Étude des caractéristiques déterminant la prise de décision dans le traitement du cancer des voies biliaires

Junyi Zhao

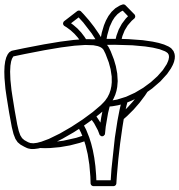
AGROCAMPUS OUEST

☐ **CFR Angers** ☑ **CFR Rennes**

Année universitaire : 2020 - 2021 Spécialité : Ingénieur agronome Spécialisation (et option éventuelle) : Sciences des données	**Mémoire de fin d'études** ☒ d'ingénieur d'AGROCAMPUS OUEST (École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage), école interne de L'institut Agro (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de master d'AGROCAMPUS OUEST (École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage), école interne de L'institut Agro (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de Montpellier SupAgro (étudiant arrivé en M2) ☐ d'un autre établissement (étudiant arrivé en M2)

Etude des caractéristiques déterminant la prise de décision dans le traitement du cancer des voies biliaires.

Par : Junyi ZHAO



Soutenu à Rennes le 6 septembre 2021

Devant le jury composé de :

Président : David CAUSEUR

Maître de stage : Audrey SCHMITT

Enseignant référent : David CAUSEUR

Autres membres du jury (Nom, Qualité) :

Sébastien LÊ, Enseignant

Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle d'AGROCAMPUS OUEST

Remerciements

J’aimerais remercier Mme Geneviève Bonnelye, CEO de Kantar Health, de m’avoir donné l‘opportunité de réaliser mon stage au sein de Kantar Health. Je remercie également Mme Audrey Schmitt, pour m’avoir confié diverses missions, et m’avoir aidé à développer mes compétences professionnelles et relationnelles.

Mes plus chaleureux remerciements vont à Mme Sabine Rizzo, ma maitresse de stage, pour écoute et sa tutelle au cours mon stage.

Je remercie également Mme Dalia Thome et Mme Ilia Triantafyllidy pour leurs conseils et leur disponibilité.

Je remercie aussi mon tuteur de stage, M. David Causeur, pour son suivi le long du stage.

Sommaire

Table des figures

Table des tableaux

Liste des abréviations :

ACM: Analyse des Correspondances Multiples
AFM : Analyse Factorielle Multiple
IMC : Indice de masse corporelle
OOB out-of-bag

Introduction

Le cholangiocarcinome, ou cancer des voies biliaires, est un cancer rare et complexe : il comprend moins de 1% des patients atteints de cancer et entre 10 à 15% des cancers primaires du foie, et inclut plusieurs sous-types (ESMO, 2016). Mondialement, il s'agit du 20ème cancer le plus fréquent seulement et la 17ème cause de décès par cancer. L'incidence du cancer des voies biliaires est de 0,35 cas pour 100 000 à 2 cas pour 100 000 par an, mais peut être jusqu'à 40 fois supérieur dans les régions endémiques (Lancet, 2021).

La complexité de ce cancer vient du fait qu'il comporte plusieurs sous-types, définis en fonction de la localisation de la tumeur primaire dans les voies biliaires : cholangiocarcinome intra-hépatique, cholangiocarcinome péri-hilaire, cholangiocarcinome distal, cancer de la vésicule biliaire ou carcinome de l'ampoule de Vater. Ces sous-types ont des caractéristiques, des étiologies et des incidences différentes, mais sont étudiés ensemble dans le cadre de ce projet car les données sur le cancer des voies biliaires sont encore rares et ne permettent pas de distinguer plus finement ces différentes indications. Mais également parce que le diagnostic du cancer des voies biliaires se fait généralement lorsque la tumeur atteint déjà un stade avancé, ou bien un stade métastatique. Le diagnostic du cancer des voies biliaire est tardif, notamment parce que les symptômes ne sont pas apparents dans les stades primaires de la maladie. On compte en effet très peu de patients diagnostiqués à un stade précoce, un tiers à un stade localement avancé et deux tiers à un stade métastatique (Kantar Health, 2021).

De plus, la diversité géographique des facteurs de risque pour les sous-types de cancer des voies biliaires entraîne de grandes différences dans l'incidence mondiale de chaque cancer. Les régions endémiques du cancer des voies biliaires sont les régions d'Asie et d'Amérique du Sud, avec une incidence supérieure aux régions d'Europe de l'Ouest et d'Amérique du Nord.

Kantar Health est entreprise réalisant des études de marché et sondages, spécialisée en oncologie. Ses principaux clients sont les laboratoires pharmaceutiques. Kantar les accompagne dans leur stratégie de R&D, d'accès au marché de leurs molécules et dans la commercialisation de leurs produits. L'étude portant sur le cancer des voies biliaires est réalisée aux Etats-Unis, au Japon, en Chine, en Allemagne et en Italie.

Nous souhaitons donc analyses les variables les plus importantes pour le choix de traiter ou non un patient, et pour le choix du traitement qui lui sera prescrit.

Afin de répondre à cette problématique, nous allons dans un premier temps présenter le contexte et les enjeux de l'étude, puis prédire si le patient sera traité ou non. Enfin, une comparaison des pays participant à l'étude sera effectuée grâce à la méthode de l'Analyse des Correspondances Multiples.

I) Contexte et enjeux de l'étude

1) Le traitement actuel du cancer des voies biliaires

Le cancer des voies biliaires est un cancer détecté tard. Les facteurs de risque multiples, dépendamment de l'emplacement de la tumeur primaire. Généralement, une cirrhose, une prédisposition génétique ou encore une infection parasitaire peuvent être facteurs de risque. Le patient moyen a plus de 50 ans, avec un statut rénal normal a modérément insuffisant, des métastases hépatiques et pas d'infection ou d'obstruction biliaire. Le taux de survie médian reste très faible : il est inférieur à 12 mois. Souvent le choix de meilleurs soins de soutien est fait pour les patients dont la maladie progresse.

La résection chirurgicale est aujourd'hui la meilleure solution thérapeutique pour les patients atteints de cancer des voies biliaires à un stade précoce. Les patients en stade localement avancé ou métastatique sont majoritaires mais ne sont souvent pas candidats à la résection chirurgicale parce que découverte du cancer se fait tard et a ainsi le temps de proliférer dans tout le corps. Les options thérapeutiques restent limitées, de plus ce n'est pas un cancer commun, les médecins ont donc des difficultés à prescrire un traitement du fait qu'il y a peu de cas précédents.

Dans le cadre de ce projet, nous nous intéressons plus particulièrement au cancer des voies biliaires de stade localement avancé à métastatique. Les traitements proposés sont des chimiothérapies, mais on observe également un essor dans les immunothérapies, les essais cliniques et les thérapies ciblées.

2) Objectifs de l'étude

L'étude s'est déroulée dans cinq pays : aux Etats-Unis, au Japon, en Chine, en Allemagne et en Italie, entre mars et mai 2021. Les enjeux de l'étude pour le client sont de mieux comprendre le marché actuel et de capturer les tendances dans le traitement du cancer des voies biliaires, à un niveau global. Ceci leur permet notamment de préparer le lancement de leur produit, dans ce rapport, noté produit X.

L'étude a permis de recueillir beaucoup de variables dans cinq pays différents, ce qui permet de comprendre les différents niveaux de prescriptions et les traitements standards actuels. Nous nous intéressions plus précisément à la prise de décision de traiter ou non les patients : sur quelles caractéristiques se fonde-t-elle ? Ensuite, nous nous intéressons aux leviers de la décision de prescription du traitement, au niveau du patient.

Il est aussi intéressant de voir si, à partir des variables recueillies sur les patients et sur les médecins, il est possible de visualiser les similarités et les différences entre les pays, et également voir s'il est possible de caractériser chaque pays. Et déterminer les spécificités de chaque pays en matière de connaissances sur le cancer des voies biliaires.

3) Le questionnaire de l'étude

L'étude est réalisée sous la forme d'un questionnaire en ligne envoyé à des médecins spécialistes. Le questionnaire de l'étude comporte un court questionnaire permettant de recueillir les attitudes et usages des médecins ainsi que leur point de vue et leurs connaissances des traitements proposés, puis de formulaires de dossiers patients. Enfin vient une partie permettant aux médecins d'évaluer les trois produits proposés et leur intention de prescription de chaque produit, sur une échelle de 1 à 10.

Au tout début de l'enquête, des questions de sélection sont posées pour vérifier l'éligibilité du médecin à l'enquête. Elles permettent également de fournir des informations de signalétique, comme la spécialité médicale principale des médecins, le nombre d'années de pratique pour évaluer leur expérience dans le métier, le niveau de l'hôpital où ils exercent, et la région dans laquelle se situe leur lieu d'exercice par exemple.

Les questions dites d'attitude et usages sont composées d'une partie concernant :

- les besoins insatisfaits de l'indication, sur une échelle de Likert allant de 1 à 5, puis avec les options « pas du tout un besoin insatisfait » et « un besoin insatisfait très important » Le médecin peut également choisir l'option « il n'y a pas de besoins insatisfaits dans cette indication ».

- la connaissance des médecins sur les essais cliniques récents, avec des réponses spontanées, puis un choix des options dans un menu déroulant,

- la perception et les attitudes envers le traitement actuel des patients, leur degré de satisfaction sur une échelle de 1 à 10, avec « pas du tout satisfait » allant jusqu'à « très satisfait »

- leur participation à des réunions multidisciplinaires, durant lesquelles les médecins discutent des dossiers de leurs patients.

Les caractéristiques des patients sont recueillies dans le dossier patient. Il est composé de plusieurs parties, concernant les caractéristiques physiologiques et les données cliniques des patients. Les données recueillies pour les patients traités et les patients en soins palliatifs sont différentes : les traitements actuels pour les patients traités, et les traitements antérieurs et la cause de l'arrêt des soins pour les patients en soins palliatifs.

4) Présentation des données recueillies

Les données du questionnaire sont séparées en deux jeux de données : les données médecins et les données patients. Les données médecin comptent 413 individus en lignes et 323 variables en colonnes. Ici, chaque variable correspond à une question du questionnaire. L'échantillon pour chaque pays est le suivant : 97 médecins répondant aux Etats-Unis, 104 au Japon, 99 en Chine, 64 en Allemagne et 49 en Italie.
Nous observons qu'il y a un nombre plus important de médecins répondant au questionnaire aux Etats-Unis, au Japon et en Chine : le nombre d'habitants y est plus élevé et l'indication est plus connue dans les pays asiatiques.

Quant aux données patients, chaque médecin remplit entre un et sept questionnaires

patient pour les patients traités, et jusqu'à deux questionnaires pour ses patients en soins palliatifs, dépendamment du nombre total de ses patients. Nous avons ainsi recueilli des informations sur 2821 patients. Il y a en tout 505 variables sur la signalétique, les caractéristiques physiologiques et le diagnostic des patients. L'échantillon pour chaque pays est le suivant : 550 médecins répondant aux Etats-Unis, 622 au Japon, 771 en Chine, 477 en Allemagne et 401 en Italie.

II) Prédiction du choix de traiter ou non un patient

1) Objectif de la prédiction

Nous souhaitons connaitre les caractéristiques qui différencient les patients traités des patients placés en soins palliatifs, ceci afin de pouvoir aisément distinguer les patients qui constituent une cible potentielle du nouveau traitement, des patients qui ne seront pas ciblés. Nous voulons également obtenir l'importance des caractéristiques des patients dans le choix du traitement.

2) Méthodologie mise en place

La variable que nous souhaitons prédire est une variable catégorielle à deux modalités : le patient est traité (Y = 1) ou non traité (Y = 0). Il y a 15 variables prédictives avec au total 53 modalités.

La méthode de prédiction choisie est la méthode *randomForest*. C'est une méthode d'agrégation d'un grand nombre de modèles d'algorithmes d'arbres de décision. Il sélectionne plusieurs modèles d'arbres de décision dans un tirage avec remise, et permet ainsi de construire un nœud de l'arbre sur un sous-ensemble de variables tirées aléatoirement (Breiman, 2001). La classe prédite est la classe la plus fréquemment observée parmi les prédictions. L'erreur de prédiction est estimée par la mesure de l'out-of-bag (OOB). Il s'agit de l'erreur de prédiction moyenne sur chaque échantillon d'apprentissage x_i, utilisant seulement les arbres qui n'ont pas l'échantillon x_i dans leur échantillon de bootstrap.
C'est une méthode qui est très performante, mais plutôt difficilement interprétable.

Nos données sont déséquilibrées : sur le nombre total de patients, 2183 individus sont traités (77%) contre 638 individus non traités (23%). Nous choisissons donc de mettre en œuvre une méthode d'oversampling, qui permet d'augmenter les observations de la classe minoritaire pour que le jeu de données soit artificiellement équilibré. Ceci permet de garder les individus de la classe à prédire. Nous utilisons pour cela le package *ROSE*.

La méthode *randomForest* comporte deux paramètres qui peuvent être optimisés, pour obtenir une meilleure performance de prédiction. Ce sont les paramètres *ntree* et *mtry*. Le paramètre *ntree* correspond au nombre d'arbres à faire pousser. Le paramètre *mtry* correspond au nombre de variables aléatoirement échantillonnés à prendre en compte à chaque nœud.
Nous optimisons d'abord le paramètre *ntree*, en effectuant plusieurs essais et en gardant celui qui minimise l'erreur OOB. Puis avec le paramètre *ntree* fixé, nous obtenons de la

même manière une valeur pour le paramètre *mtry*.

La fonction *varImpPlot* du package *randomForest* permet d'obtenir les variables par ordre d'importance. Les variables avec un MeanDecreaseAccuracy les plus important seront celles qui discriminent le plus les patients traités des patients non traités, et donc permettent d'améliorer la prédiction.

Enfin, la matrice de confusion permet d'évaluer la performance et la qualité du modèle.

	Référence	
Prédits	0	1
0	Vrais négatifs	Faux négatifs
1	Faux positifs	Vrais positifs

Tableau 1 : Exemple de matrice de confusion

3) Mise en œuvre du modèle et résultats

Nous séparons le jeu de données initial en un jeu de données d'apprentissage et un jeu de données test. Nous choisissons aléatoirement 70% des individus pour le jeu de données d'apprentissage et 30% pour le jeu de données test. Les variables choisies sont toutes les variables caractéristiques des patients.
Nous effectuons ensuite les différentes étapes citées en partie II) 2.

Tout d'abord, il est nécessaire d'équilibrer les données. En effet, si nous gardions les données non équilibrées, il y a un risque de sur-apprentissage de la classe majoritaire.

Avec la méthode d'oversampling, le gonflement artificiel des individus non traités permet d'apprendre le modèle et de prédire sur des données équilibrées. L'erreur OOB est de 18.08% lorsqu'on apprend le modèle sur des données non-équilibrées, et est diminuée à 9.82% lorsque les données sont équilibrées.

Quant à la prédiction, nous observons qu'en équilibrant les données, le Balanced Accuracy est de 90%, pour une Accuracy de 90%, alors que lorsque les données étaient déséquilibrées, le Balanced Accuracy était de 68%, pour une Accuracy de 84%.

	Référence	
Prédits	Traités	Non traité
Traités	638	99
Non traités	30	79

Tableau 2 : Matrice de confusion sur la prédiction effectuée sur données non équilibrées

	Référence	
Prédits	Traités	Non traité
Traités	358	23
Non traités	61	404

Tableau 3 : Matrice de confusion sur la prédiction effectuée sur données équilibrées

Finalement, les données permettant de mesurer la performance du modèle sont, en comparant les résultats obtenus avec des données déséquilibrées et équilibrées :

	Sans oversampling	Avec oversampling
Accuracy	84%	90%
Sensibilité	96%	85%
Spécificité	44%	95%
Précision	85%	95%

Tableau 4 : Résultats des modélisations avec données déséquilibrées et données équilibrées

Le réequilibrage des données est ainsi un étape qui semble nécessaire pour la bonne précision de la prédiction.

- Etape d'optimisation des paramètres du modèle

Le modèle initial a pour paramètres par défaut les valeurs de *ntree* = 500 et *mtry* = 7.

Nous optimisons d'abord le paramètre *ntree*, en utilisant la fonction *err.rate* et en réalisant le graphique montrant l'évolution de l'erreur OOB en fonction du nombre d'arbres pris en compte par le modèle.

Ensuite, le paramètre mtry est optimisé grâce à la fonction tuneRF du package randomForest. Un graphique est également réalisé pour montrer l'évolution de l'erreur OOB en fonction du paramètre mtry, avec une valeur de ntree fixe.

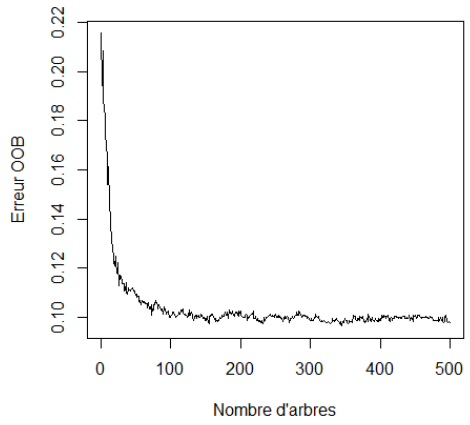


Figure 1 : Optimisation du paramètre ntree

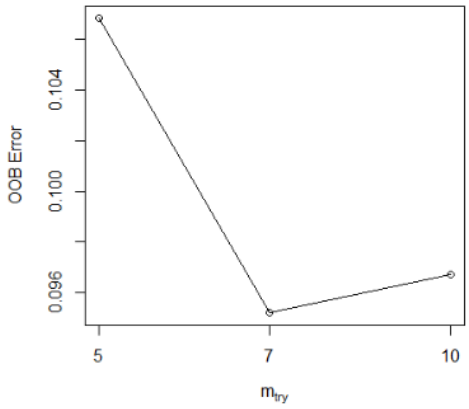


Figure 2 : Optimisation du paramètre mtry

Après plusieurs répétitions, nous obtenons comme valeurs optimales :

	ntree	mtry	Erreur OOB
Sans paramètre optimisé	500	7	9.92%
ntree optimisé	250	7	9.47%
ntree et mtry optimisés	250	6	9.42%

Tableau 5 : Optimisation des paramètres ntree et mtry

L'optimisation des paramètres du modèle a permis d'obtenir une erreur OOB plus faible, mais finalement elle n'apporte pas beaucoup de diminution d'erreur. Nous choisissons quand même les paramètres optimisés pour la suite de l'étude du modèle.

- Etape de prédiction

L'étape de prédiction, réalisée avec le jeu de données test, permet de tester si le modèle bien appris, et s'il est performant et de qualité.
Nous obtenus les résultats suivants, pour le modèle avec les données équilibrées artificiellement, et les paramètres ntree et mtry optimisés.

	Observés	
Prédits	Traités	Non traité
Traités	348	33
Non traités	50	415

Tableau 6 : Matrice de confusion sur la prédiction effectuée

L'Accuracy obtenue pour ce modèle est de 90%. Elle mesure la précision du modèle donc ici nous avons un modèle performant.
La sensibilité, mesure du nombre de vrais positifs, ici des patients réellement traités, est de 87%. Ceci décrit le fait que les patients prédits traités le sont effectivement. La spécificité est de 93%. C'est la mesure des vrais négatifs, elle évalue le nombre d'individus prédits non traités qui sont effectivement non traités.

Globalement, le modèle est performant et précis, ce qui permet d'estimer les variables discriminant le plus les patients traités des patients placés en soins palliatifs.

- Etape de mesure de l'importance des variables

Nous utilisons l'estimation du MeanDecreaseAccuracy, qui explique la perte de précision par la suppression d'une variable dans le modèle. Plus l'Accuracy décroit, plus la variable est importante dans le modèle. Les variabels sont présentées dans l'ordre decroissant d'importance. Nous préférons cette mesure à la mesure du MeanDecreaseGini, qui mesure de combien chaque variable contribue à l'homogénéité des nœuds dans la forêt aléatoire résultante. Plus la valeur du MeanDecreaseGini est grande, plus la variable est importante dans le modèle.

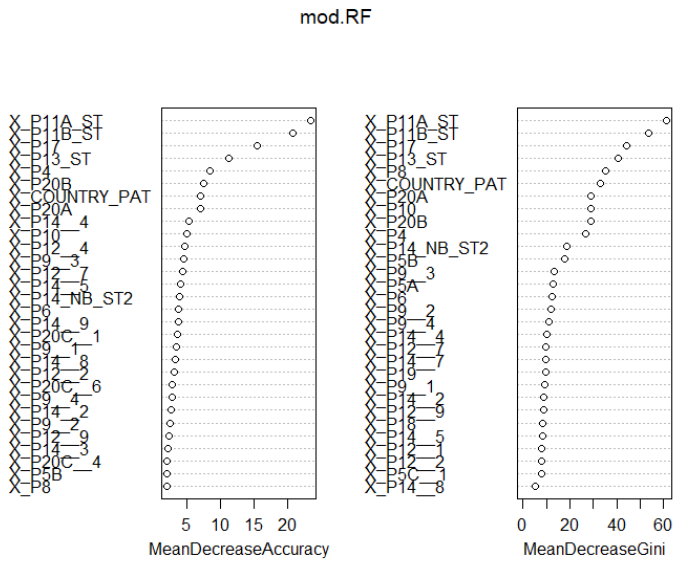


Figure 3 : Ordre d'importance des variables

Les variables qui ont le plus d'importance en termes du MeanDecreaseAccuracy sont l'ECOG, le statut rénal, le child-pugh, l'âge du patient et le statut MSI.

Effectivement, l'ECOG est un score de performance qui évalue l'état général des patients. C'est un indicateur qui est très discriminant pour la décision de suivi du traitement ou non.

Nous souhaitons effectuer une comparaison par pays pour savoir s'il y a des caractéristiques cliniques des patients plus importantes qui dirigent les décisions du médecin pour le traitement.

				
ECOG	ECOG	ECOG	ECOG	ECOG
Statut renal	Renal status	Child-pugh	Statut renal	Child-pugh
Comorbidité – cirrhose	Child-pugh	Statut renal	Mutation génétique 8	Statut renal
Age	Nombre de comorbidités	Age	Comorbidité – maladie rénale	IMC
Etiologie - cholangite sclérosante primitive	MSI status	MSI status	Nombre de comorbidités	Nombre de comorbidités

Tableau 7 : Ordre d'importance des variables selon le MeanDecreaseAccuracy

Les résultats nous montrent que l'ECOG reste l'indicateur le plus important pour le suivi du traitement. Le statut rénal et l'âge sont également des facteurs très discriminants pour savoir si le patient sera traité ou placé en soins palliatifs.

4) Discussion

Nous avons montré ici l'importance d'obtenir des données équilibrées, afin de pouvoir améliorer la performance de prédiction. En effet, l'échantillon est biaisé en faveur des patients traités activement qui sont plus jeunes, plus en forme et avec une meilleure fonction rénale que les patients placés en soins palliatifs. Ceci provient du fait que les médecins remplissent jusqu'à sept formulaires pour les patients traités, et seulement deux pour les patients en soins palliatifs. Cependant, nous avons pu obtenir une bonne performance de prédiction et obtenir les informations les plus importantes, qui discriminent le plus les patients traités des patients non traités. Ces informations, au niveau patient, sont les mêmes à travers les cinq pays de l'étude. Nous pouvons donc conclure que la décision de traitement se fait principalement avec les caractéristiques des patients. Mais certains patients subissent aussi un impact financier, notamment aux Etats-Unis et en Chine. Les inégalités dans l'accès aux assurances et donc le refus de la part du patient sont également un frein au traitement.

III) **Segmentation des patients pour définir la cible du produit X**

Nous souhaitons maintenant définir les variables les plus significatives pour segmenter les patients, et ainsi trouver les patients cibles pour le produit X. Pour cela, nous allons d'abord mettre en œuvre une méthode de sélection des données, avec un modèle de classification et une sélection pas à pas, minimisant le critère d'Akaike, puis effectuer une Analyse des Correspondances Multiples comme nous avons de très nombreuses variables, pour réduire et synthétiser l'information. Enfin, à la suite de la réalisation de l'ACM, nous effectuerons une Classification Ascendante Hiérarchique afin de classifier les patients en fonction de leur disposition à recevoir le produit X comme traitement futur.

1) Sélection des données

a) Objectifs

Les variables récupérées à partir du formulaire patient sont nombreuses et nous souhaitons à présent définir celles qui sont les plus significatives. Ici nous gardons seulement les informations des patients traités, comme les patients en soins palliatifs ne reçoivent pas de traitement futur. Ceci permet de diminuer le nombre de variables et d'inclure celles qui sont les plus significatives dans les modèles de réduction de dimension réalisés ensuite.

Les données des patients sont :

Patient demographics and diagnosis

- Gender
- Age
- Insurance coverage
- Stage at diagnosis
- Date of diagnosis
- Primary tumor location
- Metastatic sites

Current treatment

- Current treatment status, distinguish between actively treated vs BSC
- Line of treatment
- Type of active therapy
- Treatment goal for 1L
- Multidisciplinary tumor board

Patient characteristics

- Weight (current and at initiation of treatment)
- Etiology
- ECOG status
- Comorbidities/risk factors (by disease subtypes): chronic obstructive pulmonary disease, coronary heart disease, cerebrovascular disease, hypertension, autoimmune disease, diabetes, primary sclerosing cholangitis (PSC), primary biliary cirrhosis (PBC), congenital malformations such as choledochal cysts and multiple biliary papillomatosis
- PDL1
- Genetic mutations (IDH, FGFR, EGFR, KRAS…)

Figure 4 : Variables du formulaire patient

Afin de choisir les variables qui expliquent le mieux le traitement futur reçu, nous mettons en place une méthode de sélection des données.

b) Méthode retenue pour la sélection des données

Nous sélectionnons les variables qui prédisent le traitement que le patient sera le plus apte à recevoir. La variable réponse notée Y est une variable catégorielle à 4 classes : produit X, produit Y, produit Z ou encore traitement actuel.
Les variables explicatives sont les variables citées précédemment.

Le modèle logistique s'écrit donc :

$$\log\frac{P(Y=C_2)|(X=x)}{P(Y=C_1)|(X=x))} = \beta_0^{(2)} + \beta_1^{(2)}x,$$
$$\log\frac{P(Y=C_3)|(X=x)}{P(Y=C_1)|(X=x))} = \beta_0^{(3)} + \beta_1^{(3)}x,$$
$$\log\frac{P(Y=C_4)|(X=x)}{P(Y=C_1)|(X=x))} = \beta_0^{(4)} + \beta_1^{(4)}x$$

où $\beta_0^{(k)}$, $\beta_1^{(k)}x_1$, ..., $\beta_p^{(k)}x_p$ sont les p(K - 1) paramètres du modèle

Le modèle de régression logistique est réalisé avec la fonction *multinom* du package *nnet*, puis la sélection effectuée avec la fonction *stepwise* du package *RcmdrMisc.* Cette fonction permet de réaliser une sélection des variables explicatives lorsque le nombre de variables est supérieur à 15. Elle permet également de sélectionner pas à pas : en allant en avant (forward) et en arrière (backward), ce qui permet d'être encore plus précis dans la sélection. Le critère que nous choisissons est le critère d'Akaike. Le modèle choisi est le modèle avec le critère d'Akaike le plus faible, donc à chaque étape, la variable qui augmente le critère d'Akaike sera éliminée du modèle. Les mêmes étapes sont réalisées dans chacun des pays de l'étude.

Les résultats obtenus sont les suivants :

Les variables retenues par le modèle sont : le statut ECOG, l'âge, le statut child-pugh, les maladies rénales.

Les variables retenues par le modèle sont le statut ECOG, l'indice de masse corporelle, le statut child-pugh, les métastases au péritoine (membrane séreuse qui tapisse l'abdomen, le pelvis et les viscères, délimitant l'espace virtuel de la cavité péritonéale), les calculs biliaires, le statut de la tumeur (localement avancé ou métastatique) au diagnostic, l'emplacement de la tumeur primaire, l'infection biliaire, le statut rénal.

Les variables retenues par le modèle sont l'âge, le child-pugh, les métastases au foie, le statut MSI, le statut ECOG, la cirrhose, le nombre total de comorbidités, les métastases au péritoine, l'infection biliaire.

Le statut ECOG, les maladies rénales, le statut child-pugh, les maladies cardiovasculaires, les métastases au péritoine, la cirrhose, l'abus d'alcool.

Les variables retenues par le modèle sont le statut ECOG, le statut child-pugh, l'infection biliaire, les métastases au péritoine, le statut PD-L1 (positif ou négatif), les comorbidités cardiovasculaires, le statut rénal, le statut de la tumeur au diagnostic.

Résultats

Nous essayons de classer les patients selon les produits qui leur seront prescrits. Tout d'abord, le nombre de variables sélectionné est différent selon le pays. Cela peut s'expliquer par le fait que le nombre de patient qui recevront le produit X qui est différent dans chaque pays. Ces résultats sont finalement les mêmes variables que celles obtenues avec la mise en œuvre de l'algorithme de *RandomForest* et sont globalement les mêmes entre les pays.

Discussion

Entre les pays, les variables expliquant le traitement sont les mêmes.
Est-ce que ces variables permettent de bien différencier les traitements que recevront les patients et permettent de classer les patients ?

2) Segmentation des patients

L'objectif initial était de distinguer s'il y avait des variables qui différencient la décision de prescription des médecins et sont distinctes entre les pays. Pour cela, une Analyse Factorielle Multiple pourrait être réalisée, mais l'inertie total était trop faible pour pouvoir interpréter les résultats. C'est pourquoi nous réalisons une Analyse des Correspondances Multiples, afin de remplir les objectifs suivants : étudier les individus qui se ressemblent, les variables qui distinguent les individus et les modalités de ces variables. Dans un second temps, une Classification Ascendante Hiérarchique sera réalisée afin de séparer les individus en groupes, avec les individus qui se ressemblent le plus dans le même groupe et en maximisant les différences entre les groupes.

a) Réalisation de l'ACM

L'ACM est une méthode de réduction de la dimension, elle permet de synthétiser l'information contenue dans les données sur quelques plans principaux. Elle permet d'étudier des jeux de données multivariées avec des variables qualitatives.

Une ACM est réalisée pour chaque pays, et inerties sont assez élevées pour qu'elles soient interprétables. Globalement, les axes sont séparés en fonction du nombre de comorbidités sur l'axe 1 et du statut rénal sur l'axe 2. Elle consiste ici à un prétraitement pour la classification.

b) La construction de la classification et segmentation des patients

La fonction utilisée est la fonction HCPC du package FactoMineR. Elle effectue une classification hiérarchique sur composantes principales. C'est une méthode non supervisée qui va regrouper les individus comportant des caractéristiques communes. Elle a pour objet le résultat d'une analyse factorielle, ici les ACM préalablement construites. La distance de Ward calculée par cette fonction permet de maximiser l'inertie intra-classe. Le nombre de classes peut être défini par le dendogramme construit par la fonction, la coupure proposée, permettant de maximiser le gain d'inertie, est choisie.

Résultats :

Aux Etats-Unis, la classification définit trois groupes distincts et un groupe non homogène.

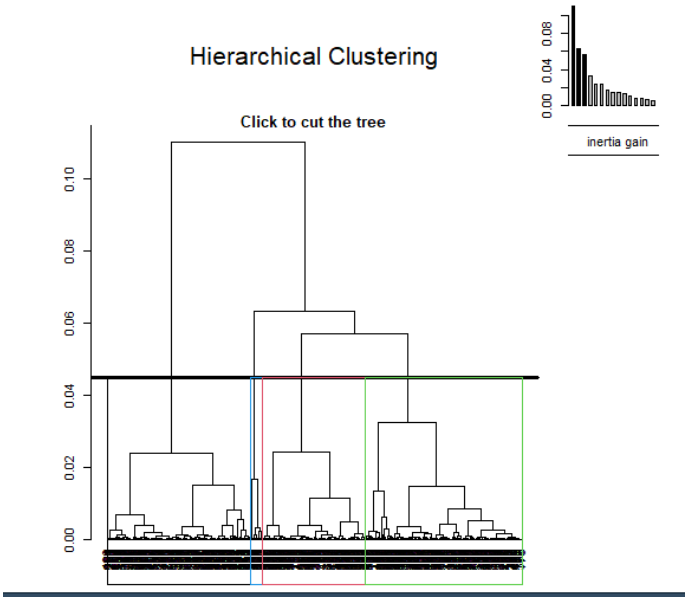


Figure 5: Dendogramme associé à la classification des patients aux Etats-Unis

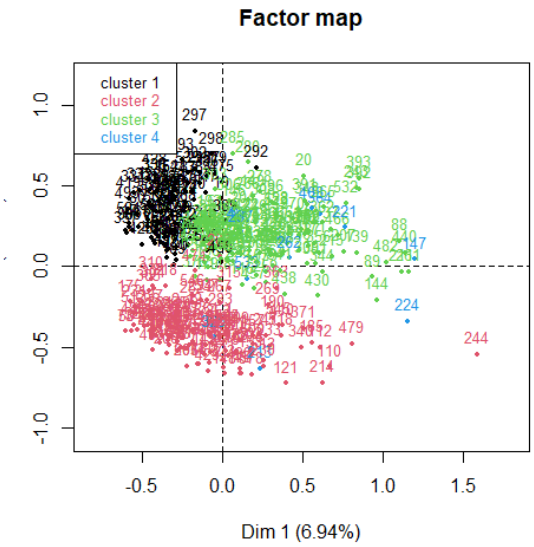


Figure 6 : Groupes issus de la classification

Au Japon, la classification permet également de trouver trois groupes bien distincts.

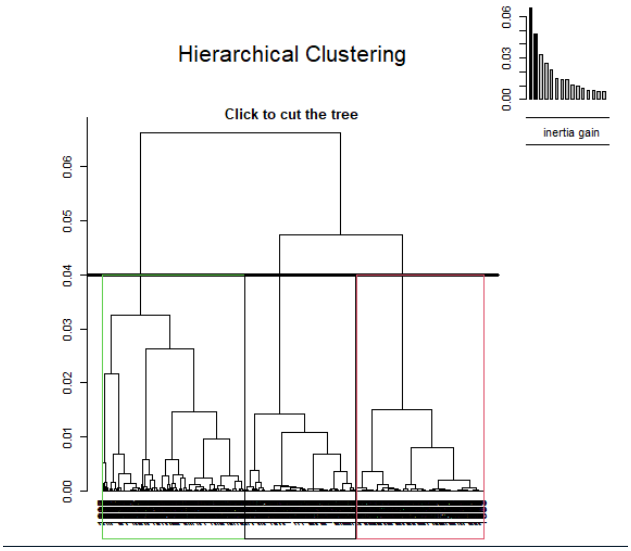


Figure 7: Dendogramme associé à la classification des patients aux Etats-Unis

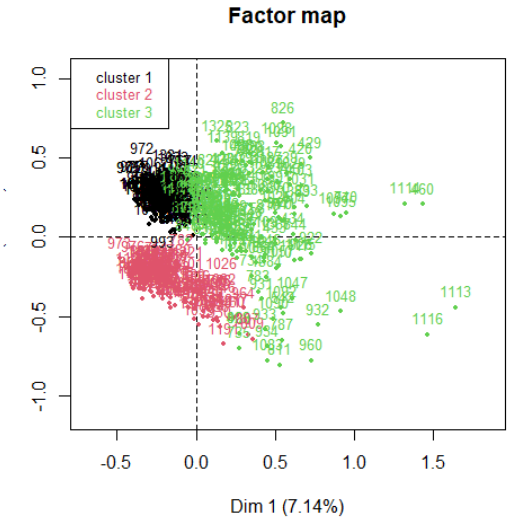


Figure 8 : Groupes issus de la classification

Dans ces deux pays, des profils communs se dégagent.

Groupe 1 : Ce groupe comporte 25% des individus aux Etats-Unis et 33% au Japon.

Ils sont caractérisés par un très bon statut en général, pas de comorbidités, un ECOG de 0, un statut rénal normal, un IMC sain, et un child-pugh de A. Ce groupe est peu probable à recevoir le produit X.

Groupe 2 : Ce groupe comporte 34% des patients aux Etats-Unis et 30% des patients au Japon.
Aux Etats-Unis, il s'agit de patients qui ont plus de 70 ans mais restent en bonne forme, avec un ECOG de 0-1 et un statut rénal normal. Au Japon, ce sont des patients qui sont également en bonne santé, mais qui ont au moins une caractéristique moins idéale, comme une comorbidité ou une fonction rénale altérée. Un segment pour le produit X est clairement démarqué ici, il s'agit d'un groupe qui constitue une opportunité pour le client.

Groupe 3 : Ce groupe est composé de 38% des individus aux Etats-Unis et 37% des individus au Japon. Il s'agit d'un groupe avec une ou deux comorbidités, plus âgés, avec un statut rénal défaillant, un ECOG 2+. Ce groupe ne reçoit pas le produit X car leur état est très dégradé.

Un 4ème groupe se démarque aux Etats-Unis, il s'agit d'un groupe constitué d'individus présentant une obésité. Ils ont été regroupés car cette caractéristique commune n'est pas présente chez les autres patients.

Discussion sur ces profils :
Contrairement à ce que l'on pourrait penser, le segment qui représenterait une opportunité pour le produit X n'est pas constitué des patients les plus en forme, qui pourraient tolérer un traitement sans doute plus agressif, mais de patients certes en relative bonne santé mais présentant un statut plus dégradé. Enfin, un dernier segment est composé des individus les plus frêles et peu propices à recevoir un traitement différent et vont plus rester sur le traitement actuel. Dans l'analyse, nous rencontrons un 4ème groupe qui se démarque et comporte des individus avec une caractéristique déterminante, qui les discrimine des

autres patients.

En EU2 (Allemagne et Italie), la fonction propose un découpage en 6 classes.

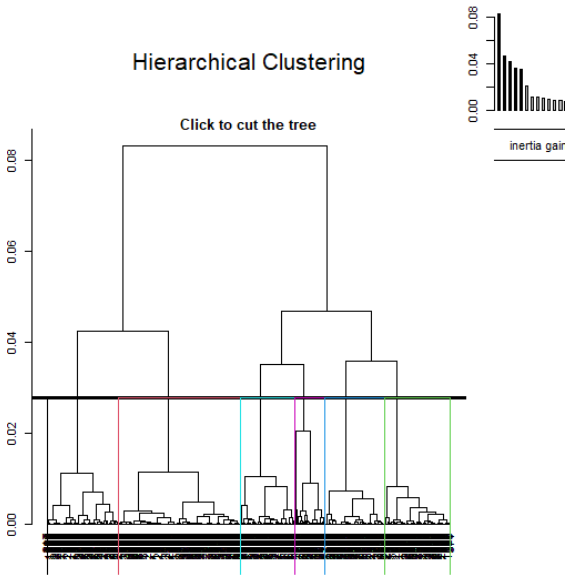


Figure 9: Dendogramme associé à la classification des patients aux Etats-Unis

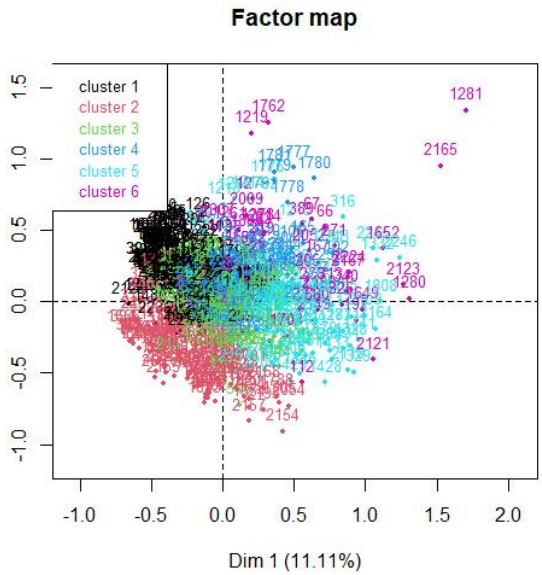


Figure 10 : Groupes issus de la classification

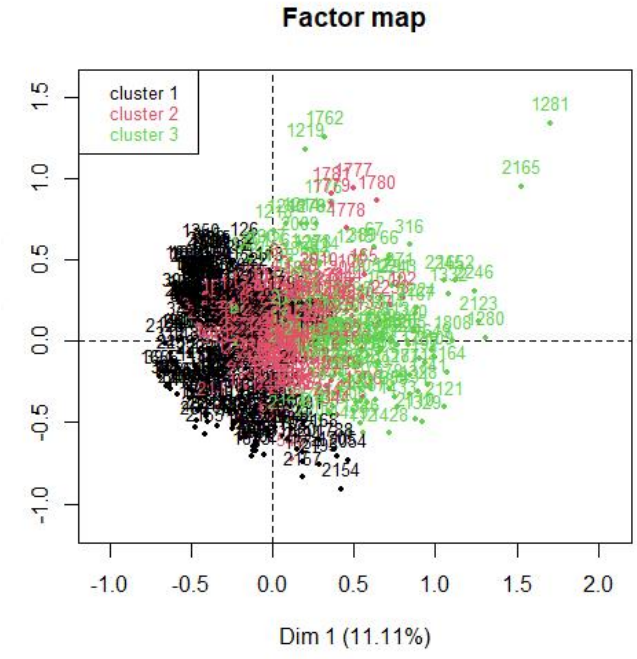


Figure 11: Dendogramme associé à la classification des patients aux Etats-Unis

En essayant un découpage en 3 classes, nous voyons que ces trois classes se recoupent encore de façon importante.

En Chine, le découpage en 3 classes résulte des groupes qui se superposent également.

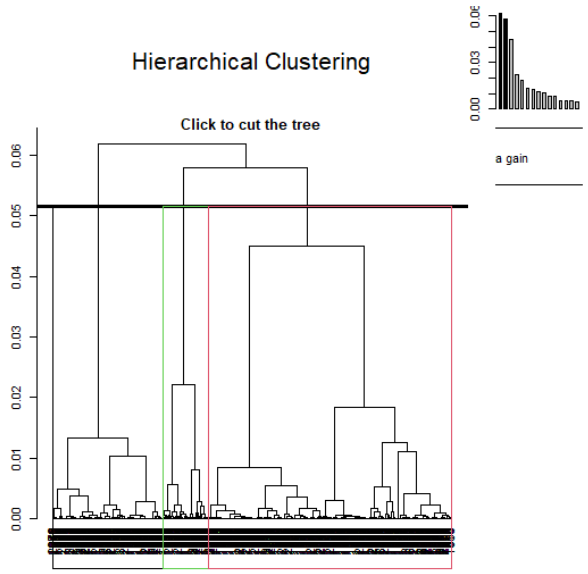


Figure 12: Dendogramme associé à la classification des patients en Chine

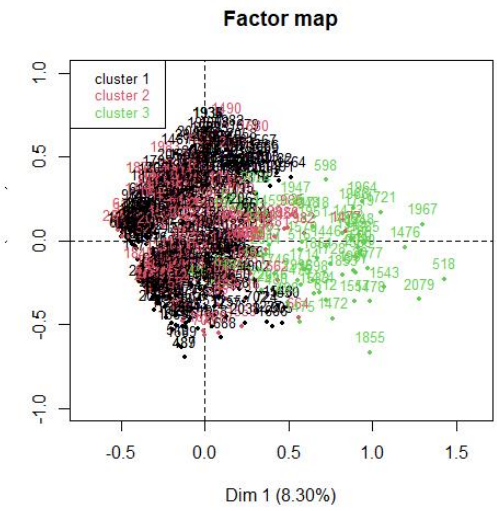


Figure 13 : Classification associée

Dans ces deux régions, n'y a pas de facteurs clairs qui influencent le choix d'un produit plutôt qu'un autre. Il est possible que ce choix provienne de facteurs externes comme les assurances, le choix du patient et des choix des médecins.

Discussion : finalement, sur les caractéristiques patients, nous n'observons pas d'influence de la région endémique.

Perspectives

Ce projet a permis de distinguer des caractéristiques des patients qui pourraient diriger l'intention de prescription des médecins. Cependant, sur le plan des caractéristiques des médecins, le questionnaire n'est pas encore assez complet pour pouvoir déterminer des groupes de médecins.

Conclusion

Notre objectif était de déterminer les caractéristiques influençant le traitement que recevrait un patient atteint de cancer des voies biliaires. Pour cela, nous avons d'abord essayé de comprendre les mécanismes de séparation des patients traités des patients placés en soins palliatifs. Il est montré que ce choix se fait sur l'état de santé général du patient, notamment pour savoir s'il est apte à supporter des traitements qui peuvent détériorer rapidement son état de santé.

Les méthodes d'ACM ont permis de réaliser une classification ascendante hiérarchique. Nous avons établi une ressemblance entre les groupes de patients entre le Japon et les Etats-Unis. Un premier groupe de patients, en très bonne santé générale, ne reçoit pas le produit X. Des patients avec un état de santé plus dégradé, que ce soit par des comorbidités ou des maladies rénales, sont la cible principale pour le produit. Enfin, un groupe de patients dont le profil d'approche des patients placés en soins palliatifs resteront sur leur traitement actuel.

Sur le niveau des caractéristiques médecin, nous n'observons pas de grandes différences entre les réponses des médecins des différents pays.

Finalement, le cancer des voies biliaires est aujourd'hui un cancer qui est rare. Les médecins n'ont pas beaucoup de cas précédents pour le choix du traitement.

Bibliographie

- Marcano-Bonilla L, Mohamed EA, Mounajjed T, Roberts LR. Biliary tract cancers: epidemiology, molecular pathogenesis and genetic risk associations. Chin Clin Oncol. 2016 Oct;5(5):61. doi: 10.21037/cco.2016.10.09. PMID: 27829275.

- Ghidini, Michele et al. “Biliary tract cancer: current challenges and future prospects.” *Cancer management and research* vol. 11 379-388. 28 Dec. 2018, doi:10.2147/CMAR.S157156

- Huifen Wang, Ping Sun, Katherine Baria; AstraZeneca, Gaithersburg, MD; AstraZeneca Pharmaceuticals, Gaithersburg, MD, Session: Poster Session B: Hepatobiliary Cancer, Neuroendocrine/Carcinoid, Pancreatic Cancer, and Small Bowel Cancer
  https://meetinglibrary.asco.org/record/182335/abstract

- https://www.cancer-environnement.fr/537-Cancer-de-la-vesicule-biliaire.ce.aspx

- Breiman, L. Random Forests. *Machine Learning* 45**,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324

l'institut Agro agriculture • alimentation • environnement	Diplôme : Ingénieur Spécialité : Sciences agronomes Spécialisation / option : Sciences des données Enseignant référent : David CAUSEUR

Auteur(s) : Junyi ZHAO Date de naissance* : 14/05/1998	Organisme d'accueil : KANTAR HEALTH Adresse : 3 avenue Pierre Masse, 75014 PARIS
Nb pages : 17 Annexe(s) : 0	
Année de soutenance : 2021	Maître de stage : Sabine RIZZO

Titre français : Etude des caractéristiques déterminant la prise de décision du traitement dans le cancer des voies biliaires.

Titre anglais: Study of the characteristics determining treatment decision making in biliary tract cancer.

Résumé (1600 caractères maximum) :

Le cancer des voies biliaires est un cancer rare et complexe. Il s'agit dans ce présent mémoire de déterminer les variables qui pourraient influencer la prise de décision pour le futur traitement, au niveau des caractéristiques patients. Cette étude étant menée sur plusieurs pays, il est intéressant de les comparer pour comprendre les déterminants dans chaque pays.

Ainsi, une méthode de prédiction avec les Forêts aléatoires a été mise en place, afin de déterminer les variables les plus importantes pour la décision de traiter ou de placer en soins palliatifs un patient.

Afin de déterminer des groupes de patientèle, une analyse exploratoire permettant la réduction de la dimension a été menée, suivie d'une classification ascendante hiérarchique.

Ces méthodes ont permis de dégager des groupes de patients, et donc une cible potentielle pour le produit présenté par le client. Un travail de communication reste pourtant à effectuer, concernant les différents traitements et leurs bénéfices, pour le cancer des voies biliaires.

Abstract (1600 caractères maximum):

Biliary tract carcinoma is a rare and complex cancer. The aim of this thesis is to determine the variables that could influence decision-making for future treatment, in terms of patient characteristics. As this study was conducted in several countries, it is interesting to compare them to understand the determinants in each country.

Thus, a prediction method with Randomized Forests was implemented to determine the most important variables for the decision to treat or place a patient in best supportive care.

To determine patient groups, an exploratory analysis with dimension reduction was conducted, followed by a hierarchical ascending classification.

These methods allowed us to identify groups of patients, and therefore a potential target for the product presented by the client. However, communication work remains to be done concerning the different treatments and their benefits for biliary cancer.

Mots-clés : Oncologie, Forêts aléatoires, Segmentation patient, Segmentation médecin, classification ascendante hiérarchique

Key Words: Oncology, Random Forest, Patient segmentation, Physician segmentation, hierarchical bottom-up classification

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*