années à venir. Le monde de demain doit se préparer. « La prévoyance ne doit pas stériliser l'avenir : elle doit préparer ses voies » comme l'écrivait le romancier français Paul Bernard à la fin du 19^{e} siècle.

Ces nombreux enjeux pourraient conduire les sociétés d'assurance à procéder de manière indirectement contrainte à une augmentation tarifaire importante dans les années à venir.
Or, pour maintenir leur compétitivité, ces dernières pourraient être amenées à se tourner vers des mécanismes de transfert de risques. Ceci ferait donc émerger de nouvelles discussions quant à la mise en place de dispositifs de réassurance adaptés à la prévoyance.

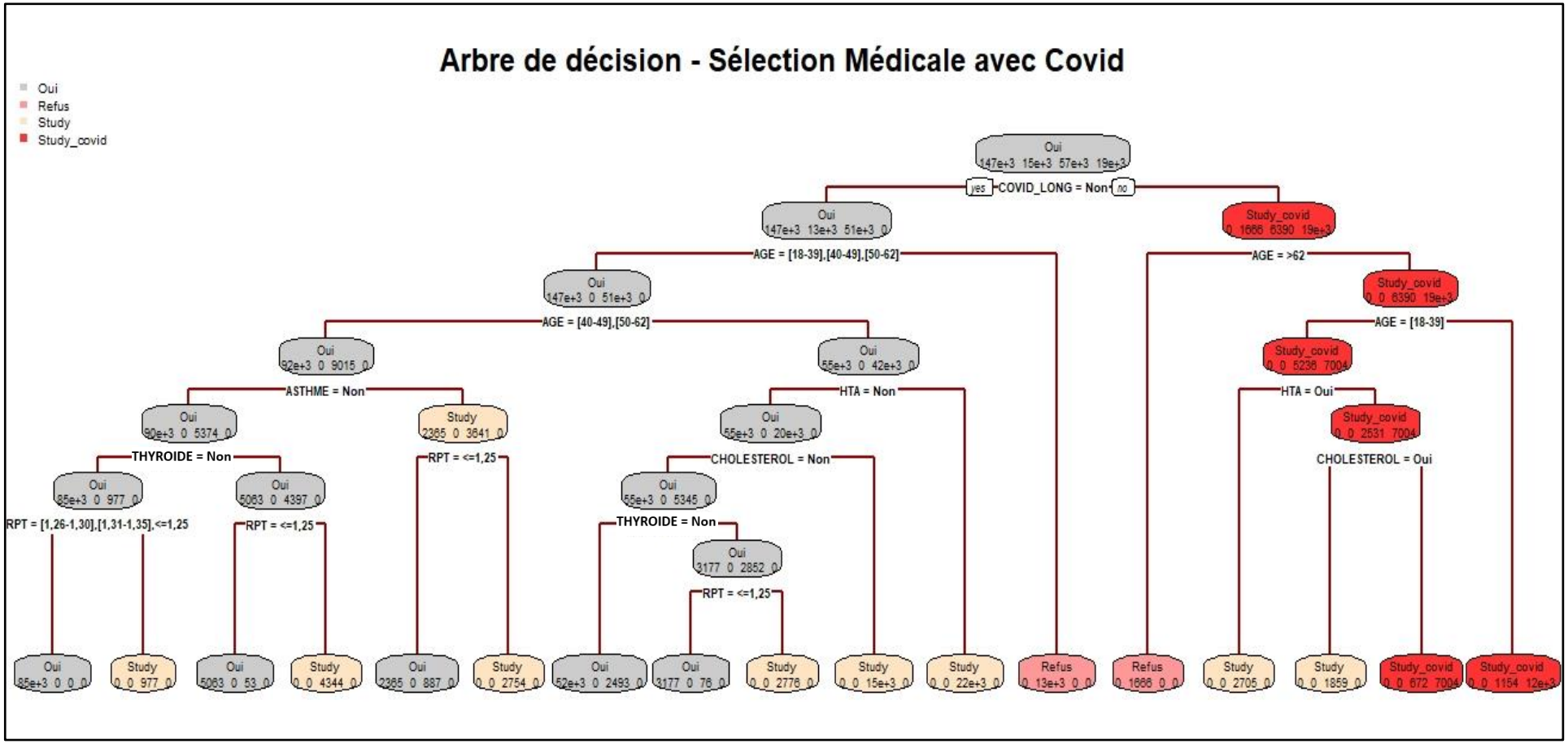
Liste des figures

Annexes

Annexe A : Arbre optimal de décision – Sélection Médicale sans Covid



Annexe B : Arbre optimal de décision – Sélection Médicale avec Covid



Bibliographie

Cours

BOYER C . (2022) *Sciences des données : Machine Learning* (ISUP)

CORNUAILLE C . (2022) *Assurance : Les principes de la prévoyance* (ISUP)

THOMAS M. (2021) *Econométrie de l'assurance non-vie* (ISUP)

Ouvrages

FRIEDMAN J., HASTIE T., TIBSHIRANI R. [2001] *"The Elements of Statistical Learning", 2ème édition. Springer*

MOHRI M., ROSTAMIZADEH A., TALWALKAR A. [2012] *"Foundations of Machine Learning", MIT Press*

MURPHY K.P [2012] *"Machine Learning : A Probabilistic Perspective", MIT Press*

NELDER J., WEDDERBURN R - [1972] *"Generalized linear models"*, Journal of Roy. Stat. Soc. B, vol. 135, 370-384

Sites internet

[1] SMR - ObservatoryEasternRegion Public Health - *Théorie SMR, site consulté le 20 juillet 2022*

[2] GLM - Théorie - *GLM Ressources Actuarielles, site consulté le 22 juillet 2022*

[3] Machine Learning - *Théorie Arbres de décision, site consulté le 29 juillet 2022*

[4] Taux d'incidence des indépendants en arrêts de travail - *Etude loi d'incidence arrêts de travail, site consulté le 26 avril 2022*

[5] Covid Tracker - Loi d'incidence - *Infos Covid Tracker, site consulté le 10 juin 2022*

[6] HTA - *Infos HTA, site consulté le 17 juin 2022*

Thyroïde - *Infos Thyroïde Cancer, site consulté le 17 juin 2022*

Asthme - *Infos Asthme, site consulté le 17 juin 2022*

Fumeur - *Infos Fumeur, site consulté le 17 juin 2022*

Cholestérol - *Infos Cholestérol , site consulté le 17 juin 2022*

[7] Construction de variables Covid - *Infos Covid (1), site consulté le 24 juin 2022*
Infos Covid (2), site consulté le 24 juin 2022

[8] Covid long - Institut Pasteur - *Qu'est-ce que le Covid long ?, site consulté le 19 août 2022*
ORSA - *Article ORSA - GALEA, site consulté le 21 septembre 2022*