UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNES

MÉMOIRE

présenté en vue d'obtenir

MASTER PARCOURS STATISTIQUE POUR L'ÉVALUATION ET LA PRÉVISION

Mention : Mathématiques et applications
Promotion : 2022 - 2023

MODÈLE DE COX POUR LE DIABÈTE DE TYPE I

Léa Pimpernelle

Sous la direction de Monsieur Amor KÉZIOU

Remerciement

Tout d'abord, je tiens à exprimer ma sincère gratitude envers mon directeur de mémoire, Monsieur KÉZIOU, pour ses précieux conseils, sa disponibilité et son écoute attentive qui ont grandement contribué à la réussite de mon travail.

J'aimerais également adresser mes remerciements au co-juré, Monsieur BLANCHARD, pour son vif intérêt à mon travail, le temps qu'il a consacré à lire mon mémoire et sa participation lors de ma soutenance.

Je souhaite exprimer ma reconnaissance envers tous les membres du projet MoGly pour leur écoute, leurs conseils et leur intégration au sein du projet.

Je tiens également à remercier Hélène NOVAKOWSKI pour ses explications détaillées sur le script de traitement de la base de données, la sécurité des données et le chiffrement d'un dossier sur VeraCrypt.

Je tiens aussi à remercier Ousmane FALL pour m'avoir généreusement fourni le script de son mémoire.

Enfin, je souhaite exprimer ma gratitude envers toutes les personnes qui ont pris le temps de m'écouter parler de mon mémoire, pour leurs conseils et leur soutien précieux.

Résumé

Ce mémoire se concentre sur l'application de l'analyse de survie et du modèle de Cox dans l'étude des diabétiques de type 1 et de la complication de la rétinopathie. La rétinopathie est l'une des complications les plus courantes du diabète, et il est crucial de comprendre les facteurs de risque et la progression de cette maladie chez les patients atteints de diabète de type 1. De ce faite, nous pourrons tirer une première conclusion sur la survenue de la rétinopathie chez les diabétiques de type 1.

Abstract

This dissertation concentrates on the application of survival analysis and the Cox model in the study of type 1 diabetes and the complication of retinopathy. Retinopathy is one of the most frequent complications of diabetes, and it is crucial to understand the risk factors and progression of this disease in patients with type 1 diabetes. In this way, we can draw a first conclusion regarding the occurrence of retinopathy in type 1 diabetics.

Liste des abréviation

CARéDIAB (Champage-Ardenne Réseau Diabète)

HbA1c : hémoglobine glyquée ou hémoglobine A1c

IMC : Indice de Masse Corporelle

LDL : Low-Density Lipoprotein

PDS : poids

Sommaire

1. Introduction

Le diabète de type 1 est une maladie chronique qui se caractérise par une destruction des cellules bêta du pancréas, entraînant une incapacité à produire de l'insuline. Le diabète de type 1 affecte environ 10% des personnes atteintes de diabète dans le monde. Selon l'Organisation Mondiale de la Santé, plus de 500 millions de personnes dans le monde souffrent de diabète en 2021 et 4,5 millions en France. Cette maladie est associée à un risque accru de complications graves, telle que la rétinopathie, qui est la principale cause de cécité chez les adultes. De ce fait, ce mémoire traite sur l'apparition de la rétinopathie[1] qui représente l'événement de l'analyse des données de survie.

L'analyse des données de survie est une branche de la statistique qui s'intéresse à l'étude des temps jusqu'à l'apparition d'un événement afin d'évaluer l'efficacité de traitements, d'identifier les facteurs de risque et de prédire la survie des patients.

[1] C'est une complication due au diabète qui s'attaque à l'œil et la rétine. L'excès de sucre dans le sang fragilise la paroi des vaisseaux sanguins de la rétine entraînant une perte d'étanchéité. En conséquence, certaines zones de la rétine ne reçoivent plus de sang (l'ischémie). Ce qui entraine la rupture puis l'éclatement des vaisseaux rétiniens. De nouveaux vaisseaux vont apparaitre, mais seront plus fragiles. Ils pourront être à l'origine de complications graves dont les dommages seront irréversibles et pourront aller jusqu'à la cécité. Les patients atteints de rétinopathie ont une sensation de flou, des taches devant les yeux et sont plus facilement éblouis. Cette complication affecte la vision de près et diminue la capacité de lire, mais aussi la vision de loin, avec l'incapacité de distinguer les formes (agnosie visuelle).
Pour limiter la progression de cette complication, il est important de bien contrôler sa glycémie. Il est également d'avoir recours à un traitement par laser, une injection dans l'œil, ou une intervention chirurgicale. Cette maladie se dégrade progressivement, lorsque le sucre s'accumule dans le sang, les vaisseaux sanguins se fragilisent et créent de petites hémorragies. Ce qui constitue la rétinopathie cependant il est possible d'être atteint d'une rétinopathie diabétique même en ayant une bonne vue et pas de symptômes.

L'histoire de l'analyse des données de survie remonte à plus d'un siècle. En 1935, le mathématicien allemand Emil Gumbel[2] a publié un article sur la distribution des temps de vie des personnes. Ce travail a jeté les bases de l'analyse de survie en introduisant la notion de fonction de survie, qui décrit la probabilité de survie au-delà d'un temps donné.
Dans les années 1950 et 1960, les chercheurs en médecine ont commencé à s'intéresser à l'analyse de survie pour étudier les facteurs de risque associés à la mortalité et à la morbidité dans diverses populations. En 1958, le statisticien anglais David Cox a introduit le modèle de régression de Cox permettant de modéliser les effets des facteurs de risque sur la survie en prenant en compte le temps depuis le début de l'étude.

Au cours des années suivantes, de nombreuses méthodes statistiques ont été développées pour l'analyse de survie, telles que les méthodes non paramétriques (par exemple : la méthode de Kaplan-Meier) et les modèles semi-paramétriques (par exemple : le modèle de Cox).

Au fil du temps, l'analyse de survie est devenue une technique statistique couramment utilisée dans de nombreux domaines, notamment la médecine, la biologie, l'épidémiologie, l'ingénierie et la finance.

En résumé, l'analyse de survie a connu une évolution significative au cours des dernières décennies, passant d'une méthode marginale à une technique statistique essentielle dans de nombreux domaines. Elle continue d'évoluer pour répondre aux besoins des chercheurs et des praticiens dans des domaines toujours plus nombreux.

Dans ce mémoire de méthodologie, la première partie portera sur la base de données de survie. Nous commencerons par expliquer la provenance de la base de données. Nous introduiront la notion de données censurées et les variables explicatives.

---

[2] L'article de Emil Gumbel intitulé "La distribution des temps de vie" a été publié en 1935 dans la revue "Jahrbuch für Nationalökonomie und Statistik". Il s'agit d'un article majeur dans le domaine de la statistique et de l'analyse de survie, où Gumbel a introduit la distribution de Gumbel (également appelée distribution de double exponentielle), qui est une distribution de probabilité continue utilisée pour modéliser les temps de survie. http://archive.numdam.org/article/AIHP_1935__5_2_115_0.pdf

Dans la deuxième partie du mémoire, nous nous concentrerons sur l'analyse de survie. Nous aborderons la fonction de survie, qui est une estimation de la probabilité qu'un individu survive au-delà d'un certain temps. Nous présenterons la courbe de survie de Kaplan-Meier, une méthode non paramétrique utilisée pour estimer la fonction de survie en présence de censure. Nous discuterons également des tests d'hypothèse couramment utilisés pour comparer les courbes de survie entre différents groupes.

Enfin, nous explorerons en détail le modèle de Cox, une méthode de régression semi-paramétrique utilisée pour modéliser l'effet des covariables sur la survie. Nous définirons le modèle de Cox puis le construirons à partir de variables explicatives. Ce modèle sera testé par des tests statistiques. Finalement, nous ferons des courbes individuelles pour l'apparition de la rétinopathie.

2. La base de données de survie

La base de données provient de CARéDIAB[3] est une base de données de suivi de patients atteints de diabète de type 1 depuis leur diagnostic précoce jusqu'à l'âge adulte. Cette base de données a été créée en 1992 par l'association française de diabétologie et de nutrition pédiatriques (AFDNP) pour améliorer la prise en charge du diabète de type 1 chez les enfants et les adolescents. Les données collectées dans CARéDIAB comprennent des informations sur les caractéristiques socio-démographiques des patients, leur histoire médicale, leur traitement, les complications et les résultats des examens de surveillance. Cette base de données est utilisée pour étudier les facteurs de risque de complications du diabète de type 1 comme la rétinopathie.

[3] Site de CARéDIAB : https://reseaux-sante-ca.org

2.1 Base de données CARéDIAB

La base de données CARéDIAB a été remplie par des praticiens médicaux à chaque consultation depuis 20 ans. Elle est constituée de 2 707 patients atteints du diabète de type 1. Le traitement de la base de données a été effectué par Hélène Novakowski pour passer d'une base de données avec une ligne par consultation représentant 63 072 lignes à une ligne par patient. Le travail d'Hélène Novakowski a été conséquent puisqu'il a fallu faire une première sélection de variables, corriger les incohérences, corriger les données aberrantes et bien d'autres.

À partir de son code R, pour cette étude, des colonnes ont été ajouté. Pour les variables du taux de triglycéride[4], l'IMC[5] et le taux d'HbA1c[6], le minimum, le maximum et la variance on était ajouté en parcourant les consultations des patients. Ce qui fait que pour le taux de triglycéride il y a 4 colonnes (la moyenne du taux des consultations, la variance du taux des consultations, le taux minimum des consultations et le taux maximum des consultations). Une fois ce code R compilé, on obtient une base de données exploitable avec une ligne par patient.

[4] Les triglycérides sont une forme de graisse présente dans notre corps et dans de nombreux aliments. Elles sont une source d'énergie essentielle pour notre organisme. Les triglycérides sont composées de trois molécules d'acides gras liées à une molécule de glycérol. Les niveaux de triglycérides dans le sang peuvent varier en fonction de notre alimentation et de notre métabolisme. Il est recommandé de maintenir des niveaux de triglycérides sains en adoptant une alimentation équilibrée, en limitant la consommation d'aliments riches en graisses saturées et en sucres ajoutés, en faisant de l'exercice régulièrement et en évitant la consommation excessive d'alcool.

[5] IMC = poids (en kg) / (taille (en m))^2

[6] L'HbA1c est un examen sanguin utilisé pour évaluer le contrôle à long terme de la glycémie chez les personnes atteintes de diabète. L'HbA1c mesure la quantité d'hémoglobine qui s'est liée aux molécules de glucose dans le sang au cours des deux à trois derniers mois. Lorsque le glucose sanguin est élevé, une partie de celui-ci se lie à l'hémoglobine, formant ainsi l'HbA1c. Des niveaux d'HbA1c élevés indiquent un mauvais contrôle de la glycémie, ce qui peut être préjudiciable à long terme et augmenter le risque de complications liées au diabète. En général, un niveau d'HbA1c inférieur à 7% est souvent recommandé pour les personnes atteintes de diabète.

2.2 Données censurées

Dans le cadre de cette étude, les patients qui constituent cette base de données n'ont pas forcément la rétinopathie, complication liée au diabète de type 1. Ces patients forment la censure. Une durée de vie est dite censurée si la durée exacte n'est pas encore connue.

La censure est appelée dans ce cas, la censure à droite. Ce qui signifie qu'à la période de fin de suivi, l'évènement d'intérêt ne s'est pas encore produit. Ainsi, on ne connaît donc pas la durée de vie X, mais que, $X > t$, où t représente le temps, entre la date de détection du diabète de type 1 du patient à la fin de l'étude. Ces informations sont essentielles pour l'étude et doivent être incluse dans la base de données exploitées.

2.3. Formalisation de la base de données

Comme déjà mentionné, la durée de vie est assimilée à la réalisation d'une variable aléatoire X, continue, qui prend des valeurs uniquement positives. En présence de censure à droite, on considère la variable latente C, correspondant à la durée écoulée avant la censure de l'information.

Dès lors, à l'issue de la collecte des données, nous disposons, pour chaque sujet, des données suivantes :

$$T_i = \min(X_i, C)$$
$$\delta = 1_{\{X \geq C\}}$$

Or, l'estimation dans le cas des modèles de survie, s'appuie sur la vraisemblance statistique. Sa maximisation en présence de censure suppose de considérer que les processus de durée X et de censure C sont indépendants. Dans le cas de cette étude, si la censure est due à l'arrêt du traitement, ou si les patients les plus malades ne sont plus suivis. À l'inverse, ce n'est pas le cas, si la censure est liée à la fin de l'étude ou occasionnée par un déménagement sans lien avec l'état de santé du sujet. Il y a en annexe un tableau regroupant la base de données finale pour l'analyse de survie.

3. L’analyse de survie

L'analyse de survie est une méthodologie statistique utilisée pour étudier le temps jusqu'à un événement d'intérêt tel que la survenue d'une maladie. Elle permet de modéliser et d'analyser les données de survie en prenant en compte la présence de censures, qui se produisent lorsque certaines observations n'ont pas atteint l'événement d'intérêt à la fin de la période d’observation. Dans l'analyse de survie, deux concepts fondamentaux sont la fonction de survie qui décrit l'évolution de la survie dans le temps; et la fonction de hasard ou fonction de risque instantané représentant le taux de défaillance instantané à un moment donné de la survie.

3.1. Fonctions de survie

La fonction de survie ou probabilité de survie est un concept en analyse de survie. Elle représente la probabilité qu'un événement d'intérêt (la survenue de la rétinopathie chez les diabétique de type 1) ne se soit pas encore produit à un moment donné. En d'autres termes, elle quantifie la proportion d'individus ou de sujets qui survivent au-delà d'un certain point dans le temps.

3.1.1. Fonction de survie S

La fonction de survie est notée : $S(t) = \mathbb{P}(X > t),\ t \geq 0$. C’est la probabilité que l’événement se produise après t, $S(t) \in [0,1]$. La fonction est décroissante, la probabilité à l’origine est égale à 1 ($S(0) = 1$). La probabilité après en temps infini est nulle ($S(\infty) = 0$).

3.1.2. Fonction de répartition F

La fonction de répartition $F(t)$ est la probabilité que l’événement se produise avant t, $F(t) \in [0,1]$. Elle est noté : $F(t) = \mathbb{P}(X \leq t) = 1 - S(t)$.

La fonction est croissante, la probabilité à l'origine est égale à 0 ($F(0) = 0$). La probabilité après en temps infini est égale à 1 ($F(\infty) = 1$).

3.1.3. Densité de probabilité f

Pour t fixé, la fonction f représente la probabilité de l'événement dans un petit intervalle de temps après l'instant t. La fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$:

$$F(t) = \int_0^t f(u)\,\mathrm{d}u$$

Si la fonction de répartition admet une dérivée au point t alors :

$$f(t) = \lim_{h \to \infty} \frac{\mathbb{P}(t \leq X \leq t + h)}{h} = F'(t) = -S'(t)$$

3.2. Estimer la fonction de survie

La méthode de Kaplan-Meier est la plus courante. Elle permet d'analyser le temps jusqu'à l'événement d'intérêt et construire la courbe de survie de Kaplan-Meier.

3.2.1. Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est une méthode non paramétrique[7] utilisée pour estimer la fonction de survie à partir de données de survie censurées à droite. La méthode ajuste les estimations de survie en prenant en compte les individus censurés. Cette méthode divise le temps de participation observé en intervalles et estime la probabilité de survie pour chaque intervalle, ce qui donne à la courbe une apparence « d'escalier".

[7] C'est type de modèle statistique qui ne fait pas d'hypothèses spécifiques sur la forme ou la distribution des données

L'estimateur de Kaplan-Meier découle de l'idée que si l'événement n'est pas apparu 2 ans après le début de l'étude alors l'événement n'était pas présent la première année et n'est pas apparu la deuxième année. Pour tout temps t_1 et t_2 tel que $t_2 > t_1$:

$$\begin{aligned} S(t_2) &= \mathbb{P}(X > t_2) \\ &= \mathbb{P}(X > t_2, X > t_1) \\ &= \mathbb{P}(X > t_2 \,|\, X > t_1)\mathbb{P}(X > t_1) \end{aligned}$$

Dans le cas général, en prenant les temps d'événement distinct $X_{(i)}$ avec $i = 1,...,n$ rangé par ordre croissant, on obtient

$$\begin{aligned} S(t_n) &= \mathbb{P}(X > t_n \,|\, X > t_{n-1}) \times \ldots \times \mathbb{P}(X > t_1) \\ &= \mathbb{P}(X > t_n \,|\, X > t_{n-1}) \times \ldots \times \mathbb{P}(X > t_1 \,|\, X > t_0) \\ &= \prod_{i=1}^{n} \mathbb{P}(X > t_i \,|\, X > t_{i-1}) \end{aligned}$$

avec $t_0 = 0$.

La probabilité de connaitre la probabilité t_i sachant qu'on n'a pas subi l'événement avant t_{i-1} est estimé par le nombre d'événements entre t_{i-1} et t_i parmi les sujets à risque au temps t_i :

$\mathbb{P}(X \leq t_i \,|\, X > t_{i-1}) = 1 - \mathbb{P}(X > t_i \,|\, X > t_{i-1}) = \dfrac{d_i}{Y_i}$, où Y_i est le nombre d'individus à risque de subir l'événement avant le temps t_i et d_i est le nombre de sujets subissant l'événement au t_i. Ainsi, $d_i = 0$ en cas de censure au temps t_i et $d_i = 1$ en cas de survenu de l'événement.

On obtient l'estimateur de Kaplan-Meier :

$$\hat{S}(t) = \prod_{\substack{i = 1,\ldots,n \\ t_i \leq t}} \left(1 - \frac{d_i}{Y_i}\right)$$

3.3. Fonctions de hasard cumulé

La fonction de hasard cumulé ou fonction de risque cumulé est une mesure statistique qui décrit comment le risque de défaillance évolue au fil du temps.

La fonction de hasard est définie comme le rapport entre la fonction de densité de défaillance instantanée ou taux de défaillance instantané et la fonction de survie. Elle représente la probabilité de défaillance à un instant donné, sachant que le sujet a survécu jusqu'à cet instant. La fonction de hasard cumulé est définie par :

$$H(t) = \int_0^t h(u)\,\mathrm{d}u,$$

où $h(t)$ est la fonction de hasard à l'instant t qui est définit par :

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq X < t + dt \,|\, X \geq t)}{dt} = \frac{f(t)}{S(t)}$$

où $f(t)$ est la fonction de densité de défaillance instantanée, et $S(t)$ est la fonction de survie, à l'instant t.

3.4. Estimer la fonction de hasard cumulé

Après avoir obtenu une estimation de la fonction de survie, il est souvent souhaitable d'obtenir une estimation de la fonction de hasard correspondante. La méthode de Nelson-Aalen est la méthode la plus couramment utilisée pour estimer la fonction de hasard cumulée à partir des données de survie.

3.4.1 Estimateur de Nelson-Aalen

La méthode de Nelson-Aalen est une méthode non paramétrique utilisée pour estimer la fonction de hasard cumulée.

L'estimation de la fonction de hasard cumulée à l'instant t avec la méthode de Nelson-Aalen est basée sur le calcul des taux de défaillance cumulés jusqu'à cet instant. Elle peut être utilisée pour des données de survie censurées en ajustant les taux de défaillance cumulés en fonction des individus encore à risque au moment de chaque événement de défaillance. L'estimation de la fonction de hasard cumulée est calculée comme suit :

$$\hat{H}(t) = \sum_{i:X_i \leq t} \left(\frac{d_i}{Y_i} \right)$$

où Y_i représente le nombre d'individus à risque juste avant X_i et d_i représente le nombre de décès en X_i. La fonction sera en « escalier » où chaque saut de taille $\frac{d_i}{Y_i}$ représentera chaque instant de survenue de la rétinopathie.

3.4. Application

En utilisant un programme R dédié à l'analyse de survie, nous allons estimer des fonctions de survie et de hasard cumulé à partir de la base de données.

D’un point de vue général, la fonction de survie ci-dessous représente l’ensemble des sujets de l’étude, soit 2248 personnes. L’axe des abscisses représente l'échelle du temps (en année). Il indique la durée écoulée depuis le début de l'étude jusqu'à un certain point dans le temps. L’axe des ordonnés représente la probabilité de survie à un moment donné. Cette probabilité indique la proportion d'individus qui n’ont pas la rétinopathie jusqu'à ce point spécifique dans le temps. La courbe décroissante indique que la probabilité de survie diminue avec le temps. Il n’y a pas de points d'inflexion, pas de changements brusques dans la forme de la courbe.

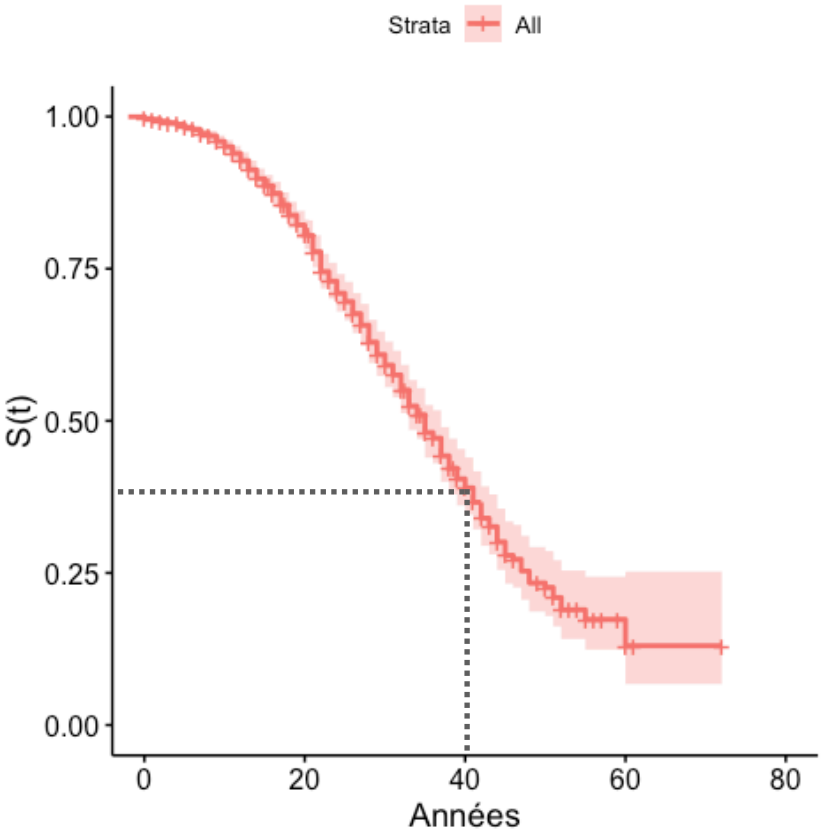

Figure 1 : Courbe de survie de l'ensemble de la base de données

Cohorte : 2248 personnes **Lecture :** Au bout de 40 ans, plus de 50% des personnes de cette étude sont atteint de la rétinopathie.

3.4.1 Selon l'âge de détection du diabète de type 1

La fonction de survie peut servir à comparer des groupes comme sur le graphique ci-dessous. On retrouve deux courbes l'une pour les personnes ayant été diagnostiqué du diabète de type avant la majorité (= 0) et l'autre pour les personnes ayant été diagnostiquées du diabète de type après la majorité (= 1). On peut visualiser les différences de survie entre ces deux groupes et évaluer l'impact de cette variable sur la survie. Ainsi, on peut dire qu'au début c'est plutôt les personnes ayant été diagnostiqué du diabète de type 1 qui ont une meilleure survie. Puis à partir d'une trentaine d'années, il y a une inversion.

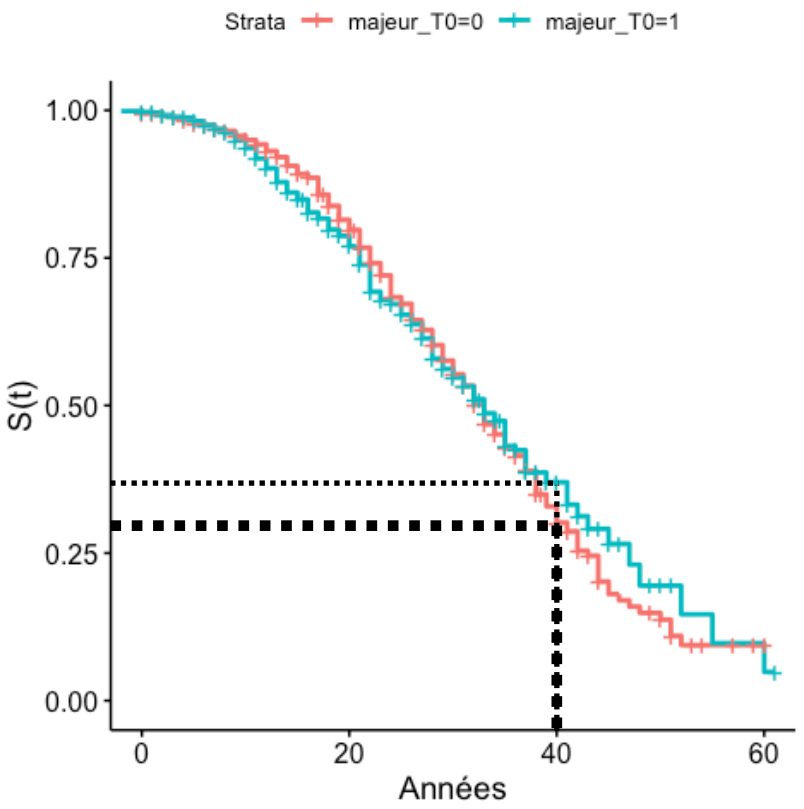

Figure 2 : Courbe de survie selon l'âge de détection du diabète de type 1

Cohorte : 890 personnes pour la courbe bleue et 1358 personnes pour la courbe rouge **Lecture :** Au bout de 40 ans, environ 33% des personnes ayant été diagnostiqué du diabète de type 1 après leurs majorités sont atteint de la rétinopathie. Les personnes ayant été diagnostiqué du diabète de type 1 avant leurs majorités ont environ 27% de survie.

Après avoir discuté de la probabilité de survie, il est également essentiel de considérer le risque cumulé. La fonction de hasard cumulé mesure l'accumulation du risque d'un événement jusqu'à un moment donné. Sur le graphique ci-dessous, la fonction de hasard cumulé se trouve à gauche. On constate que plus le temps augmente plus le risque d'avoir la rétinopathie augmente pour les personnes ayant été diagnostiquées du diabète de type 1 après leurs majorités. On se rend compte avec la fonction du de survie à droite de la liaison entre ces deux courbes.

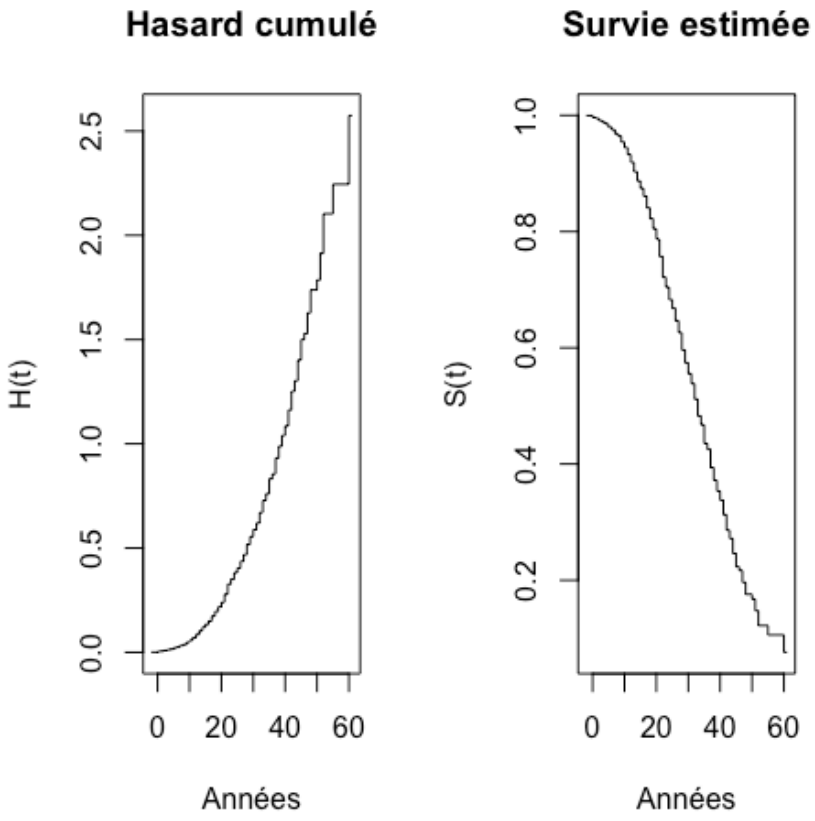

Figure 3 : Courbe du hasard cumulé et de survie chez les personnes ayant été diagnostiqué du diabète de type 1 après leurs majorités

Cohorte : 890 personnes ayant été diagnostiqué du diabète de type 1 après leurs majorités **Lecture :** L'événement d'intérêt est plus susceptible de se produire à 40 années de diabète de type 1 qu'à 60 ans de diabète de type 1.

En combinant l'analyse de la fonction de survie et de la fonction de hasard cumulé, on peut obtenir une vision globale de la dynamique du risque de survie dans l'étude considérée. Cela permet de mieux comprendre les facteurs qui influencent la survie des individus

4. Le modèle de Cox

Le modèle de Cox, où modèle de régression de Cox, est une méthode utilisée en analyse de survie pour évaluer l'impact des variables explicatives sur le risque de survie. Il a été développé par David Cox en 1972. Le modèle est adapté pour étudier des données censurées. Il permet de prendre en compte à la fois les individus qui ont subi l'événement et ceux qui sont encore en cours de suivi à la fin de l'étude.

Le modèle de Cox est un modèle multivarié. Il permet d'exprimer la relation entre le risque instantané de survenue de l'événement étudié h(t) et des facteurs de risque, les variables explicatives Le modèle de Cox s'exprime par la formule suivante :

$$h(t \mid Z) = \lambda_0(t)\exp(\beta' Z)$$

où Z est un vecteur de covariables de dimension $p \times 1$ et β un vecteur de dimension $p \times 1$ de coefficients de régression.

5. Construction du modèle de Cox

La construction du modèle de Cox implique plusieurs étapes. La première étant la sélection des variables explicatives par le biais de procédures de sélection de modèles. Les variables potentiellement associées au risque de survie sont en annexe. Une fois le modèle construit, on vérifie les hypothèses de non-significativité de chacune des variables ainsi que le modèle en lui-même. Enfin, grâce à ce modèle des courbes individuelles de survie par patient seront présentées.

5.1. Sélection de modèles

La sélection de modèles joue un rôle essentiel dans l'obtention de résultats précis et interprétables. L'objectif de la sélection de modèles est de choisir les variables explicatives les plus appropriées qui ont un impact significatif sur le risque de survie. On cherche à éviter le sur-ajustement ou le sous-ajustement du modèle. Un modèle sur-ajusté est trop complexe et peut entraîner une mauvaise généralisation des résultats. Tandis qu'un modèle sous-ajusté peut négliger des variables importantes et réduire la puissance de l'analyse. Les variables explicatives testées sont basées sur la connaissance du domaine par les professionnels présents pour ce projet MoGly.

5.1.1. Test du rapport de vraisemblance

Pour évaluer l'effet de chaque variable sur le risque de survie en utilisant un modèle de régression de Cox, le code R parcourt chaque variable afin de déterminer celles qui ont un impact significatif sur le risque de survie. Cette étape permet d'identifier les facteurs les plus influents et d'exclure les variables non pertinentes du modèle. Pour se faire, un test ANOVA est effectué entre le modèle excluant la variable testée et le modèle complet. Pour que ces tests aient lieu, il faut filtrer les lignes de données présentant des valeurs manquantes pour la variable testée. Le code

itère automatiquement toutes variables à tester grâce à une boucle « for ». Les résultats obtenus sont stockés et se trouvent dans le tableau ci-dessous.

Variables	P-value	Nombre de patients
HbA1c moyen	1.44439396252078e-05	2219
Triglycéride minimum	5.24022356514377e-04	2248
Variance triglycéride	1.26471607286465e-03	1918
Âge de déctection du Diabète de type 1	5.18094694461574e-03	2248
Triglycéride moyen	6.91112252009529e-03	1918
HbA1c minimum	1.81738133762111e-02	2248
Variance IMC	6.89532942012372e-02	2041
IMC maximum	1.54624309286822e-01	2248
Sexe	1.70041427326827e-01	2248
LDL moyen	1.80495210196468e-01	1893
Variance HbA1c	1.91490846052814e-01	2219
Dose insuline en fonction du poids	1.92222548756519e-01	1636
IMC minimum	3.62364423795496e-01	2248
HbA1c maximum	3.89547713319565e-01	2248
Triglycéride maximum	5.04149581268429e-01	2248
IMC moyen	5.79643901148704e-01	2041
Hypertension	8.58007334353938e-01	1642

Tableau 1 : Test du rapport de vraisemblance

Cohorte : voir colonne nombre de patients **Lecture :** La variable HbA1c moyen a un échantillon de 2219. La p-value du test ANOVA entre le modèle complet et le modèle complet moins cette variable est d'environ

0.14e-05. Ce qui fait que cette variable est très significative car elle est inférieur a 0.05.

Pour ce test, on peut retenir 6 variables : le taux d'HbA1c moyen des patients, le taux de triglycéride minimum des patients, la variance du taux de triglycéride des patients, l'âge de détection du diabète des patients, le taux de triglycéride moyen des patients et le taux d'HbA1c minimum des patients.

5.1.2. Méthode forward stepwise

La méthode forward stepwise, méthode de sélection de variables progressives. La méthode forward stepwise est une approche itérative de sélection de variables qui ajoute séquentiellement les prédicteurs les plus prédictifs au modèle. Elle ajoute la variable qui améliore la plus la performance du modèle en termes de critère d'évaluation (ici utilisé : le critère d'information bayésien (BIC) et le critère d'information d'Akaike (AIC)). Cette variable est sélectionnée en effectuant une régression avec chaque variable explicative restante et en choisissant celle qui apporte la plus grande amélioration. Après avoir ajouté la première variable, la méthode forward stepwise continue à ajouter séquentiellement les variables une par une, en choisissant à chaque étape celle qui apporte la plus grande amélioration. Le processus s'arrête lorsque l'ajout de nouvelles variables n'améliore plus le critère de sélection.

5.1.2.1. Méthode forward stepwise avec le critère d'évaluation BIC

Avec le critère d'évaluation d'Akaike (AIC), l'échantillon de patients est de taille 929. Dans le tableau ci-dessous, on retrouve les sept premières variables les plus significative dans l'ordre décroissant. Avec une

concordance de 0.723 (indice de concordance de Harrell[8]), cela indique que le modèle de régression de Cox a une capacité modérée à prédire le risque de survie des individus. La p-value est obtenue à partir du test de rapport de vraisemblance (likelihood ratio test en anglais). Ce test compare deux modèles : le modèle complet qui inclut toutes les variables et le modèle réduit qui exclut une variable spécifique. Le test complet se trouve en annexe.

Variables	P-value	Coefficient
HbA1c moyen	1.27e-05	0.48
Triglycéride minimum	0.002	− 1.4
Âge de détection du diabète de type 1	0.004	0.14
Variance triglycéride	0.006	− 0.68
Triglycéride moyen	0.0171	1.12
HbA1c minimum	0.0179	− 0.21
Variance IMC	0.07	− 0.003

Tableau 2 : Méthode forward stepwise AIC

Cohorte : 929 patients **Lecture :** Le taux d'HbA1c minimum a une p-value de 0.0179 par le test de vraisemblance. Cette variable est significative. Son coefficient négatif indique une relation négative entre la variable explicative et la variable de réponse.

5.1.2.2. Méthode forward stepwise avec le critère d'évaluation BIC

Avec le critère d'évaluation bayésien (BIC), l'échantillon de patients est de taille 929. Dans le tableau ci-dessous, on retrouve les sept premières variables les plus significative dans l'ordre décroissant. La concordance est de 0.723 ce qui indique que le modèle de régression de Cox a une capacité

[8] C'est une mesure utilisée pour évaluer la performance prédictive d'un modèle de régression de survie, tel que le modèle de régression de Cox. Un indice de concordance de Harrell supérieur à 0,5 indique que le modèle a une capacité prédictive meilleure que le hasard. Plus l'indice est proche de 1, meilleure est la performance prédictive du modèle.

modérée à prédire le risque de survie des individus. On obtient le même résultat que le test précédent. Le test complet se trouve en annexe.

Variables	P-value	Coefficient
HbA1c moyen	1.27e-05	0.48
Triglycéride minimum	0.002	− 1.4
Âge de détection du diabète de type 1	0.004	0.14
Variance triglycéride	0.006	− 0.68
Triglycéride moyen	0.0171	1.12
HbA1c minimum	0.0179	− 0.21
Variance IMC	0.07	− 0.003

Tableau 3 : Méthode forward stepwise BIC

Cohorte : 929 patients **Lecture :** Le taux d'HbA1c minimum a une p-value de 0.0179 par le test de vraisemblance. Cette variable est significative. Son coefficient négatif indique une relation négative entre la variable explicative et la variable de réponse.

5.1.3. Méthode backward stepwise

À l'inverse de la méthode précédente, la méthode backward stepwise est une approche itérative de sélection de variables qui élimine progressivement les variables non significatives d'un modèle statistique. La méthode backward stepwise commence avec un modèle comprenant toutes les variables explicatives et les élimine progressivement. Le processus de la méthode backward stepwise commence par ajuster un modèle avec toutes les variables explicatives. À chaque étape, la variable qui a le moins d'impact significatif sur le modèle est éliminée. Cette décision est généralement basée sur un critère d'évaluation statistique (ici utilisé : le critère d'information bayésien (BIC) et le critère d'information d'Akaike (AIC)). Après avoir éliminé une variable, le modèle est réajusté et les variables restantes sont réévaluées. Le processus se répète jusqu'à ce que

toutes les variables non significatives aient été éliminées et que le modèle final ne contienne que les variables les plus significatives.

5.1.3.1. Méthode backward stewise avec le critère de sélection AIC

Avec le critère d'évaluation d'Akaike (AIC), l'échantillon de patients est de 929. Dans le tableau ci-dessous, on retrouve les sept premières variables les plus significative dans l'ordre décroissant. La concordance est de 0.716 ce qui indique que le modèle de régression de Cox a une capacité modérée à prédire le risque de survie des individus. Le test complet se trouve en annexe.

Variables	P-value	Coefficient
HbA1c moyen	1.38e-06	0.51
Triglycéride minimum	0.0003	− 1.45
HbA1c minimum	0.0004	− 0.29
IMC maximum	0.0005	0.046
Triglycéride moyen	0.0006	1.3
Âge de détection du diabète	0.003	0.01
Variance triglycéride	0.005	− 0.67
HbA1c maximum	0.053	0.086

Tableau 4 : Méthode backward stepwise AIC

Cohorte : 929 patients **Lecture :** Le taux d'HbA1c moyen a une p-value de 1.38e-06 par le test de vraisemblance. Cette variable est significative. Son coefficient positif indique une relation positive entre la variable explicative et la variable de réponse.

5.1.3.2. Méthode backward stewise avec le critère de sélection BIC

Avec le critère d'évaluation bayésien (BIC), l'échantillon de patients est de 929. Dans le tableau ci-dessous, on retrouve les sept premières

variables les plus significatives dans l'ordre décroissant. La concordance est de 0.708 ce qui indique que le modèle de régression de Cox a une capacité modérée à prédire le risque de survie des individus. Le test complet se trouve en annexe.

Variables	P-value	Coefficient
HbA1c moyen	< 2e-16	0.63
HbA1c minimum	2.8e-07	− 0.33
Triglycéride minimum	0.0001	− 1.5
Triglycéride moyen	0.0003	1.33
Âge de détection du diabète	0.001	0.01
IMC maximum	0.0026	0.03
Variance triglycéride	0.003	− 0.69

Tableau 5 : Méthode backward stepwise BIC

Cohorte : 929 patients **Lecture :** Le taux d'HbA1c moyen a une p-value inférieur à 2e-16 par le test de vraisemblance. Cette variable est significative. Son coefficient positif indique une relation positive entre la variable explicative et la variable de réponse.

Pour ces quatre tests, on retrouve souvent les mêmes variables lors des résultats. En se basant sur la valeur de la concordance des tests et la redondance des variables explicatives, on peut construire un modèle optimale.

5.2. Modèle optimale

Les variables sélectionnées sont : le taux d'HbA1c moyen et minimum, le taux de triglycérides minimum, moyen et leur variance, l'âge de détection du diabète de type 1 et la variance de l'IMC. Ces variables explicatives forment le modèle optimal. Par des tests de non-significativité, chaque variable sera re-testée puis classée selon sa p-value.

5.2.1. test des variables

Pour évaluer l'impact de chaque variables sur le modèle de Cox optimal, plusieurs tests statistiques ont été réalisés.

Le test de rapport de vraisemblance avec ANOVA, le code effectue un test entre le modèle excluant la variable testée et le modèle complet qui inclut toutes les variables. Cela permet de comparer la performance des deux modèles et de déterminer si l'exclusion de la variable testée conduit à une perte significative d'information.

Avant de réaliser les tests ANOVA, les lignes des variables testées sont filtrées en présence de valeurs manquantes. Cette étape garantit que l'analyse soit basée sur des données complètes et fiables.
Ces résultats sont stockés dans un tableau dans l'ordre croissant, comme illustré ci-dessous :

Variables	P-value	Nombre de patients	Signe des coefficients
HbA1c moyen	2.62082392986035e-24	2219	+
HbA1c minimum	5.26988874072328e-10	2248	−
Âge de déctection du Diabète de type 1	9.82433719267694e-05	2248	+
Triglycéride moyen	0.027387508974948 9	1918	+
Variance triglycéride	0.071610096461916 7	1918	−

Tableau 6 : Test du rapport de vraisemblance (modèle optimal et modèle optimal sans la variable à tester)

Cohorte : colonne Nombre de patients (max = 2248) **Lecture :** Le taux d'HbA1c minimum a une p-value de 2.6e-24 en comparant le modèle

optimal et le modèle optimal sans la variable du taux d'HbA1c moyen. Cette variable est significative.

Un deuxième test est pris en compte, il s'agit du test de Wald. Il teste l'hypothèse nulle selon laquelle les coefficients de régression pour chaque variable indépendante sont égaux à zéro. Cela signifierait qu'il n'y a pas d'effet de ces variables sur le risque relatif. Si la p-value est inférieure à un seuil de significativité (0.05), l'hypothèse nulle est rejetée et on peut conclut suggérer que le coefficient de régression est significativement différent de zéro. Ainsi, la variable a un effet significatif sur le risque relatif. Le tableau ci-dessous regroupe les variables associées à leurs résultats du test de Wald par ordre croissant.

Variables	P-value	Coefficient
HbA1c moyen	< 2e-16	0.597534
HbA1c minimum	2.31e-10	– 0.333067
Âge de déctection du Diabète de type 1	6.28e-05	0.015755
Triglycéride moyen	0.0147	0.265846
Variance triglycéride	0.0708	– 0.006117

Tableau 7 : Test de Wald

Cohorte : 1622 patients **Lecture :** Le taux d'HbA1c moyen a une p-value inférieur à 2e-16 par le test de Wald. Cette variable est significative. Son coefficient positif indique une relation positive entre la variable explicative et la variable de réponse.

Les variables avec des valeurs de p-value faibles (inférieures à 0.05) sont considérées comme significativement liées au risque de survie, tandis que les variables avec des valeurs de p-value élevées pourrait être exclues car elles ne contribuent pas de manière significative à la prédiction du risque de survie. Pour la suite, les variables ne sont pas retirées, le modèle de Cox optimal va être testé dans son ensemble.

5.3. Vérification du modèle de Cox optimal

Pour évaluer le modèle de Cox optimal, plusieurs tests[9] statistiques ont été réalisé également.

Les tests de rapport de vraisemblance (Likelihood ratio test), de Wald et du score évaluent si l'ensemble des variables du modèle ont un effet significatif sur le risque relatif. Voici les résultats :

Test	**P-value**
Rapport de vraisemblance	< 2e-16
Wald	< 2e-16
Score	< 2e-16

Tableau 8 : Tests du Modèle de Cox optimal

Cohorte : 1622 patients **Lecture :** Le test de vraisemblance a une p-value inférieur à 2e-16.

Les p-value sont toutes inférieur à 2e-16 indiquant le rejet l'hypothèse nulle et suggère qu'il existe une association significative entre les variables et le risque de survie. Le résultat complet du test se trouve en annexe.

5.4. Courbe individuelle de survie

Une courbe individuelle de survie représente la probabilité de survie d'un individu spécifique. En utilisant le modèle optimal et une ligne d'un patient de la base de données finale, on peut tracer sa courbe de survie.

Prenons le patient 8, il n'est pas atteint de la rétinopathie. Son diabète de type 1 a commencé en 2008 à l'âge de 7 ans. Il est suivi depuis

[9] Les tests évaluent l'hypothèse nulle selon laquelle les coefficients de régression du modèle de Cox sont tous égaux à zéro.

le début de l'étude et comptabilise 55 questionnaires[10] sur 14 ans. Les valeurs des variables explicatives sont dans le tableau suivant :

Variables	Valeurs
HbA1c moyen	9.316279 %
HbA1c minimum	7 %
Âge de déctection du Diabète de type 1	7 ans
Triglycéride moyen	1 mmol
Variance triglycéride	0.71755 mmol
Variance IMC	4.44767834 kg/m^2
Triglycéride minimum	0.43 mmol

Tableau 9 : Variables explicatives du patient 8

Cohorte : patient 8 **Lecture :** La variance de l'IMC du patient 8 est d'environ 4.45 kg/m^2.

En appliquant le modèle optimal avec les valeurs explicatives du patient 8, on obtient cette courbe de survie, sur le graphique de gauche ci-dessous. On constate qu'à partir de 50 années de diabète de type 1, la probabilité de survie est quasiment nulle pour l'apparition de la rétinopathie. Le graphique de droite permet une visualisation du patient 8 par rapport à l'ensemble des personnes de l'étude, qui ont ou pas la rétinopathie.

[10] Il y a 21 types de questionnaires dont le majoritaire est la consultation. On peut retrouver également de l'ophtalmologie, de la podologie, de la neurologie, du diététique, de la cardiologie, de la néphrologie, des suivis rétinopathie et des suivis pompe insuline.

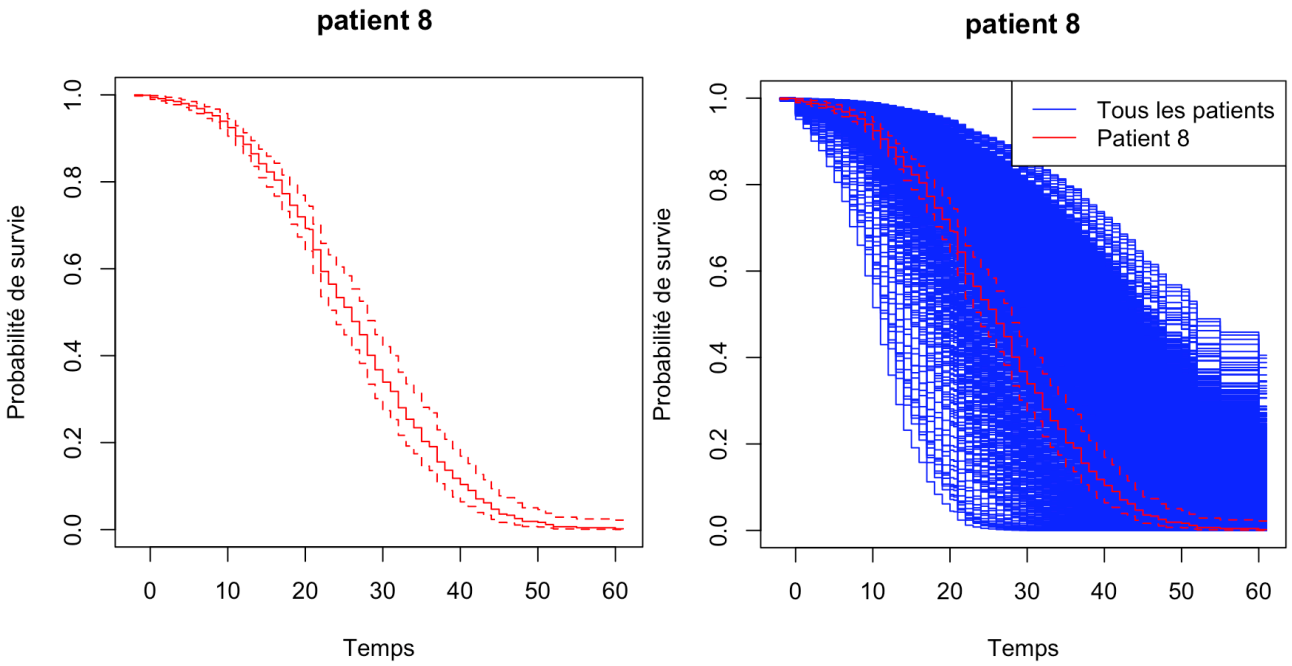

Figure 4 : Courbe de survie du patient 8

Cohorte : patient 8 **Lecture :** La probabilité de survie 20 ans après avoir était diagnostiqué du diabète de type 1 est d'environ 0.7.

Cohorte : 2248 patients **Lecture :** Position du patient 8 comparé aux restes des patients.

Les courbes individuelles de survie prennent en compte les caractéristiques propres du patient tout en visualisant la variation de la durée de survie entre les individus. Elles permettent également de comprendre la possible évolution de la maladie.

6. Conclusion

La rétinopathie est une complication fréquente chez les patients atteints de diabète de type 1. Dans ce mémoire, nous avons utilisé le modèle de Cox pour étudier les facteurs associés à l'apparition de la rétinopathie chez les patients diabétiques de type 1.

Nos résultats ont montré que l'âge de détection du diabète de type 1, la variance de l'IMC, les niveaux de triglycérides et les niveaux d'HbA1c étaient des facteurs significatifs au risque de développer la rétinopathie.

Nous avons observé une augmentation du risque de rétinopathie avec l'âge et l'étude a démontré une utilité du modèle de Cox dans l'analyse de la relation entre les covariables et l'apparition de la rétinopathie chez les patients diabétiques de type 1.

Pour aller plus loin, le modèle de Cox pourrait inclure des covariables dépendantes du temps, ce qui permettait de prendre en compte les variations temporelles des facteurs de risque. Cela pourrait fournir des estimations plus précises sur la survie, en tenant compte des fluctuations temporelles des covariables.

7. Bibliographie

Statistique sur le diabète

Le diabète, une maladie qui progresse (SANTI Pascale, 30 mai 2022), Le Monde
https://www.lemonde.fr/sciences/article/2022/05/30/le-diabete-une-maladie-qui-progresse_6128229_1650684.html#:~:text=La%20maladie%2C%20qui%20touche%20plus,5%20%25%20par%20an%20en%20France.

Théorie

Introduction aux analyses de survie (FOUCHER Yohann), Université de Nantes Master 2 - Modélisation en Pharmacologie Clinique et Epidémiologie
https://www.divat.fr/images/Biostats/Teaching/mpce-introduction_analyse_de_survie.pdf

Introduction à l'analyse des durées de survie (SAINT PIERRE Philippe, 2015), Université Pierre et Marie Curie

Programmation

Analyse de survie : le modèle de Cox (BOUAZIZ Olivier), Université Paris Cité
https://helios2.mi.parisdescartes.fr/~obouaziz/CoxSurv.pdf

Modèles semi-paramétriques de survie en temps continu sous R (QUANTIN Simon), INSEE
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjDje3A8of_AhUGUqQEHUnEDzUQFnoECCMQAQ&url=https%3A%2F%2Fwww.insee.fr%2Ffr%2Fstatistiques%2Ffichier%2F3695681%2FM2018-02.pdf&usg=AOvVaw3Ev9DgielXKoJwjRCbUlSi

8. Table des figures

9. Table des tableaux

10. Table des annexes

11. Annexes

Variables utilisées de la base de données finale

Variable	Descrition
IdPatient	Numéro du patient
Sexe	Sexe
AnneeNaissance2	Année de Naissance
CDDateDecouverteDiabete	Année de découverte du diabète de type 1
AgeDecDiabete	Âge de découverte du diabète de type 1
Retino_vrai	Le patient a la rétinopathie (=1)
nbconsult	Nombre de questionnaires
full_suivi	Dans l'étude depuis la détection du diabète de type 1 (=1)
annee_fin_suivi	Fin de suivie dans l'étude
moy_imc	IMC moyen
max_imc	IMC maximum
min_imc	IMC minimum
var_imc	Variance de l'IMC
ldl_moy	LDL moyen
dose_pds_moy	Dose d'insuline en fonction du poids
trigly_moy	Triglycéride moyen
trigly_min	Triglycéride maximum
trigly_max	Triglycéride minimum
trigly_var	Variance du taux de Triglycéride (mmol)
HbA1c_moy	HbA1c moyen
HbA1c_min	HbA1c maximum
HbA1c_max	HbA1c minimum
HbA1c_var	Variance du taux d'HbA1c
Hypertension	Hypertension (=1)
Fumer_vrai	Fumeur (=1)
Survie	Années entre la découverte du diabète de type 1 et l'apparition de la rétinopathie
Censure	durée écoulée avant la censure de l'information
T_i	min(Censure, Survie)

Variable	Descrition
majeur_T0	Majeur au moment de la découverte du diabète de type 1

Variables	
HbA1c moyen	Sexe
Triglycéride minimum	LDL moyen
Variance triglycéride	Variance HbA1c
Âge de déctection du Diabète de type 1	Dose insuline en fonction du poids
Triglycéride moyen	IMC minimum
HbA1c minimum	HbA1c maximum
Variance IMC	Triglycéride maximum
IMC maximum	IMC moyen
Hypertension	

Variables potentielles aux modèle de Cox

Note : La prise en compte de la consommation de tabac est rejetée car l'échantillon de réponse est insuffisant pour l'étude.

Méthode backward stepwise AIC

```
> summary(model_back_aic_b)
Call:
coxph(formula = Surv(T_i, Retino_vrai == 1) ~ AgeDecDiabete +
    max_imc + var_imc + ldl_moy + trigly_moy + trigly_min + trigly_var +
    HbA1c_moy + HbA1c_max + HbA1c_min + factor(Sexe), data = Diabdata)

  n= 929, number of events= 303

                    coef exp(coef)  se(coef)      z Pr(>|z|)
AgeDecDiabete  0.013587  1.013680  0.004649  2.922 0.003475 **
max_imc        0.045671  1.046730  0.013018  3.508 0.000451 ***
var_imc       -0.015904  0.984221  0.010122 -1.571 0.116133
ldl_moy       -0.251560  0.777587  0.178353 -1.410 0.158403
trigly_moy     1.290328  3.633978  0.375350  3.438 0.000587 ***
trigly_min    -1.454086  0.233614  0.397276 -3.660 0.000252 ***
trigly_var    -0.673681  0.509828  0.238822 -2.821 0.004790 **
HbA1c_moy      0.509685  1.664767  0.105580  4.827 1.38e-06 ***
HbA1c_max      0.085857  1.089651  0.044358  1.936 0.052922 .
HbA1c_min     -0.282935  0.753569  0.079303 -3.568 0.000360 ***
factor(Sexe)M  0.175177  1.191458  0.118299  1.481 0.138659
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
AgeDecDiabete    1.0137     0.9865    1.0045    1.0230
max_imc          1.0467     0.9554    1.0204    1.0738
var_imc          0.9842     1.0160    0.9649    1.0039
ldl_moy          0.7776     1.2860    0.5482    1.1030
trigly_moy       3.6340     0.2752    1.7413    7.5837
trigly_min       0.2336     4.2806    0.1072    0.5089
trigly_var       0.5098     1.9614    0.3193    0.8142
HbA1c_moy        1.6648     0.6007    1.3536    2.0475
HbA1c_max        1.0897     0.9177    0.9989    1.1886
HbA1c_min        0.7536     1.3270    0.6451    0.8803
factor(Sexe)M    1.1915     0.8393    0.9449    1.5024

Concordance= 0.716  (se = 0.018 )
Likelihood ratio test= 149  on 11 df,   p=<2e-16
Wald test            = 160.9  on 11 df,   p=<2e-16
Score (logrank) test = 164.3  on 11 df,   p=<2e-16
```

Lecture :

Les coefficients (coef) représentent les estimations de l'effet de chaque variable indépendante sur le risque relatif. Par exemple, pour la variable « AgeDecDiabete » (l’âge de détection du diabète), le coefficient est de 0.013587 signifie une augmentation 1.4% du risque relatif de l'événement étudié par années.

Les valeurs (Pr(>|z|)) indiquent la probabilité d'observer une relation entre la variable indépendante et le risque relatif. Pour la variable de l’âge de détection du diabète, la p-value est inférieure à 0.05, ce qui suggère une

relation significative entre l'âge au diagnostic du diabète et le risque relatif de l'événement.

Les ratios exp(coef) représentent l'interprétation exponentielle des coefficients et correspondent aux rapports de risques instantanés. Pour la variable « HbA1c_moy » (le taux d'HbA1c moyen), le ratio exp(coef) est de 1.0137, le risque de l'événement d'intérêt est multiplié par 1.0137

Les valeurs « se(coef) » représente l'écart-type estimé du coefficient de régression. Il mesure la précision de l'estimation du coefficient.

La concordance est une mesure de l'ajustement global du modèle. Une valeur proche de 1 indique une meilleure adéquation du modèle aux données.

Les tests de rapport de vraisemblance (Likelihood ratio test), de Wald (Wald test) et du score (Score (logrank) test) évaluent si l'ensemble des variables indépendantes du modèle ont un effet significatif sur le risque relatif.

Méthode backward stepwise BIC

```
> summary(model_back_bic_b)
Call:
coxph(formula = Surv(T_i, Retino_vrai == 1) ~ AgeDecDiabete +
    max_imc + trigly_moy + trigly_min + trigly_var + HbA1c_moy +
    HbA1c_min, data = Diabdata)

  n= 929, number of events= 303

                  coef exp(coef)  se(coef)      z Pr(>|z|)
AgeDecDiabete  0.014484  1.014589  0.004529  3.198 0.001382 **
max_imc        0.031286  1.031781  0.010401  3.008 0.002631 **
trigly_moy     1.332851  3.791840  0.365181  3.650 0.000262 ***
trigly_min    -1.492658  0.224774  0.384977 -3.877 0.000106 ***
trigly_var    -0.689538  0.501808  0.233616 -2.952 0.003162 **
HbA1c_moy      0.631102  1.879681  0.065662  9.611  < 2e-16 ***
HbA1c_min     -0.325072  0.722475  0.063293 -5.136 2.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
AgeDecDiabete    1.0146     0.9856    1.0056    1.0236
max_imc          1.0318     0.9692    1.0110    1.0530
trigly_moy       3.7918     0.2637    1.8536    7.7570
trigly_min       0.2248     4.4489    0.1057    0.4780
trigly_var       0.5018     1.9928    0.3175    0.7932
HbA1c_moy        1.8797     0.5320    1.6527    2.1378
HbA1c_min        0.7225     1.3841    0.6382    0.8179

Concordance= 0.708  (se = 0.018 )
Likelihood ratio test= 138.4  on 7 df,   p=<2e-16
Wald test            = 154.7  on 7 df,   p=<2e-16
Score (logrank) test = 155  on 7 df,   p=<2e-16
```

Méthode forward stepwise AIC

```
> summary(model_back_aic_f)
Call:
coxph(formula = Surv(T_i, Retino_vrai == 1) ~ AgeDecDiabete +
    moy_imc + max_imc + min_imc + var_imc + ldl_moy + dose_pds_moy +
    trigly_moy + trigly_max + trigly_min + trigly_var + HbA1c_moy +
    HbA1c_max + HbA1c_min + HbA1c_var + factor(Sexe) + factor(Hypertension),
    data = Diabdata)

  n= 929, number of events= 303

                           coef exp(coef) se(coef)      z Pr(>|z|)
AgeDecDiabete           0.01442   1.01452  0.00503  2.867  0.00414 **
moy_imc                 0.02957   1.03001  0.05589  0.529  0.59680
max_imc                 0.05675   1.05839  0.03697  1.535  0.12473
min_imc                -0.04199   0.95888  0.04572 -0.918  0.35840
var_imc                -0.02982   0.97062  0.01670 -1.785  0.07421 .
ldl_moy                -0.23618   0.78964  0.18448 -1.280  0.20046
dose_pds_moy            0.29787   1.34698  0.23721  1.256  0.20921
trigly_moy              1.11553   3.05118  0.46775  2.385  0.01708 *
trigly_max              0.09110   1.09538  0.13897  0.656  0.51212
trigly_min             -1.40332   0.24578  0.44417 -3.159  0.00158 **
trigly_var             -0.67783   0.50772  0.24733 -2.741  0.00613 **
HbA1c_moy               0.47585   1.60939  0.10900  4.366 1.27e-05 ***
HbA1c_max               0.04676   1.04787  0.05014  0.933  0.35097
HbA1c_min              -0.21478   0.80671  0.09068 -2.369  0.01785 *
HbA1c_var               0.04387   1.04485  0.02994  1.465  0.14281
factor(Sexe)M           0.16178   1.17560  0.12197  1.326  0.18470
factor(Hypertension)Oui 0.02460   1.02490  0.13821  0.178  0.85875
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                        exp(coef) exp(-coef) lower .95 upper .95
AgeDecDiabete              1.0145     0.9857    1.0046    1.0246
moy_imc                    1.0300     0.9709    0.9231    1.1493
max_imc                    1.0584     0.9448    0.9844    1.1379
min_imc                    0.9589     1.0429    0.8767    1.0488
var_imc                    0.9706     1.0303    0.9394    1.0029
ldl_moy                    0.7896     1.2664    0.5500    1.1336
dose_pds_moy               1.3470     0.7424    0.8462    2.1442
trigly_moy                 3.0512     0.3277    1.2199    7.6316
trigly_max                 1.0954     0.9129    0.8342    1.4383
trigly_min                 0.2458     4.0687    0.1029    0.5870
trigly_var                 0.5077     1.9696    0.3127    0.8244
HbA1c_moy                  1.6094     0.6214    1.2998    1.9927
HbA1c_max                  1.0479     0.9543    0.9498    1.1561
HbA1c_min                  0.8067     1.2396    0.6754    0.9636
HbA1c_var                  1.0448     0.9571    0.9853    1.1080
factor(Sexe)M              1.1756     0.8506    0.9256    1.4931
factor(Hypertension)Oui    1.0249     0.9757    0.7817    1.3438

Concordance= 0.723  (se = 0.017 )
Likelihood ratio test= 153  on 17 df,   p=<2e-16
Wald test            = 165.4  on 17 df,   p=<2e-16
Score (logrank) test = 176.1  on 17 df,   p=<2e-16
```

Méthode forward stepwise BIC

```
> summary(model_back_bic_f)
Call:
coxph(formula = Surv(T_i, Retino_vrai == 1) ~ AgeDecDiabete +
    moy_imc + max_imc + min_imc + var_imc + ldl_moy + dose_pds_moy +
    trigly_moy + trigly_max + trigly_min + trigly_var + HbA1c_moy +
    HbA1c_max + HbA1c_min + HbA1c_var + factor(Sexe) + factor(Hypertension),
    data = Diabdata)

  n= 929, number of events= 303

                           coef exp(coef) se(coef)      z Pr(>|z|)
AgeDecDiabete           0.01442   1.01452  0.00503  2.867  0.00414 **
moy_imc                 0.02957   1.03001  0.05589  0.529  0.59680
max_imc                 0.05675   1.05839  0.03697  1.535  0.12473
min_imc                -0.04199   0.95888  0.04572 -0.918  0.35840
var_imc                -0.02982   0.97062  0.01670 -1.785  0.07421 .
ldl_moy                -0.23618   0.78964  0.18448 -1.280  0.20046
dose_pds_moy            0.29787   1.34698  0.23721  1.256  0.20921
trigly_moy              1.11553   3.05118  0.46775  2.385  0.01708 *
trigly_max              0.09110   1.09538  0.13897  0.656  0.51212
trigly_min             -1.40332   0.24578  0.44417 -3.159  0.00158 **
trigly_var             -0.67783   0.50772  0.24733 -2.741  0.00613 **
HbA1c_moy               0.47585   1.60939  0.10900  4.366 1.27e-05 ***
HbA1c_max               0.04676   1.04787  0.05014  0.933  0.35097
HbA1c_min              -0.21478   0.80671  0.09068 -2.369  0.01785 *
HbA1c_var               0.04387   1.04485  0.02994  1.465  0.14281
factor(Sexe)M           0.16178   1.17560  0.12197  1.326  0.18470
factor(Hypertension)Oui 0.02460   1.02490  0.13821  0.178  0.85875
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                        exp(coef) exp(-coef) lower .95 upper .95
AgeDecDiabete              1.0145     0.9857    1.0046    1.0246
moy_imc                    1.0300     0.9709    0.9231    1.1493
max_imc                    1.0584     0.9448    0.9844    1.1379
min_imc                    0.9589     1.0429    0.8767    1.0488
var_imc                    0.9706     1.0303    0.9394    1.0029
ldl_moy                    0.7896     1.2664    0.5500    1.1336
dose_pds_moy               1.3470     0.7424    0.8462    2.1442
trigly_moy                 3.0512     0.3277    1.2199    7.6316
trigly_max                 1.0954     0.9129    0.8342    1.4383
trigly_min                 0.2458     4.0687    0.1029    0.5870
trigly_var                 0.5077     1.9696    0.3127    0.8244
HbA1c_moy                  1.6094     0.6214    1.2998    1.9927
HbA1c_max                  1.0479     0.9543    0.9498    1.1561
HbA1c_min                  0.8067     1.2396    0.6754    0.9636
HbA1c_var                  1.0448     0.9571    0.9853    1.1080
factor(Sexe)M              1.1756     0.8506    0.9256    1.4931
factor(Hypertension)Oui    1.0249     0.9757    0.7817    1.3438

Concordance= 0.723  (se = 0.017 )
Likelihood ratio test= 153  on 17 df,   p=<2e-16
Wald test            = 165.4  on 17 df,   p=<2e-16
Score (logrank) test = 176.1  on 17 df,   p=<2e-16
```

Tests du modèle optimal et ses variables

```
> summary(modele_cox)
Call:
coxph(formula = Surv(T_i, Retino_vrai == 1) ~ AgeDecDiabete +
    var_imc + trigly_moy + trigly_min + trigly_var + HbA1c_moy +
    HbA1c_min, data = Diab)

  n= 1622, number of events= 409
   (626 observations effacées parce que manquantes)

                  coef exp(coef)  se(coef)      z Pr(>|z|)
AgeDecDiabete  0.015755  1.015879  0.003937  4.002 6.28e-05 ***
var_imc        0.008907  1.008947  0.006452  1.381   0.1674
trigly_moy     0.265846  1.304535  0.108971  2.440   0.0147 *
trigly_min    -0.121972  0.885173  0.150456 -0.811   0.4175
trigly_var    -0.006117  0.993902  0.003386 -1.807   0.0708 .
HbA1c_moy      0.597534  1.817631  0.054845 10.895  < 2e-16 ***
HbA1c_min     -0.333067  0.716722  0.052539 -6.339 2.31e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
AgeDecDiabete    1.0159     0.9844    1.0081    1.0237
var_imc          1.0089     0.9911    0.9963    1.0218
trigly_moy       1.3045     0.7666    1.0537    1.6151
trigly_min       0.8852     1.1297    0.6591    1.1888
trigly_var       0.9939     1.0061    0.9873    1.0005
HbA1c_moy        1.8176     0.5502    1.6324    2.0239
HbA1c_min        0.7167     1.3952    0.6466    0.7945

Concordance= 0.672  (se = 0.018 )
Likelihood ratio test= 141.5  on 7 df,   p=<2e-16
Wald test            = 165.2  on 7 df,   p=<2e-16
Score (logrank) test = 168.4  on 7 df,   p=<2e-16
```

Code R

```
library(dplyr)
library(survival)
library(survminer)
library(lubridate)
library(magrittr)
library(timereg)
library(MASS)

# Importation de la base ----
base_1_lea <- read.csv("/Volumes/NO NAME/Lea_MoGly/
base_1_lea.csv")
base_1_lea<-base_1_lea[,c(-1)] # retire la ligne ajouté automatiquement

## Modif base ----
base_1_lea %>%
  mutate(Survie = Retino_year - CDDateDecouverteDiabete) %>% # X_i
  mutate(Censure = annee_fin_suivi - CDDateDecouverteDiabete) %>% #
C_i
  mutate(Retino_year_or_end = ifelse(is.finite(Retino_year),Retino_year,
annee_fin_suivi)) %>%
  mutate(T_i = Retino_year_or_end - CDDateDecouverteDiabete) %>%
#T_i
  mutate(majeur_T0 = ifelse(AgeDecDiabete<18,0,1)) %>% # majeur au
moment de la découverte du diabete
  identity() -> Diab


Diab <- mutate_at(Diab, "Retino_vrai", ~replace(., is.na(.), 0)) #sinon pb
graphique
Diab$Retino_vrai <- as.numeric(Diab$Retino_vrai)
```

```
# Kaplan Meier ----
# Tout les individus
fit=coxph(Surv(T_i,Retino_vrai==1)~ 1, data = Diab)
Hazcum=basehaz(fit,centered=FALSE)
par(mfrow=c(1,2))
plot(Hazcum$time,Hazcum$hazard,type="s",xlab="Années",ylab="H(t)")
title("Hasard cumulé")
plot(Hazcum$time,exp(-
Hazcum$hazard),type="s",xlab="Années",ylab="S(t)")
title("Survie estimée")

ggsurvplot(
  fit = survfit(Surv(T_i, Retino_vrai) ~ 1, data = Diab),
  xlab = "Années",
  ylab = "S(t)")

# survie selon le genre
fit=coxph(Surv(T_i,Retino_vrai==1)~ Sexe, data = Diab)
Hazcum=basehaz(fit,centered=FALSE)
par(mfrow=c(1,2))
plot(Hazcum$time,Hazcum$hazard,type="s",xlab="Années",ylab="H(t)")
title("Hasard cumulé")
plot(Hazcum$time,exp(-
Hazcum$hazard),type="s",xlab="Années",ylab="S(t)")
title("Survie estimée")

ggsurvplot(
  fit = survfit(Surv(T_i, Retino_vrai) ~ Sexe, data = Diab),
  xlab = "Années",
  ylab = "S(t)")


## majeur ou pas au moment de la découverte du diabete
```

```
fit=coxph(Surv(T_i,Retino_vrai==1)~ majeur_T0, data = Diab)
Hazcum=basehaz(fit,centered=FALSE)
par(mfrow=c(1,2))
plot(Hazcum$time,Hazcum$hazard,type="s",xlab="Années",ylab="H(t)")
title("Hasard cumulé")
plot(Hazcum$time,exp(-
Hazcum$hazard),type="s",xlab="Années",ylab="S(t)")
title("Survie estimée")

summary(fit)
ggsurvplot(
  fit = survfit(Surv(T_i, Retino_vrai) ~ majeur_T0, data = Diab),
  xlab = "Années",
  ylab = "S(t)")

# Ajustement du modèle de régression de Cox
# Modele de Cox ----
Diab <- Diab[, -which(names(Diab) == "Fumer_vrai")] # trop de NA
# Modele complet
mod_full <- coxph(Surv(T_i, Retino_vrai==1) ~   AgeDecDiabete +
moy_imc + max_imc
                  + min_imc + var_imc + ldl_moy + dose_pds_moy + trigly_moy

+trigly_max+trigly_min+trigly_var+HbA1c_moy+HbA1c_max+HbA1c_
min
                  +HbA1c_var+ factor(Sexe)+ factor(Hypertension),
                  data = Diab)
summary(mod_full)
cox.zph(mod_full)

## Premier test des varaibles ----
# Liste de variables à tester
```

```
variables <- c("AgeDecDiabete", "moy_imc", "max_imc", "min_imc",
"var_imc", "ldl_moy", "dose_pds_moy", "trigly_moy", "trigly_max",
"trigly_min", "trigly_var", "HbA1c_moy", "HbA1c_max", "HbA1c_min",
"HbA1c_var", "Sexe", "Hypertension")
res <- data.frame("var" = numeric(0), "test_anova" =
numeric(0),"nombre_patient"=numeric(0))
# Boucle for pour retirer une variable à chaque itération
for (i in 1:length(variables)) {
  # Sélection de toutes les variables sauf la variable à retirer
  x<-variables[i]
  variables_selectionnees <- variables[-i]

  # filtrer les lignes avec des valeurs manquantes pour la colonne "col3"
  ligne_complete <- !is.na(Diab[,x])

  # sélectionner les lignes complètes pour la colonne "col3" et toutes les
colonnes
  data_filtre <- Diab[ligne_complete, ]
  # Construction de la formule de survie en utilisant les variables
sélectionnées
  formule_survie <- as.formula(paste("Surv(T_i, Retino_vrai == 1) ~",
paste(variables_selectionnees, collapse = " + ")))

  # Application de la régression de Cox à la formule de survie
  mod <- coxph(formule_survie, data = data_filtre)
  mod_full <- coxph(Surv(T_i, Retino_vrai==1) ~   AgeDecDiabete +
moy_imc + max_imc
                + min_imc + var_imc + ldl_moy + dose_pds_moy +
trigly_moy

+trigly_max+trigly_min+trigly_var+HbA1c_moy+HbA1c_max+HbA1c_
min
                +HbA1c_var+ factor(Sexe)+ factor(Hypertension),
```

```
                data = data_filtre)

  a <- anova(mod,mod_full)

  # ajouter une ligne de valeurs
  nouvelle_ligne <- c(x, a[2,4], nrow(data_filtre))
  res <- rbind(res, nouvelle_ligne)
}
names(res) <- c("var", "test_anova","nombre_patient")
res

## test AIC et BIC ----
Diabdata <- na.omit(Diabdata) #929 obs

modele_complet <- coxph(Surv(T_i, Retino_vrai==1) ~ AgeDecDiabete +
moy_imc + max_imc +
                  min_imc + var_imc + ldl_moy + dose_pds_moy +
trigly_moy +
                  trigly_max + trigly_min + trigly_var + HbA1c_moy +
HbA1c_max +
                  HbA1c_min + HbA1c_var + factor(Sexe) +
factor(Hypertension),
                data = Diabdata)

model_back_aic_b <- stepAIC(modele_complet, scope = list(lower = . ~ 1,
upper = . ~ .),
                    direction = "backward")
model_back_aic_b
summary(model_back_aic_b)

model_back_aic_f <- stepAIC(modele_complet, scope = list(lower = . ~ 1,
upper = . ~ .),
                    direction = "forward")
```

```
model_back_aic_f
summary(model_back_aic_f)


model_back_aic_both <- stepAIC(modele_complet, scope = list(lower = . ~
1, upper = . ~ .),
                               direction = "both")
model_back_aic_both
summary(model_back_aic_both)

model_back_bic_b <- stepAIC(modele_complet, scope = list(lower = . ~ 1,
upper = . ~ .),
                            direction = "backward",k=log(929))
model_back_bic_b
summary(model_back_bic_b)

model_back_bic_f <- stepAIC(modele_complet, scope = list(lower = . ~ 1,
upper = . ~ .),
                            direction = "forward",k=log(929))
model_back_bic_f
summary(model_back_bic_f)

model_back_bic_both <- stepAIC(modele_complet, scope = list(lower = .
~ 1, upper = . ~ .),
                               direction = "both",k=log(929))
model_back_bic_both
summary(model_back_bic_both)


model_back_bic
summary(model_back_aic_f)

# modele optimal ----
```

```
modele_cox<-coxph(Surv(T_i, Retino_vrai==1) ~ AgeDecDiabete +
var_imc + trigly_moy +
       trigly_min + trigly_var + HbA1c_moy + HbA1c_min,data = Diab)

## test des variables du modele opti ----
# Liste de variables à tester
variables_opt <- c("AgeDecDiabete", "var_imc", "trigly_moy",
"trigly_min", "trigly_var", "HbA1c_moy", "HbA1c_min")
res_opt <- data.frame("var" = numeric(0), "test_anova" =
numeric(0),"nombre_patient"=numeric(0))
# Boucle for pour retirer une variable à chaque itération
for (i in 1:length(variables)) {
  # Sélection de toutes les variables sauf la variable à retirer
  x<-variables_opt[i]
  variables_selectionnees <- variables_opt[-i]

  # filtrer les lignes avec des valeurs manquantes pour la colonne "col3"
  ligne_complete <- !is.na(Diab[,x])

  # sélectionner les lignes complètes pour la colonne "col3" et toutes les
colonnes
  data_filtre <- Diab[ligne_complete, ]
  # Construction de la formule de survie en utilisant les variables
sélectionnées
  formule_survie <- as.formula(paste("Surv(T_i, Retino_vrai == 1) ~",
paste(variables_selectionnees, collapse = " + ")))

  # Application de la régression de Cox à la formule de survie
  mod <- coxph(formule_survie, data = data_filtre)
  mod_full <- coxph(Surv(T_i, Retino_vrai==1) ~ AgeDecDiabete +
var_imc + trigly_moy +
                trigly_min + trigly_var + HbA1c_moy + HbA1c_min,
              data = data_filtre)
```

```
  a <- anova(mod,mod_full)

  # ajouter une ligne de valeurs
  nouvelle_ligne <- c(x, a[2,4], nrow(data_filtre))
  res_opt <- rbind(res_opt, nouvelle_ligne)
}
names(res_opt) <- c("var", "test_anova","nombre_patient")
res_opt

summary(modele_cox)

# Courbe par patient ----
par(mfrow=c(1,1))
# Sélectionner les données pour le patient 8
patient_8 <- Diab[Diab$IdPatient == 8, ]

# Calculer la courbe de survie pour l'ensemble des patients
survival_curve_all <- survfit(modele_cox, newdata = Diab)

# Calculer la courbe de survie pour le patient 8
survival_curve_8 <- survfit(modele_cox, newdata = patient_8)

# Afficher les patients et le patient 8 sur un même graphique avec une
légende
plot(survival_curve_all, xlab = "Temps", ylab = "Probabilité de survie", col
= "blue")
lines(survival_curve_8, col = "red")
legend("topright", legend = c("Tous les patients", "Patient 8"), lty = 1, col =
c("blue", "red"))
title("patient 8")

# Afficher la courbe de survie pour le patient 8
```

```
par(mfrow=c(1,1))
plot(survival_curve_8, xlab = "Temps", ylab = "Probabilité de survie", col
= "red")
title("patient 8 ")
```