Titre: Title:	Phénotypage du patient chirurgical et prédiction de trajectoire post-opératoire par apprentissage machine
Auteur: Author:	Pascal Laferrière-Langlois
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Laferrière-Langlois, P. (2022). Phénotypage du patient chirurgical et prédiction de trajectoire post-opératoire par apprentissage machine [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10511/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/10511/
Directeurs de recherche: Advisors:	Nadia Lahrichi, Maxime Cannesson, & Philippe Richebé
Programme: Program:	Maîtrise recherche en génie industriel

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Phénotypage du patient chirurgical et prédiction de trajectoire post-opératoire par apprentissage machine

PASCAL LAFERRIÈRE-LANGLOIS

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Août 2022

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Phénotypage du patient chirurgical et prédiction de trajectoire post-opératoire par apprentissage machine

présenté par **Pascal LAFERRIÈRE-LANGLOIS**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquée*

a été dûment accepté par le jury d'examen constitué de :

Louis-Martin ROUSSEAU, président

Nadia LAHRICHI, membre et directrice de recherche

Maxime CANNESSON, membre et codirecteur de recherche

Philippe RICHEBÉ, membre et codirecteur de recherche

Julie HUSSIN, membre

DÉDICACE

À la perle de ma vie, avec laquelle j'ai explorée le sable vénitien des plage californiennes et les montagnes enneigées de l'Ouest, sur lesquelles ont fleuries certaines idées ici-présentées.

REMERCIEMENTS

D’abord, merci à Nadia Lahrichi, ma directrice de thèse que j’ai dû solliciter trop souvent pour résoudre des complexités d’horaire liées à ma maîtrise. Merci pour ton expertise, ton dévouement à nous faire croître, et l’ambiance de collaboration que tu sais créer.

Merci à Dr Maxime Cannesson, qui m’a accueilli et intégré dans son équipe, et avec qui j’ai pu explorer la complexité entourant l’intégration de la recherche en données, la pratique médicale clinique et l’industrie, ainsi qu’à toute l’équipe de UCLA que j’ai eu la chance de connaître. Merci à Dr Philippe Richebé, qui est toujours présent pour m’aiguiller, me partager sa riche expérience, et supporter mes projets les plus fous. Merci à Olivier Verdonck, le chef de service clinique le plus dévoué que j’ai connu, toujours prêt à faire l’effort supplémentaire pour rendre service, tout en demeurant équitable et diplomate. Merci à toute l’équipe d’anesthésie et d’inhalothérapie de l’hôpital Maisonneuve-Rosemont qui est rapidement devenue une 2e famille, particulièrement à Dr. Louis Morisson avec qui je sais que je formerai un superbe binôme en clinique.

Merci à Marc-André Geraldo et Fergus Imrie pour la générosité de votre temps et le partage de votre expertise *Python*, de laquelle j’ai pu énormément apprendre jusqu’à atteindre le statut « programmeur débutant+ » dont je suis aujourd’hui très fier.

Enfin, merci à ma famille (oui, incluant toi, Nori), ma belle-famille et mes amis, pour votre support tout au long de mon intégration à Montréal puis à Los Angeles, pour toutes les discussions et partage d’idées et pour m’avoir changé les idées lorsque nécessaire.

RÉSUMÉ

Plus de 50 millions de chirurgies sont pratiquées annuellement en Amérique du Nord. Bien que seulement 12% des patients soient considérés à risque périopératoire élevé, ils expliquent 80% des complications graves causant un stress pour le patient et sa famille, ainsi que pour le système de santé œuvrant dans un contexte de ressources limitées. Inspiré de domaines connexes, le phénotypage digital est une approche identifiant des sous-groupes de patients partageant des caractéristiques clés au sein d'une population hétérogène et, en se basant sur le profil des données informatiques des patients, permet d'anticiper l'évolution clinique. L'hypothèse à la base de cette étude est que le dossier médical électronique permet de phénotyper les patients en période préopératoire pour stratifier leur risque. Les objectifs sont 1) identifier pour la première fois des phénotypes avec les données médicales préchirurgicales; 2) comparer la performance prédictive des phénotypes au score ASA (*American Society of Anesthesiologists*) actuellement utilisé; 3) ainsi qu'à des modèles prédictifs supervisés entraînés sur les mêmes données.

À partir de la base de données PDW (*Patient Data Warehouse*) provenant du dossier médical électronique du réseau hospitaliser de UCLA (*University of California in Los Angeles*), tous les patients adultes ayant subi une laminectomie, colectomie ou chirurgie thoracique depuis 2013, instauration du PDW, ont été extraits. Cinq issues cliniques ont été utilisées pour décrire la trajectoire postopératoire : mortalité hospitalière, mortalité à 30 jours, réopération à 30 jours, admission en soins intensifs (USI) et durée d'hospitalisation postopératoire prolongée (DHP). Considérant la présence de 4,000 variables par chirurgie dans le PDW, une première sélection de variables préopératoires a été effectuée par le consensus de trois experts cliniques et, à partir de celle-ci, un jeu de données par chirurgie fut créé en retenant uniquement les variables présentant une corrélation de Pearson statistiquement significative (p <0.05) avec l'une des issues cliniques. Après le retrait des variables fortement corrélées, les trois jeux de données comportaient respectivement 34, 36 et 33 variables, qui furent normalisées par transformation standard. Chaque jeu de donnée était ensuite séparé en ensemble de dérivation et test, basé sur l'année de chirurgie, afin de faire la validation temporelle de la performance prédictive pour chaque issue clinique.

La segmentation par quatre algorithmes (DB Scan, hiérarchique*, k-means*, *consensus k-means*) a été explorée avant de retenir le *consensus k-means* pour créer trois phénotypes. Une fois le jeu de dérivation segmenté, une forêt aléatoire a été entraînée pour attribuer prospectivement un

phénotype aux patients du jeu de données test et analyser les résultats. Pour chacune des chirurgies, le phénotype 0 était le plus fréquent (total de 73.6%) et regroupait les patients typiquement plus jeunes, avec moins de comorbidités et subissant une chirurgie non-urgente, dite élective. Le phénotype 1 regroupait des patients plus âgés et plus malades subissant typiquement une chirurgie élective, alors que le phénotype 2 était principalement caractérisé par une chirurgie urgente, une hospitalisation préopératoire plus longue et une douleur préopératoire plus grande. Les 5 issues cliniques mesurées présentaient une progression croissante à travers les phénotypes (mortalité hospitalière : 0.2%, 2.3% et 7.3%; réopération : 2.8%, 5.4% et 9.3%; admission en USI : 8%, 36.1% et 48%). Lorsque la performance prédictive des phénotypes était mesurée à l'ASA, l'aire sous la courbe ROC (*Receiver Operating Characteristics*) du phénotype digital était similaire ou légèrement supérieure (mortalité hospitalière : 0.85 et 0.84; réopération : 0.62 et 0.59; admission USI : 0.76 et 71). Les profils de courbe ROC du phénotype digital et de l'ASA suggéraient leur complémentarité, et la combinaison des deux performait de façon supérieure (0.91, 0.63 et 0.80).

Trois architectures de modèles supervisés ont été explorées : régression logistique (RL), forêt aléatoire (RF) et perceptron multicouche (MLP). À partir des mêmes jeux de données que ceux utilisés pour le phénotypage, les architectures ont été explorées avec le jeu de dérivation et une validation croisée avec 5 replis. Le modèle le plus performant pour chaque chirurgie et pour chaque issue clinique était ensuite appliqué au jeu de test. L'architecture la plus performante (11 des 15 modèles) était le MLP. Alors que l'AUROC et le score F1 ont respectivement atteint des valeurs moyennes élevées dans les 5 replis de la validation (mortalité hospitalière : 0.91 et 0.4; réopération : 0.64 et 0.22; admission USI : 0.99 et 0.95), la performance dans le groupe test était faible. La meilleure valeur d'AUROC en test est 0.74 et F1, est 0.25. Cette différence de performance est au moins partiellement expliquée par le débalancement important des classes puisque les issues cliniques utilisées sont rares. Des stratégies de balancement des classes auraient pu être explorées, mais la même transformation aurait dû être appliquée à la base de données de phénotypage pour répondre à l'objectif de comparer le signal extrait par les deux approches.

En résumé, cette recherche présente les premiers phénotypes préchirurgicaux et démontre leur capacité prédictive sur la trajectoire de soins postopératoires pour trois différentes chirurgies fréquemment pratiquées en Amérique du Nord. Si la méthode est confirmée sur davantage de chirurgies, les phénotypes ont le potentiel d'automatiser l'analyse de risque, contrairement aux scores actuellement utilisés qui dépendent d'une évaluation par le clinicien.

ABSTRACT

More than 50 million surgeries are annually performed in North America. Even if a mere 12% of the patients are considered high-risk, they account for 80% of the significant complications that will cause distress both for the patient and his family, and for the healthcare systems operating in an actual context of resource scarcity. Inspired from related fields, digital phenotyping allows the identification of clusters sharing key characteristics among a wider heterogeneous population and based on the medical data of the individual patients, can anticipate the care trajectory. The underlying hypothesis of this research electronic medical record (EMR) can be used to phenotype patients before their surgery to stratify their risk. The objectives are 1) to create the first digital phenotypes of surgical patients using the preoperative data, 2) to compare the predictive ability for the care of these phenotypes to the ASA (American Society of Anesthesiologists) score currently used in clinical environment as we as 3) with predictive supervised model trained on the same datasets.

Using the Patient Data Warehouse (PDW), a custom database created from the EMR used in the hospital network of the University of California in Los Angeles (UCLA), all the adult patients undergoing a laminectomy, colectomy of thoracic surgery since 2013, the inception of 2013, were extracted for analysis. Five clinical outcomes were used to describe the care trajectory: hospital mortality, 30-days mortality, 30-days reoperation, admission at the intensive care unit (ICU), and prolonged postoperative hospitalization. With more than 4000 different features recorded for each surgery in the PDW, a first selection of preoperative variables was made by consensus among 3 medical experts, and, from this selection, three surgery-specific datasets were created by keeping only the variables significantly correlated (p <0.05) with at least one clinical outcome. After removing highly correlated variables the three datasets respectively contained 34, 36 and 33 features, which were rescaled to a standardized distribution. Each dataset were then separated in. derivation set and test set based on the year of the surgery, allowing prospective temporal validation of the predictive models created for each dataset.

The segmentation results of 4 different algorithms (DB Scan, hierarchical, *k-means*, *consensus k-means*) were analyzed before retaining *consensus k-means* to create 3 distinct surgery-specific phenotypes. Once the final segmentation completed on the derivation dataset, a random forest algorithm was trained to prospectively attribute a phenotype to the patients in the test set and

analyse the results. For all surgeries, phenotype 0 was the most frequently attributed (63.6%) and mostly contained younger and healthier patient undergoing an elective surgery. Phenotype 1 typically contained older patient with more comorbidities undergoing elective procedure, while phenotype 2 was characterized by the urgency of the surgery, longer preoperative hospitalisation and increased preoperative pain. The 5 clinical issues increased progressively among the phenotypes (hospital mortality: 0.2%, 2.3% and 7.3%; reoperation: 2.8%, 5.4% and 9.3%; ICU admission: 8%, 36.1% and 48%). When compared to the ASA score, the area under the receiver operating characteristics curve (AUROC) of the digital phenotype was similar or slightly superior to ASA (hospital mortality: 0.85 and 0.84; reoperation: 0.62 and 0.59; ICU admission: 0.76 and 71). The profiles of the 2 ROC curves suggested complimentary in the signal extracted, and the linear combination of the 2 scores achieved higher performance than each alone (0.91, 0.63 and 0.80).

Three architectures of supervised models were explored: logistic regression (LR), random forest (RF) and multilayer perceptron (MLP). Starting from the same datasets as used for phenotyping, the architectures were explored in the derivation set with a 5-fold cross validation. The most performing model for each surgery and each clinical outcome was used on the test set. MLP was provided the best performance in 11 of the 15 models. While AUROC and F1 score both reached high average performance in the 5-fold validations (hospital mortality: 0.91 and 0.4; reoperation: 0.64 and 0.22; ICU admission: 0.99 and 0.95), the prediction in the test set remained low. The highest AUROC and F1 score obtained in the test set were respectively 0.74 and 0.25. This performance discordance can be partially explained by the significant class imbalance considering that all clinical outcomes explored are rare. Strategies as over- and under-sampling could have been explored to improve the metrics, but the same transformation would have to be made on the datasets before phenotyping the patients to respect our objective of comparing the signal extracted by the two different methods.

In conclusion, this research presents the first presurgical phenotypes and demonstrated their predictive ability on the postoperative care trajectory of three different frequently performed surgeries. When compared to ASA score which requires expert assessment, these phenotypes strictly use readily extractable features from the EMR and therefore present the potential of automating risk stratification.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ASA	*American Society of Anesthesiology*
AUROC	*Area Under Receiver Operating Characteristics curve*
CRISP-DM	Cross Industry Standard Process for Data Mining (CRISP-DM)
COVID-19	*Coronavirus Disease of 2019*
DDR	*Discovery Data Repository*
DME	Dossier medical électronique
ICD9	*International Classification of Disease, 9th edition*
ICD10	*International Classification of Disease, 10th edition*
KNN	*K-nearest neighbors*
MLP	Percepton multicouche
NMI	Score d'information mutuelle normalisée
OMS	Organisation mondiale de la santé
PCA	Analyse en composant principale
PDW	*Patient Data Warehouse*
POSPOM	Score préopératoire de mortalité postopératoire
RL	Régression logistique
SVM	Machine à vecteur de support
t-DSME	*t-Distributed Stochastic Neighbor Embedding*
UCLA	Université de Californie à Los Angeles
USI	Unité de soins intensifs

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

La chirurgie est une période critique dans l'épisode de soins d'un patient [1]. Bien que la majorité se déroule sans particularité, une complication peut sévèrement altérer la qualité de vie du patient, voire directement menacer la vie. La mortalité postopératoire, définie par un décès survenant dans les 30 jours suivant une chirurgie, explique 7.7% des décès à l'échelle de la planète, soit 4.2 millions annuellement [2]. Bien que ce nombre comprenne les traumas et les chirurgies tentées en dernier recours, une récente publication classifie la mortalité post-opératoire comme 3^{e} cause de mortalité globale, et la mortalité à 6.1 millions si toutes chirurgies requises étaient pratiquées [2]. Bien que la mortalité postopératoire globale demeure sous les 2%, plus de 80% de celle-ci est attribuable au 12% des patients jugés à risque chirurgical élevé [3]. Les complications et le prolongement non planifié d'un séjour hospitalier induisent un stress significatif pour le patient, pour sa famille, ainsi que pour le système de santé. Ce dernier point est significatif au Québec et au Canada, considérant le nombre de lits et les ressources humaines limitées, cela s'ajoutant au coût croissant des traitements de santé. Avec la démocratisation du dossier médical électronique (DME) aux États-Unis et certaines provinces canadiennes, ainsi que la croissance de la puissance de calcul des systèmes informatiques, les futurs patients peuvent être triés en analysant le comportement et la trajectoire de soin de patients similaires ayant déjà été traités [1, 4-7]. Si appliquée de manière systématique et rigoureuse, une telle stratification de risque peut 1) servir d'aide cognitive à la prise de décision thérapeutique; 2) faciliter la collaboration décisionnelle avec le patient; 3) améliorer l'allocation des ressources; 4) servir de métrique pour analyser les soins de santé prodigués dans différents milieux; et 5) aider les compagnies d'assurance et de facturation pour équilibrer les primes [1, 8].

Plusieurs publications récentes ont tenté de stratifier le risque chirurgical des patients, et la majorité ciblait certaines issues cliniques postopératoires telles que la mortalité, l'insuffisance rénale aiguë ou le besoin de réintubation. Le score POSPOM - *Preoperative Score to Predict Post-Operative Mortality* – utilise l'âge, le type de chirurgie, ainsi que 26 comorbidités sélectionnées par des experts du domaine [9]. L'index Charlson de comorbidité, initialement créé pour prédire la mortalité à 10 ans et les coûts associés, a récemment été transposé au contexte périopératoire [10]. Néanmoins, aucun outil de stratification n'est parvenu à remplacer le score ASA – *American*

Society of Anesthesiologists – développé en 1941 [1]. Ce score demeure l'outil de stratification de risque de prédilection autant auprès des hôpitaux que des compagnies d'assurances. Celui-ci pose toutefois plusieurs limitations, incluant le besoin d'une expertise clinique pour être attribué à un patient, en plus d'être indépendant du type de chirurgie que subira le patient.

Le phénotypage digital est une méthode nouvellement appliquée en contexte clinique afin d'identifier des sous-groupes partageant des caractéristiques communes au sein d'une population hétérogène [11-13]. Sans influence humaine, les algorithmes d'apprentissage machine non-supervisés peuvent segmenter la population en sous-groupes, en identifiant les caractéristiques discriminantes de la population. Un expert peut ensuite extraire les caractéristiques clés associées à chaque sous-groupe [14] pour mieux comprendre le profil type de chaque phénotype digital. Tout comme la couleur des yeux définit le phénotype physique d'un être humain, les caractéristiques clés de chaque sous-groupe définissent le phénotype digital, ou profil type. À partir de ce phénotype, il devient possible d'anticiper certains comportements telles les complications postopératoires ou la réponse à des traitements prodigués. Par exemple, une récente publication présente l'analyse de 20,189 patients septiques aux soins intensifs, au sein de laquelle 4 phénotypes ont été identifiés pour leur différent profil de biomarqueurs et leur type d'atteinte d'organes à l'admission[15]. Ces phénotypes étaient associés à une mortalité variant de 5% à 40% et avaient une réponse variable au traitement de remplissage volémique. Similairement, trois phénotypes ont été identifiés à la présentation initiale de patients atteints de COVID-19 (*Coronavirus Disease of 2019*) et étaient associés à une admission en soins intensifs et/ou une mortalité respectivement de 8, 18 et 43%. L'attribution du phénotype se basait principalement sur l'âge, les comorbidités du patient et les symptômes de présentation. Deux autres exemples notables sont 1) la dépression majeure, avec une présentation très hétérogène et un objectif bien établi de prévenir l'autodangerosité; et 2) les cardiomyopathies, ou pathologies du cœur, qui ont un traitement désormais radicalement différent basé sur la nature obstructive, restrictive ou ischémique. Jusqu'à maintenant, aucune tentative n'a été faite pour phénotyper la population hétérogène des patients allant subir une chirurgie.

L'hypothèse à la base de cette recherche est que nous pouvons créer des phénotypes digitaux préchirurgicaux permettant, en se basant sur le phénotype prospectivement attribué aux patients, d'anticiper leur trajectoire de soins. Pour tester cette hypothèse, l'ensemble des patients ayant subi une laminectomie (exérèse de la lame d'une vertèbre), une colectomie (exérèse d'une portion de

colon) ou une chirurgie thoracique ont été extraits de la base de données périopératoire de UCLA – *University of California in Los Angeles* – depuis l'instauration du dossier médical électronique en 2013. Ces chirurgies ont été sélectionnées puisqu'elles sont parmi les plus souvent pratiquées aux États-Unis et qu'elles présentent une gamme de risque allant de faible à modérément élevé.

Les objectifs de ce projet sont donc de :

1) créer des phénotypes digitaux préopératoires pour les patients subissant une chirurgie;

2) de comparer la performance prédictive de ces phénotypes au score de l'ASA par rapport à la trajectoire de soins postopératoires, telle que définie par les issues cliniques suivantes : la mortalité hospitalière, la mortalité à 30 jours, la réopération, l'admission postopératoire en soins intensif; et la durée d'hospitalisation prolongée;

3) de comparer la performance prédictive des phénotypes digitaux à des modèles d'apprentissage supervisés cherchant à prédire les mêmes issues postopératoires.

Ce mémoire débutera au chapitre 2 par une revue de littérature exposant les connaissances actuelles sur le phénotypage en médecine, ainsi que sur l'utilisation d'algorithmes d'apprentissage machine en périopératoire. La chapitre 3 présentera l'approche méthodologique standard pour un projet d'exploration de données, alors que le chapitre 4 détaillera l'application de cette méthodologie dans le contexte du projet actuel pour l'entraînement de modèles d'apprentissage machine. Le chapitre 5 présentera les résultats des expériences menées en lien avec les objectifs, ainsi que l'analyse de ces résultats. Finalement, le chapitre 6 fera la synthèse de cette thèse et proposera des avenues de recherches futures.

CHAPITRE 2 REVUE DE LITTÉRATURE

Cette revue de littérature présente l'état actuel des connaissances liées à cette recherche. Les outils de stratification opératoire actuellement utilisés ou récemment développés seront présentés et l'emphase sera mise sur l'application actuelle du phénotypage digital en santé. Les variables influençant la trajectoire de soins, ainsi que les stratégies d'extraction de ces informations, seront présentées à partir des modèles actuellement publiés. Finalement, la dernière section traitera spécifiquement de la gestion des données et des types d'algorithmes utilisés dans le cadre de cette recherche.

2.1 Définir la trajectoire de soins postopératoires

La trajectoire de soins peut être décrite au moyen de différentes issues cliniques mesurables, dont les plus fréquentes sont la mortalité intrahospitalière, à 30 jours, ou à 90 jours, la durée de séjour en hôpital, la durée de séjour en soins intensifs et le pourcentage d'incidence de complications significatives tels les évènements cardiaques majeurs. Ces issues cliniques sont dites « centrées sur le patient », puisqu'elles décrivent directement l'impact sur la vie ou la qualité de vie [16]. Elles aident ainsi le patient dans sa prise de décision avant l'opération, tout comme elle aide l'hôpital dans sa gestion des ressources[17]. C'est pour cette raison que beaucoup d'efforts sont mis pour adéquatement prédire la trajectoire de soins anticipées. Pour ce faire, différentes stratégies existent, et le phénotypage présenté dans cette recherche jette un nouveau regard sur les outils existants.

2.2 Outils de stratification de risques préopératoires

Bien que la mortalité postopératoire soit la troisième cause de mortalité à l'échelle de la planète, les deux dernières décennies ont vu une diminution drastique de ce nombre de 7.8 fois entre 1990 et 2019, passant de 100.85 à 12.98 par 100,000 chirurgies [2, 18]. Les explications sont multiples et la meilleure sélection de patients, issue de l'amélioration de la stratification préopératoire, est l'une d'entre elles.

Une revue systématique publiée en 2021 a permis d'identifier 21 différents scores de risques prédisant la mortalité [19]. Ces scores se divisent en deux grandes familles : 1) les scores classiques, qui calcule un score à partir de certains paramètres et qui, selon ce score, associe un pourcentage

de risque, par exemple la mortalité ou l'infarctus cardiaque, et 2) les algorithmes prédictifs et/ou de classification exploitant l'apprentissage machine, les bases de données périopératoires et analysant habituellement un nombre plus élevé de variables.

2.2.1 Scores classiques

Score ASA

Développé en 1941 et encore utilisé aujourd'hui à travers toute l'Amérique du Nord et l'Europe, le score ASA se distingue par son omniprésence. Étant validé comme facteur de risque indépendant de mortalité, il est utilisé à la fois par les gouvernements pour guider la distribution des budgets à travers les hôpitaux, et par les compagnies d'assurance pour déterminer les primes demandées aux patients [1]. Le Tableau 2.1 décrit les scores et l'incidence de complications associées ainsi que l'aire sous la courbe *Receiver Operating Characteristic* (AUROC) qui est quantifiée à 0.81 pour la mortalité [19]. Néanmoins, plusieurs limitations y sont associées : 1) l'expertise d'un anesthésiste est requise pour l'attribution du score, et 2) il ne prend pas en compte le degré de risque de la chirurgie, pourtant considéré par d'autres scores comme le facteur de risque le plus influent [9].

Tableau 2.1 Classification du score ASA, inspiré de Horvath et al. [1] et Hackett et al. [20]

Score ASA	**Définition**	**Complications**	**Décès**
ASA 1	Patient normal et en santé	2%	0.02%
ASA 2	Patient ayant une pathologie systémique mineure	5%	0.14%
ASA 3	Patient ayant une pathologie systémique majeure sans notion d'incapacité	14%	1.41%
ASA 4	Patient ayant une pathologie systémique majeure provoquant une menace constante à la vie	37%	11.14%
ASA 5	Patient moribond avec risque mortalité imminente en absence de chirurgie	71%	50.87%

Score POSPOM

Le score POSPOM fut publié et développé en 2016 à partir de 5,507,834 chirurgies, avant d'être validé sur la moitié de ce nombre [9]. À partir de 29 variables préopératoires, il prédit la mortalité

de l'ensemble de validation avec un AUROC de 0.929 dans la publication initiale, et entre 0.78 et 0.86 dans des publications subséquentes [21, 22]. Brièvement, des points sont attribués selon l'âge du patient (maximum 16 points), selon les comorbidités (de 1 à 4 points par comorbidités) et selon le type de chirurgie (maximum 22 points). La somme de ces points résulte en un score de 0 à 51+, respectivement associé à <0.001% et >97.865% de mortalité. La stratégie d'extraction des comorbidités est intéressante pour ce mémoire puisqu'elle permet de binariser les codes ICD10 (*International Classification of Disease*-10) créés par l'Organisation mondiale de la santé (OMS) et retrouvés dans les dossiers des patients [23]. À partir d'un groupe d'experts, une liste de 26 comorbidités fut établie et une liste de codes ICD10 fut associée à chacune d'elle. Si un de ces scores ICD10 est présent dans le dossier du patient avant la date de la chirurgie, il est considéré comme ayant la comorbidité. Le Tableau 2.2 présente un exemple, avec la maladie ischémique cardiaque et les 3 codes ICD10 associés.

Tableau 2.2 Exemple de classification des comorbidités du score POSPOM pour la maladie ischémique cardiaque, adaptation du matériel supplémentaire de Le Manach et al. [9].

Codes ICD10	**Pathologie associée**
I20.x	Angine de poitrine
I25.x	Maladie cardiaque ischémique chronique
Z95.5	Présence d'un implant et/ou greffe par angioplastie des coronaires

Autres scores

Au total, trois autres outils prédictifs de mortalité ont à la fois obtenu une performance prédictive de mortalité avec AUROC supérieure à 0.90 et ont été validés dans des publications subséquentes. Les scores SMPM [24], SORT (*Surgical Outcome Risk Tool*) [25] ainsi que le NZRISK [26] issu du score SORT et adapté à la population de Nouvelle-Zélande, sont similaires puisqu'ils utilisent tous le score ASA auquel ils ajoutent un risque opératoire lié à la chirurgie et un degré d'urgence, mais ils diffèrent entre eux par l'utilisation de variables comme l'âge et les comorbidités.

De manière intéressante, tous ces scores s'intéressent uniquement à la mortalité, alors que la trajectoire de soins inclut également les notions de complications non mortelles, perte de fonctions et/ou séjour prolongé en hôpital. D'autres scores spécifiques existent pour ces issues cliniques, tel

le score PROMIS (*Patient-Reported Outcome Measurement Information System*) s'intéressant à la récupération fonctionnelle nécessaire au retour à la maison [27].

2.2.2 Scores développés par apprentissage machine

Avec l'amélioration de la puissance de calcul informatique et la croissance des bases de données provenant de la démocratisation du DME, l'apprentissage machine est devenu un sujet florissant de la littérature médicale. En 2010, seulement 713 citations comprenant le terme *Machine Learning* ont été publiées dans PubMed, contre plus de 25 000 en 2021.

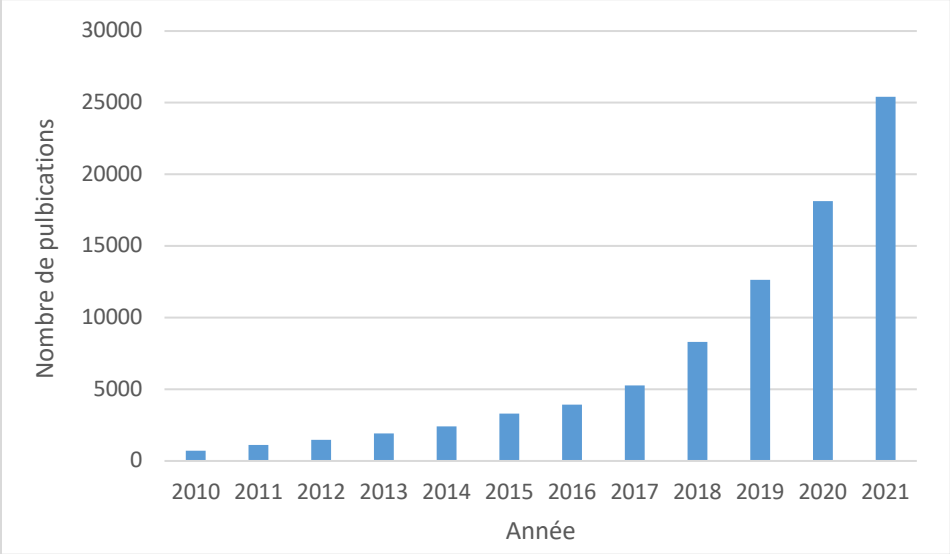


Figure 2.1 Nombre de publications comprenant le terme "Machine Learning" dans PubMed

De nouvelles publications ont ainsi exploré des modèles prédictifs plus complexes ne dépendant plus d'un score final auquel s'associe linéairement une incidence de complications. La plupart des publications explorent des modèles supervisés prédisant une complication postopératoire unique [4, 6, 28]. Les modèles fréquemment explorés incluent la régression logistique (RL), la forêt aléatoire, le perceptron multicouche (MLP), le *boosting* de gradient et la machine à vecteurs de support (SVM), qui offrent généralement des niveaux de performance rapprochées [4, 28-30]. Plus récemment, les publications se sont complexifiées en prédisant des complications multiples et moins explorées avec les scores de risque classiques [30, 31].

Les modèles d'apprentissage supervisé prédisant la mortalité rapportent des valeurs AUROC optimales dans l'intervalle 0.90-0.94 et des scores F1 aux environs de 0.25-0.3 [4, 6, 31]. De façon surprenante, les performances des scores ASA et POSPOM sont significativement inférieures à celles rapportées dans les publications discutées précédemment, descendant parfois jusqu'à une AUROC de 0.74 [6]. Cette observation pourrait s'expliquer par un biais de publication favorisant les articles d'apprentissage machine qui batte les scores classiques. Les autres complications postopératoires explorées, avec leur AUROC, incluent la pneumonie (AUROC : 0.905), l'insuffisance rénale aigue (=0.848), la thrombophlébite (0.881), l'embolie pulmonaire (0.831) et le délirium (0.762) [30]. Ces métriques de performance proviennent de l'ensemble d'une validation croisée, sans être ensuite testée dans un groupe test de patient duquel les modèles étaient complètement aveugles.

De façon intéressante, un groupe d'auteurs a souhaité prédire le score ASA plutôt qu'une complication spécifique [32, 33]. À partir de l'âge, des signes vitaux et du glucose sanguin, soit tous des paramètres non utilisés par l'anesthésiste pour attribuer un score ASA, des modèles SVM et MLP ont pu atteindre respectivement 93.25% et 91.43% de précision. Cette stratégie s'approche du phénotypage proposé dans ce mémoire puisqu'elle classifie le patient dans un groupe qui est lui-même associé à des trajectoires de soins postopératoires variables. Néanmoins, la démarche est significativement différente puisque l'existence préalable du score ASA permet l'utilisation d'apprentissage supervisé, alors que le phénotypage recherche des sous-groupes par apprentissage non supervisé.

En conclusion, peu de outils de stratification récemment développés se distinguent et le score ASA, malgré sa simplicité, continue d'offrir des performances prédictives suffisamment grande pour être utilisé en clinique. Néanmoins, la littérature se développe et les modèles se complexifient. La majorité des scores classiques n'adressaient que la mortalité, alors que les nouveaux modèles prédisent des complications spécifiques. Il est intéressant de constater que le phénotypage proposé dans cette recherche s'approche davantage du score ASA puisqu'il regroupe les patients basés sur des caractéristiques clés, sans chercher à spécifiquement prédire des issues post-opératoires au moment où le phénotype est attribué.

2.3 Application actuelle du phénotypage en médecine

Le phénotype digital, même si encore peu appliqué au contexte périopératoire ou aux soins aïgus, demeure une approche décrite dans le domaine de la santé, notamment en psychiatrie. Il s'agit d'un concept se démocratisant dans le monde actuel où la trace digitale d'un individu est enregistrée à partir de plusieurs sources, tels un téléphone cellulaire intelligent, les réseaux sociaux, les signes vitaux d'une montre d'analyse sportive, ou un DME [11]. Tout comme nous apprécions la taille et la couleur des yeux définissant le phénotype d'un individu, il est possible d'analyser les données informatiques liées à un patient pour explorer le temps passé sur chaque application d'un téléphone, les pages visitées sur un média social, et la modulation des signes vitaux liées à une activité physique [34].

L'ensemble de ces données et de ces signatures digitales définissent le phénotype digital d'un individu, qui peut servir à catégoriser une population en différent sous-groupes partageant des caractéristiques clés.

2.3.1 Segmentation en sous-groupes

Avec la quantité et la complexité grandissantes des données accumulées sur les patients, il devient intéressant d'utiliser des algorithmes de segmentation pour identifier les patrons de caractéristiques au sein d'une population hétérogène, et assigner les patients à ces patrons que l'on nommera « phénotype 1,2,3 » ou « *alpha, bêta, gamma* ». Cette approche demeure théoriquement similaire à celle utilisée depuis l'aube de la médecine pour diagnostiquer telle ou telle maladie basée sur tels symptômes (le phénotype de symptômes), mais les algorithmes de segmentation permettent d'analyser une plus grande quantité d'informations sur plus de dimensions, permettant ainsi une délimitation des sous-groupes à des granularités plus profondes [12].

Cette approche a récemment été appliquée pour phénotyper des patients hospitalisés en soins intensifs dans le cadre de sepsis, une infection systémique sévère [15]. En utilisant un algorithme de *consensus k-means* à une population de 20,189 patients septiques, quatre phénotypes, nommés *alpha*, *bêta*, *gamma* et *delta*, ont été identifiés à partir de 29 variables incluant les données démographiques, les biomarqueurs sanguins et les dysfonctions d'organes. Le profil type de chaque phénotype était analysé par un expert du domaine, permettant d'associer certaines caractéristiques à un phénotype, par exemple la dysfonction hépatique au phénotype delta. Les auteurs ont ensuite validé leur phénotype dans trois études randomisées contrôlées étudiant le sepsis. À travers

l'ensemble de dérivation (entraînement) et les trois ensembles de tests, le phénotype delta se démarquait par une mortalité plus élevée (32%), alors que le phénotype alpha, plus fréquent, était associé à 2% de mortalité.

Une segmentation similaire a récemment été appliquée dans le contexte de patient atteint de -COVID-19 [13]. Trois phénotypes, nommés 1,2 et 3 ont été identifiés par le même algorithme de *consensus k-means.* Le profil type de chacun était établi par des analyses descriptives. Par exemple, le phénotype 1 était typiquement une jeune femme (74.9%) avec peu de comorbidités et présentant de la fièvre (56%) et douleur musculaire (82%), alors que le phénotype 3 était des hommes (70.7%) plus âgés (moyenne 73 ans) avec une moyenne de 2.2 comorbidités par patient et présentant surtout des symptômes respiratoires (45%). La mortalité de chaque phénotype était respectivement 3, 9 et 22% de mortalité. Bien que cette analyse ne présente aucun groupe de validation, cette étude présente l'un des intérêts de la segmentation en phénotypes comparativement à des modèles supervisés : le phénotype attribué est corrélé à plus d'une issue clinique, soit l'admission en soins intensifs et la mortalité. Cette approche est celle utilisée dans le cadre de ce projet.

2.3.2 Phénotypage et prédiction d'évènement

Dans le contexte médical, la combinaison du DME aux technologies portatives permet de suivre l'évolution clinique d'un patient en temps réel, voire de solliciter des actions par des notifications ou questions envoyées sur le téléphone cellulaire [35]. En psychiatrie, l'analyse par technologie portative d'une semaine de mouvement corrélait significativement avec la sévérité des symptômes évalués par un clinicien ($r = 0.855$, $p = 0.017$) [36] et les profils de pensées suicidaires ont été subdivisés en 5 catégories pour quantifier le risque de tentative de suicide, permettant ainsi au clinicien d'adapter sa prise en charge [37]. Similairement, la récidive de psychose en contexte de schizophrénie était associée à une hausse de 71% des anomalies comportementales dans les deux semaines précédant le diagnostic [38]. La majorité de la littérature psychiatrique utilise des modèles supervisés (RF, SVM, MLP) pour prédire l'occurrence d'une tentative de suicide, ou la récurrence d'une psychose [39].

En contexte périopératoire et de soins aïgus, l'approche diffère puisque la majorité des données sont disponibles directement dans le DME, et le monitoring continu génère le même type de données temporelles que les technologies portatives. En exemple, l'analyse de séries temporelles

de haute fréquence (100 Hz) de la tension artérielle permet d'anticiper l'apparition de basse tension (AUROC 0.95), elle-même liée à des complications postopératoires [40].

2.4 Variables explicatives de la trajectoire postopératoire

Une revue exhaustive de la littérature permet d'identifier plusieurs variables influençant l'évolution postopératoire d'un patient. Dans le cadre de ce projet, certaines ne seront pas utilisées, soit par non-disponibilité des données, soit par manque de généralisabilité, soit par décision éthique [41].

2.4.1 Facteurs liés à l'état de santé

L'état de santé d'un patient est indubitablement lié à sa trajectoire postopératoire, et la complexité réside dans le choix des variables décrivant cet état de santé. Dans le cas du score ASA, l'état de santé est directement évalué par la médecine sur une échelle de 1 à 5 [20]. Pour s'émanciper de l'analyse humaine, il importe d'utiliser les variables sous-jacentes à l'évaluation médicale, incluant sans toutefois s'y limiter, des données démographiques (âge, sexe, taille, poids), comorbidités connues du patient ainsi qu'à un système de catégorisation de ces comorbidités, signes vitaux du patient, valeurs de laboratoire, médicaments pris chroniquement ou ponctuellement, etc [42]. Les modèles d'apprentissage supervisé prédisant la mortalité utilisent tous une combinaison de ces variables [4, 6, 28, 30, 31]. Certains modèles utilisent des données temporelles pour analyser l'évolution de certaines variables, par exemple les données liées aux signes vitaux [40].

Étant donné l'interdépendance des variables de l'état de santé, des corrélations significatives peuvent exister au sein de la première base de données. Peu d'auteurs détaillent rigoureusement leur stratégie de sélection, mais plusieurs auteurs rapportent appliquer une telle stratégie [4, 6, 29-31].

2.4.2 Facteurs liés au soins médicaux

Au-delà de l'état de santé du patient, le type de soins prodigués affecte aussi la trajectoire de soins. Les scores SMPM [24], SORT (*Surgical Outcome Risk Tool*) [25] et NZRISK [26] ajoutent au score ASA la notion d'urgence chirurgicale, parfois binaire, parfois catégorique en subdivisant le degré d'urgence, ainsi qu'un risque attributaire à la chirurgie. Certaines publications vont jusqu'à attribuer au risque chirurgical la plus grande influence, dépassant l'état de santé du patient [9]. Étant donné la variabilité entourant les procédures, trois stratégies principales sont utilisées pour

quantifier ce risque. En accord avec les lignes directrices d'évaluation préopératoire, certains scores divisent la dangerosité chirurgicale en risque faible, intermédiaire et élevé [24, 42], alors que d'autres, tel le score POSPOM [9], attribuent un risque spécifique à chaque catégorie chirurgicale (gastro-intestinale, vasculaire, etc.). Également, certains auteurs exploitent la classification HCUP (*Healthcare Cost and Utilization Project*) qui, un peu comme la classification ICD10 des pathologies, permet de catégoriser les types de chirurgie [7]. Indépendamment de la stratégie, l'ajout d'un risque chirurgical améliore la performance des modèles.

Une autre question entoure l'inclusion ou non dans le modèle des variables intraopératoires, tels les signes vitaux monitorés durant la chirurgie, *versus* se limiter uniquement aux données préopératoires. Ce choix dépend évidemment de l'objectif de recherche, mais certains auteurs ont comparé la performance de modèles avec et sans données intraopératoires [30]. Pour prédire l'occurrence de pneumonie, insuffisance rénale aïgue et embolie pulmonaire, l'ajout des données intraopératoires aux données préopératoires améliorait les performances de 0.019, 0.032 et 0.009 respectivement pour les trois chirurgies. À l'inverse, certains modèles atteignent des valeurs d'AUROC supérieures à 0.90 en utilisant principalement des données intraopératoires [6, 7, 31]. Bref, le bénéfice d'ajouter les variables intraopératoires semble avoir un effet variable sur la performance.

2.4.3 Autres facteurs

Plusieurs autres facteurs ont démontré influencer la trajectoire de soins et les complications postopératoires. La performance individuelle du médecin chirurgien bien sûr en fait partie, mais le fait d'inclure cette variable dans un modèle pose à la fois un problème de généralisabilité à d'autres hôpitaux, ainsi qu'un enjeu éthique par rapport à l'utilisation qui en sera faite [43]. À un niveau moins granulaire, les patients des femmes chirurgiennes ont statistiquement moins de complications que ceux des hommes chirurgiens [44]. Quoique réfuté par certains auteurs, le jour de la semaine pourrait également influencer l'incidence de complications [45, 46]. Dans un autre ordre d'idées, un statut socioéconomique faible a été associé à des issues cliniques défavorables, incluant une plus forte prévalence de complications postopératoires, de mortalité à 30 jours, et une plus longue durée d'hospitalisation postopératoire [47].

Malgré leur influence, les variables présentées dans cette sous-section ne semblent pas utilisées par les modèles d'apprentissage machine prédisant les issues postopératoires. Il en résulterait plusieurs

questions éthiques ou enjeux d'accessibilité aux soins de santé, considérant l'immuabilité de certaines d'entre elles.

2.5 Données et modélisation

2.5.1 Base de données et dossier médical électronique

La compagnie *International Business Machine* (IBM) décrit les mégadonnées en utilisant l'acronyme des « 4V » [48]. Le volume et la variété correspondent respectivement au nombre total de valeurs dans la table et au nombre de différentes variables, ou colonnes de la table. La vélocité correspond à la fréquence d'enregistrement des données. Celle-ci peut atteindre 150 Hz par le monitoring haute-fidélité des signes vitaux, ou être mesurée de façon inconstante lorsque le patient fait ses analyses sanguines en externe [4]. Le dernier V correspond à la véracité et adresse la fiabilité des données enregistrées : à quel point les données reflètent-elles réellement la réalité? Un groupe a démontré qu'à travers trois bases de données médicales regroupant les mêmes patients, le coefficient de corrélation κ moyen entre STS (*Society of Thoracic Surgery*) et le DME était de 0.391, et entre STS et les codes de comorbidités ICD9 (*International Classification of Disease, 9th Edition)*, de 0.225 [49]. En exemple, le diabète était rapporté chez 29.1%, 37.1% et 65.5% des mêmes patients, selon chaque banque de données. Ces chiffres démontrent aisément la limite de fiabilité des bases de données médicales dans le développement et l'utilisation de modèles, ce qui peut aussi expliquer certaines discordances de performance enregistrées lors des validations externes de ces modèles. Dans la cadre de ce projet, ces observations confirment l'importance de valider la concordance des variables rapportant une information similaire, par exemple le degré d'urgence de la chirurgie. En cas de discordance entre les variables, il vaut mieux sélectionner la moins susceptible aux erreurs d'enregistrement.

2.5.2 Stratégie de validation des modèles

Bien que la véracité des données soit un élément critique, il importe de s'assurer que nos modèles performent bien lorsque de nouveaux jeux de données lui sont présentés. Différentes stratégies de validation sont rapportées dans la littérature. La plus simple consiste à séparer le jeu de données en entraînement et validation, selon un ratio 50 : 50 à 70 : 30 [50, 51]. Suivant l'apprentissage à partir du jeu d'entrainement, le modèle est testé dans le jeu de validation et le modèle le plus performant en validation est retenu. Cette stratégie est peu utilisée puisqu'un modèle ayant aléatoirement

surappris l'ensemble de validation offrira une grande performance et sera sélectionné comme modèle malgré une faible performance dans de nouveaux jeux de données. Le surapprentissage survient lorsqu'un modèle devient trop spécifique aux détails, ou bruit, du jeu d'entraînement et perd son aptitude à généraliser ses conclusions dans de nouveaux de données.

Pour y pallier, plusieurs auteurs utilisaient la stratégie de validation croisée avec replis (k-fold), dans laquelle la base de données est divisée en 5 sous-groupes [4, 28, 30, 31]; le modèle est testé 5 fois sur chacun des sous-groupes, après avoir été entraîné sur les 4 autres [50, 51]. Le modèle est donc entraîné sur 80% des données et testé sur 20%, à 5 reprises, et les performances finales correspondent à la moyenne des performances des 5 modèles.

Une approche également observée dans la littérature médicale est de combiner l'une des stratégies précédentes à une validation temporelle[29]. Un jeu de données est isolé basé sur une année de procédure plus récente, par exemple 2020-2022. Le modèle est entraîné avec une validation croisée sur le jeu de données préalable, et une fois le meilleur modèle sélectionné, sa performance est testée dans le jeu de données isolées. Cette approche permet de reproduire l'application prospective du modèle sur de nouveaux patients, simulant un déploiement du modèle en hôpital.

2.5.3 Algorithmes non-supervisés et segmentation

Tel que discuté préalablement, identifier les phénotypes au sein d'une population hétérogène implique d'abord de segmenter la population en sous-groupes. Les algorithmes non-supervisés effectuent cette tâche, sans influence humaine directe outre le choix d'hyperparamètres, en regroupant les patients partageant des caractéristiques clés. Ils visent à minimiser la distance interne d'un sous-groupe et maximiser la distance entre deux sous-groupes. Les deux publications adressant le phénotypage des patients en sepsis aux soins intensifs et des patients atteints de COVID-19 ont utilisé l'algorithme de *consensus k-means*, dérivé de l'algorithme *k-means*, pour y parvenir [13, 15].

K-means

L'algorithme *k-means* divise les patients en calculant la distance de chaque patient, ou point, par rapport au centre de gravité du sous-groupe [52]. Le nombre *k* de sous-groupes est établi comme hyperparamètre du modèle qui est initié par la sélection aléatoire de *k* points. À chaque itération, tous les points sont attribués au sous-groupe ayant le centre de gravité le plus proche, et le nouveau centre de gravité de ce sous-groupe est recalculé pour la prochaine itération. Le traitement s'arrête

lorsque les centres de gravité cessent de se déplacer. Bien que rapide et puissant, la segmentation du *k-means* varie selon les points aléatoirement sélectionnés à l'initialisation [52, 53]. La reproductibilité est donc inconstante. Pour y pallier, le *consensus k-means* ajoute une étape.

Consensus k-means

Pour pallier aux résultats variables du *k-means*, le *consensus k-means* répète *n* fois le *k-means* en variant les points sélectionnés en début d'exécution de l'algorithme [14]. En bâtissant une matrice de l'attribution des patients à chaque itération, l'algorithme analyse le nombre de fois que chaque paire de points s'est vu attribué le même sous-groupe. Le consensus d'attribution sera établi en enchaînant un algorithme hiérarchique sur cette matrice, déterminant donc si chaque paire de patients sera dans le même sous-groupe ou sera séparée. Cet algorithme peut s'approches du modèle supervisé de forêt aléatoire, décrit par la suite, qui utilise également la notion de consensus.

Techniques de réduction de dimensionalité

Bien qu'à ce jour peu appliqués dans la recherche périopératoire et en soins intensifs, les modèles de réduction de dimensionnalité sont grandement utilisés dans la génomique et protéinomique. L'analyse de composante principale (PCA) tente d'éliminer certaines redondances en créant un nouveau système de dimensions expliquant mieux les données observées alors que l'algorithme t-DSNE (*t-Distributed Stochastic Neighbor Embedding*) minimise l'écart entre une distribution de probabilité calculée dans l'espace original, et celle créée dans un second espace de plus petite dimension. Ces deux approches sont également utilisées pour créer une représentation graphique en deux dimensions des jeux de données en comportant beaucoup plus.

2.5.4 Algorithmes supervisés

Les algorithmes supervisés permettent la prédiction et la classification. Le terme supervisé provient du choix, par les auteurs, de la variable *y* à prédire à partir des variables données aux modèles. Trois architectures différentes seront ici traitées : les modèles linéaires (logistique simple, *lasso*, *ridge* et *elastic net*), les arbres de classification et forêt aléatoires, et les réseaux de neurones.

Modèles linéaires

La régression logistique est un modèle supposant la linéarité des variables à l'étude. Elle attribue un coefficient b à chaque variable x du modèle, afin d'optimiser la capacité prédictive du modèle sur la variable binaire y.

$$P(y = 1|X) = \frac{e^{b_0 + b_1 x_1 + \ldots + b_n x_n}}{1 + e^{b_0 + b_1 x_1 + \ldots + b_n x_n}}$$

Les variables peu influentes obtiendront un coefficient b tendant vers 0, alors que les variables influentes s'en éloigneront progressivement vers le négatif ou le positif, selon sa corrélation positive ou négative à la variable y [54]. Les variables explicatives sont ainsi facilement obtenues par la réorganisation décroissante des valeurs absolues des coefficients $b_1 \ldots b_n$. La régression logistique simple a tendance à surapprendre lorsque les bases de données sont grandes.

Pour la corriger cette tendance, deux pénalités, ou régularisations, peuvent être ajoutées afin d'ajuster le biais et la variance du modèle [54]. L'application de la pénalité L_1, nommée régression *lasso*, écrase les coefficients moins influents vers zéro de manière à limiter le surapprentissage, ou une forte variance. De son côté, la pénalité L_2, nommée régression *ridge*, évite qu'une variable devienne trop influente en limitant la valeur qu'un coefficient peut prendre. Un modèle appliquant les deux pénalités porte le nom *elastic net*. La plupart des publications explorent l'un de ces trois modèles plutôt que la régression logistique simple, étant donné l'ampleur des jeux de données médicaux [4, 29, 30].

Arbres et forêts aléatoires

L'arbre de classification est constitué d'une série d'embranchements menant à des feuilles finales. À chaque embranchement, le patient est classifié selon la question posée, par exemple si l'âge est supérieur ou inférieur à 65 ans. Le dernier embranchement mène à la feuille finale, et la prédiction de cette feuille est donnée par la classe majoritaire s'y retrouvant [55]. Pour bâtir l'arbre, l'algorithme teste séquentiellement les variables et embranchements possibles. L'index Gini et la mesure d'entropie testent à leur manière la pureté des feuilles nouvellement créées, et l'embranchement optimisant la pureté est sélectionné par l'algorithme. La faiblesse inhérente à l'arbre de classification provient de la dichotomie des choix et l'instabilité des frontières de décision, par exemple si deux patients de 64 et 66 ans ont des prédictions significativement

différentes, car l'embranchement divise à 65 ans. Par la même explication, les arbres profonds, définis selon les hyperparamètres d'entrée, tendent au surapprentissage.

Pour y pallier, l'algorithme de forêts aléatoires génère un nombre prédéfini d'arbres de classification à partir d'une portion réduite des variables d'entrées et des observations [55]. Il crée ainsi des ensembles d'apprentissage différents, entraîne un arbre sur chacun d'eux, et détermine la classe finale d'une observation en sélectionnant la prédiction la plus populaire dans les arbres de classification.

Tout comme la régression logistique, la forêt aléatoire est un algorithme fréquemment inclus dans l'ensemble des modèles testés. Ce modèle est typiquement performant en périopératoire. Certains auteurs ont obtenu leur meilleure prédiction de mortalité par cette approche [4], et d'autres issues cliniques telles la réadmission postopératoire à l'urgence, avaient une performance très près d'optimale [29].

Bien que la forêt aléatoire ne crée pas un arbre facile à dessiner et analyser, l'importance relative de chaque variable du modèle peut être extraite par différentes stratégies, assurant l'interprétabilité des valeurs explicatives.

Réseaux de neurones et perceptron multicouche

Les réseaux de neurones sont constitués d'une couche d'entrée correspondant aux variables de la base de données, d'un nombre prédéterminé de couches cachées ayant chacune un nombre prédéterminé de neurones, et d'une couche de sortie correspondant à la valeur y à prédire [56]. Tel qu'inspiré des neurones du corps humain, les neurones des couches cachées reçoivent une stimulation correspondante à la somme, pondérée par le vecteur de poids w des neurones x activées à la couche précédente. Une fonction d'activation g détermine l'activité de ce neurone, et transmet une stimulation z à la couche suivante selon la formule :

$$z = g(w^T x + w_0)$$

Le réseau de neurones est initié avec des poids aléatoires qui sont réajustés à chaque itération d'apprentissage par rétropropagation afin de minimiser une fonction de perte. Étant donné la manipulation non linéaire et complexe des couches cachées, les réseaux de neurones n'offrent pas d'interprétabilité des variables explicatrices et portent ainsi le nom de « boîte noire », ce qui rebute certains experts du domaine [56]. Bien que les réseaux de neurones soient grandement utilisés pour prédire la trajectoire de soins postopératoires, telle la mortalité dans le contexte de ce travail,

plusieurs publications cherchent à éclaircir cette boite noire pour augmenter l'acceptabilité et faciliter l'implantation de ce type d'algorithme en pratique clinique [6, 7, 57].

2.6 Synthèse de la revue de littérature

En conclusion, la définition de trajectoire de soins peut inclure toutes les complications, que ce soit la mortalité, le séjour prolongé, ou la survenue d'insuffisance rénale aiguë, pourvu qu'elle soit à la fois intéressante pour le patient et pour le système de santé. Le score ASA demeure l'outil de stratification de risque préopératoire le plus utilisé, bien que les dernières décennies aient vu plusieurs nouveaux outils apparaître. Ces outils prédisaient typiquement la mortalité, mais l'intérêt grandissant de la communauté médicale pour l'apprentissage machine a contribué à développer de nouveaux scores se basant sur le DME. Inspiré d'autres disciplines comme la psychiatrie, le phénotypage est l'un de ces nouveaux outils et deux récentes publications ont confirmé l'intérêt de son application en soins critiques [13, 15]. Ce projet de recherche s'inscrit donc dans le contexte d'un intérêt grandissant pour l'apprentissage machine et pour l'applicabilité du phénotypage à des groupes de patients hétérogènes pouvant présenter des trajectoires de soins compliqués.

CHAPITRE 3 MÉTHODOLOGIE

Une méthodologie rigoureuse est la pierre angulaire d'un projet de recherche réussi. Elle permet la révision scientifique critique par les pairs, la transparence sur les forces et les faiblesses des résultats et la reproductibilité des résultats nécessaire à la validité scientifique. La majorité des disciplines scientifiques ont des lignes directrices fréquemment révisées encadrant la démarche scientifique et facilitant l'écriture des travaux de recherche. La méthodologie de cette thèse sera présentée en accord avec 1) la méthode *Cross Industry Standard Process for Data Mining* (CRISP-DM) issues de la science des données, et 2) les lignes directrices encadrant le développement et la présentation des recherches biomédicales exploitant l'apprentissage machine [58].

Ce chapitre présentera séquentiellement les étapes cruciales de la compréhension des données, la préparation des données, la modélisation, et l'évaluation des résultats, alors que le chapitre 4 appliquera cette méthodologie aux données périopératoires de UCLA en vue de créer et évaluer les phénotypes de patients chirurgicaux.

3.1 Compréhension des données

3.1.1 Source des données

La source de tout projet d'apprentissage machine est la base de données. Prendre le temps de maîtriser les caractéristiques de cette base de données est un investissement, certes chronophage, mais rentable. Il incombe de comprendre la définition des variables, la ou les stratégies de saisie de données et les sources d'erreurs d'entrée, pour ne nommer que quelques caractéristiques, afin d'exploiter les forces de la base de données, et compenser ses limites.

Avec la conscientisation de la valeur des données, plusieurs institutions publiques et privées en font désormais la collecte à des fins d'analyse, d'amélioration de processus, d'archives légales, ou parfois même sans objectif préétabli avec l'ambition d'éventuellement les exploiter. Il est maintenant connu qu'elles sont l'or du 21^{e} siècle, et il en résulte une pléiade de bases de données avec parfois des millions, voire des milliards de valeurs.

La première étape d'un projet est donc la sélection, ou création, d'une base de données permettant de répondre à nos questions de recherche. Cette base de données, structurée sous forme de tables, doit comporter suffisamment d'attributs, ou colonnes, pour permettre des résultats intéressants, et suffisamment de lignes, ou objets, pour le type de modèle que nous souhaitons entraîner. L'expert du domaine sera le mieux outillé pour identifier les variables pertinentes à conserver dans la base de données, alors que l'ingénieur de données aura une meilleure intuition sur le nombre d'exemples nécessaires. Un modèle prédictif par apprentissage supervisé pourra converger et générer des résultats intéressants avec quelques centaines ou milliers d'entrées, alors qu'un apprentissage profond pourra nécessiter des centaines de milliers d'entrées selon l'architecture sélectionnée. Le choix final doit donc reposer sur une communication entre l'expert du domaine et l'ingénieur en données. Comme le dit l'adage « *Junk in, junk out* », une mauvaise sélection de bases de données condamnera dès lors les résultats du projet de recherche.

Dans les grandes banques de données, telles celles issues des dossiers médicaux électroniques, il est fréquent de devoir regrouper plusieurs tables, initialement séparées afin d'en faciliter la lecture ou l'écriture. Ces jointures de tables se font à partir de clés d'identification uniques qui doivent se retrouver dans les deux tables jointes. Plusieurs clés d'identification de différentes hiérarchies peuvent exister, et il est primordial de les comprendre et maîtriser leur degré de granularité avant de créer des jointures, sans quoi plusieurs erreurs en découleront.

3.1.2 Description et validation des données

Une fois la banque de données sélectionnée, il faut apprivoiser ses caractéristiques. Il est primordial de comprendre l'origine des données, la stratégie de collecte, les sources d'erreurs ou artéfacts, en plus d'avoir une description de chaque variable. Une donnée saisie automatiquement est à risque d'enregistrer des artéfacts, par exemple lorsqu'un moniteur de pression artérielle enregistre une valeur faussement élevée liée à un mouvement inapproprié lors du gonflement du brassard pneumatique. Inversement, une variable saisie manuellement comportera plus souvent des erreurs de négligence (erreurs de frappe, entrées incomplètes) ou même des erreurs intentionnelles. En effet, une valeur à la limite de la normale pourrait être manuellement modifiée pour éviter de justifier une valeur limite, par exemple.

Exploration individuelle des variables

Une analyse sommaire permettra d'apprécier le type de variable, son amplitude, sa dispersion, les nombres de valeurs manquantes ou nulles, etc. Les variables cibles du projet de recherche doivent être présentes ou doivent pouvoir être construites.

Les différents types des variables sont :

1) Binaire ou booléenne : définie par {1,0} ou {True, False};
2) quantitative (numérique) : discrète ou continue;
3) qualitative (catégorique) : ordinale ou nominale, respectivement avec ou sans notion de progression ou hiérarchie entre les catégories;
4) ou libre : lorsque l'entrée est un champ de texte libre limité ou non par des règles alphanumériques.

L'exploration de chaque variable dépendra directement de son type. Une variable binaire se résumera principalement à la proportion de « 1 », ou « True », alors qu'une variable continue permettra d'analyser le type de distribution, les mesures de tendances centrales (moyenne, médiane, mode), ainsi que la dispersion (déviation standard, écart interquartile, min, max). Nous pourrons analyser les variables extrêmes et leur comportement comparativement au reste de la distribution.

Dans le cas d'une variable qualitative, nous analyserons la répartition dans chaque catégorie et, conjointement avec l'expert si nécessaire, pourrons rechercher une ambiguïté ou chevauchement entre différentes catégories. Finalement, le texte libre peut être plus difficile à analyser. Après avoir uniformisé la casse (majuscules ou minuscules), retiré les accents, et repéré des synonymes, nous pourrons rechercher, par exemple, le taux d'incidence d'une certaine expression ou groupe de caractères dans le texte (ex : le terme « douleur » se retrouve dans 12% des entrées), ou utiliser des algorithmes de segmentation ou apprentissage profond pour en extraire d'autres informations quantitatives. Ici également, la collaboration avec l'expert du domaine sera très utile.

Exploration de l'interdépendance des variables

Analyser l'interdépendance des variables permet d'identifier les erreurs de logique et les corrélations. Une erreur de logique survient lorsque plusieurs différentes données pour un même objet sont discordantes, par exemple lorsque le calcul de l'IMC basé sur la taille et le poids du

patient ne correspond pas à la valeur d'IMC enregistrée pour le patient. Ce type d'erreur est plus difficile à repérer puisqu'individuellement, chaque variable peut sembler adéquate.
La corrélation des variables peut être investiguée par différentes approches. La corrélation linéaire de Pearson est calculée avec la formule :

$$r_{x,y} = \frac{E(XY) - E(X) \cdot E(Y)}{\sigma_X \cdot \sigma_Y}$$

où X et Y sont respectivement la valeur explicative et la valeur cible. Aucune corrélation linéaire n'existe si $r_{X,Y}$ vaut 0, alors qu'elle augmentera progressivement si elle tend vers -1 (corrélation fortement négative) ou +1 (corrélation fortement positive). Une corrélation est jugée faible dans l'intervalle [-0.5; 0.5], alors qu'elle sera jugée forte autrement.
La librairie *SciPy* offre la fonction *pearsonr*, permettant d'automatiquement extraire la corrélation et la valeur *p*, quantifiant la probabilité que la corrélation observée soit liée au hasard. En établissant la valeur seuil de p à 0.05, nous pouvons établir la valeur $r_{X,Y}$ minimale pour considérer la corrélation statistiquement significative.

Il est possible d'utiliser le coefficient de Pearson à la fois pour identifier les variables fortement corrélées aux variables *y* que nous étudions, donc pertinentes à notre modèle, et pour identifier des variables redondantes qui pourraient alourdir le modèle sans l'améliorer. Ces dernières seront retirées au moment de nettoyer les données.

3.2 Préparation des données

Suivant l'observation et la validation des données, l'étape de préparation consiste à manipuler les données en éliminant, si besoin, celles non pertinentes. La préparation des données implique aussi l'émission d'hypothèses de recherche qu'il sera important de bien décrire puisqu'elles peuvent altérer, voir invalider, les résultats de nos modèles. À la fin de cette étape, les données seront prêtes pour la modélisation.

3.2.1 Sélection/restriction des données

Les banques de données sont souvent bâties avant l'élaboration de questions de recherche, entraînant un superflu de colonnes, ou variables, non pertinentes à notre question de recherche. Afin de les éliminer, nous pouvons faire appel à l'expert du domaine et utiliser notre exploration statistique.

L'expert du domaine aidera à filtrer les variables inappropriées pour notre projet. Si nous souhaitons faire une prédiction, nous devons limiter nos variables à celles disponibles au moment de faire la prédiction. Par exemple, nous ne pouvons pas utiliser les signes vitaux d'un patient durant une chirurgie si nous créons un modèle prédictif utilisé avant l'opération.

Ensuite, nos explorations statistiques peuvent retenir uniquement les variables d'intérêt atteignant un niveau de corrélation minimal avec la ou les variables y que nous souhaitons prédire. Nous pouvons faire le choix d'éliminer complètement les autres variables de notre modèle. Selon le type de modèle, un débalancement des classes peut influencer le résultat de nos modèles. Un modèle prédictif de mortalité postopératoire aura souvent une faible proportion de cas positifs, auquel cas des approches de sur- ou sous-échantillonnage peuvent être appliquées pour rebalancer les classes. Dans notre exemple, un sous-échantillonnage pourrait restreindre le nombre de patients ayant survécu, alors qu'un suréchantillonnage dupliquerait les patients décédés.

3.2.2 Modification et création de variables

Alors qu'un réseau de neurones pourra identifier comment combiner deux ou plusieurs variables entre elles, un modèle linéaire pourra nécessiter la création manuelle de variables afin de bien performer. L'exemple de l'IMC demeure pertinent puisqu'il implique une division et le carré d'une variable. La base de données initiale pourrait contenir le poids et la taille, à partir de laquelle nous pourrons créer la nouvelle variable « IMC ». Il faut demeurer prudent puisque chaque variable créée est à risque de créer des corrélations avec les variables initiales, dont il faudra peut-être même se débarrasser.

Une variable binaire sera uniformisée à {1,0} pour faciliter l'interprétation par le modèle. De son côté, une variable catégorique peut être transformée en variables indicatrices, c'est-à-dire créer une variable binaire pour chacune des classes que la variable catégorique peut prendre. Cette approche standardise la distance entre deux catégories, ce qui est essentiel pour les variables catégoriques nominales, mais peut entraîner une perte d'information pour les variables ordinales. Les

algorithmes d'apprentissage non supervisés sont particulièrement sensibles à de telles distances. Un autre exemple concerne l'extraction d'informations d'un texte libre. Il serait utile d'extraire la présence d'une certaine expression dans le texte libre, tel le nom d'une chirurgie et ainsi créer une nouvelle variable binaire rapportant cette information. Une fois de plus, l'apport de l'expert et celui de l'exploration initiale des données offriront une intuition sur ces potentielles variables à créer.

3.2.3 Nettoyage des données

Aux étapes précédentes, nous travaillions davantage au niveau des lignes et des colonnes. Au nettoyage, nous corrigeons les données et identifions les erreurs, les valeurs éloignées et les valeurs manquantes.

Tel qu'abordé, trois types d'erreurs peuvent survenir : des erreurs de négligence, des artéfacts, ou des erreurs de logique. Lorsqu'identifiées, ces erreurs ne reflètent pas la réalité et devraient être considérées comme une valeur manquante. Il incombe alors de gérer ces données manquantes puisque certains algorithmes ne peuvent pas les tolérer. Nous pouvons soit retirer complètement la ligne ou la colonne, ou imputer la valeur manquante. La stratégie d'imputation doit être choisie avec soin, puisqu'elle affectera les résultats si le nombre de données manquantes est élevé. Il est possible

1) d'établir une valeur par défaut;
2) d'utiliser une mesure de tendance centrale (moyenne, mode, médiane);
3) de modéliser la distribution de la variation afin de tirer aléatoirement une valeur dans la distribution; ou
4) d'imputer une estimation de la variable basée sur d'autres variables avec laquelle elle est typiquement corrélée.

Une combinaison de ces approches est souvent appliquée. Par exemple, si une variable comporte un pourcentage de valeurs manquantes dépassant le seuil préétabli, la variable sera éliminée; alors que la valeur sera imputée en deçà de ce seuil.

3.2.4 Gestion de la corrélation entre les variables

En accord avec nos analyses d'interdépendance des variables, nous pouvons supprimer celles ayant une trop forte corrélation entre elles. Il est possible de laisser l'expert sélectionner les colonnes à

éliminer basé sur l'interdépendance, ou de standardiser l'approche en extrayant chaque paire de variables corrélées, et conserver uniquement celle ayant la plus grande corrélation avec les variables y. Une autre approche est de débuter par l'élimination de la variable ayant le plus grand nombre de corrélations significatives, réduisant ainsi le nombre de variables éliminées. Ces deux approches « glouton » peuvent être combinées à une sélection par l'expert assurant la bonne manipulation des variables.

3.2.5 Séparation des données : dérivation et validation

Un modèle d'apprentissage machine est aussi bon que sa capacité de maintenir son efficacité sur d'autres données que celles initialement utilisées pour sa création. En d'autres mots, il doit avoir une bonne validité externe. L'approche standard de validation croisée implique de diviser la base de données en un ensemble de dérivation et un ensemble de test, ce dernier comportant entre 30 et 50% des données. L'ensemble de dérivation sera utilisé pour créer le modèle, alors que l'ensemble de test servira uniquement à tester le modèle avec un jeu de données jusqu'alors jamais vu par le modèle. Un modèle performant en entraînement et peu performant en test aura surappris, et ne pourra jamais être déployé étant donné son inaptitude à analyser de nouvelles données. L'attribution de chaque ligne à l'un des deux ensembles peut être faite aléatoirement ou en utilisant l'une des variables de la base de données, par exemple l'année de la chirurgie du patient.

Pour créer le modèle et selon l'architecture, l'ensemble de dérivation peut ensuite être divisé en ensemble d'entraînement et de validation. Ces derniers seront utilisés pour trouver les hyperparamètres optimisant l'efficacité du modèle, sans atteindre un surapprentissage, tel que décrit dans la section « Modèle ». Afin d'éviter de scinder les données à nouveau en deux groupes, certaines alternatives existent. La stratégie de repli (*k-fold*) implique de séparer le jeu de données en k groupes égaux, par exemple 5 groupes avec 20% des données chacun. Pour un même jeu d'hyperparamètres, le modèle est séquentiellement entrainé sur 4 groupes (80% des données) et testé sur un seul (20% des données). La performance du modèle correspond à la moyenne des 5 performances enregistrées sur chacun des 5 groupes, après avoir été entraîné sur les 4 autres. Ce modèle permet une meilleure généralisabilité des modèles développés. Finalement, une dernière approche est disponible pour le modèle de forêt aléatoire, avec l'approche « *out-of-bag* » qui testera chaque arbre, au fur et à mesure de leur création, sur les données non utilisées par cet arbre de la forêt.

3.2.6 Standardisation des données

Les variables continues peuvent être normalisées de manière à retirer l'influence de leur amplitude, par exemple avec un score de douleur gradé de 1 à 10 comparativement à un âge gradé de 1 à 100. Plusieurs approches, comportant chacune leurs forces et faiblesses, peuvent être utilisées et il est souvent utile d'en explorer plusieurs pour optimiser notre modélisation :

1) La normalisation standard est la plus classique, dans laquelle la distribution est centralisée à 0 et normalisée pour un écart-type de 1 (équivalent de la cote Z) selon la formule :

$$v'(i) = \frac{v(i) - \text{mean}\,(v(i))}{std\_dev(v(i))}$$

2) La normalisation *MinMax* positionne chaque valeur entre 0 et 1, avec la plus petite valeur à 0 et la plus grande à 1, selon la formule :

$$v'(i) = \frac{v(i) - \min\,(v(i))}{\max\big(v(i)\big) - \min\,(v(i))}$$

3) La catégorisation, dans laquelle des intervalles sont créés avec une variable binaire correspondante, reproduisant les variables indicatrices décrites précédemment. Cette approche peut se faire avec ou sans lissage permettant de limiter le nombre de valeurs possibles et faciliter le regroupement, telle la troncature des décimales.

Une normalisation standard aura un comportement différent de celle *MinMax* et influera inévitablement les résultats de nos modèles. Par exemple, une valeur éloignée située à 4 déviations standards de la norme, serait transformé à une valeur 4 par la normalisation standard, alors qu'elle vaudrait 1 avec l'algorithme *MinMax*, qui au lieu, écraserait toutes les autres valeurs vers 0. L'algorithme de standardisation doit se créer uniquement dans l'ensemble de dérivation avant d'être appliqué sur l'ensemble de test. Cette approche respecte l'idée que le modèle est initialement aveugle aux données sur lesquelles il sera testé.

3.3 Modélisation

Une fois rendu à l'étape de modélisation, le jeu de données est prêt pour créer, entrainer et valider nos modèles d'apprentissage machine. Plusieurs différents modèles et architectures sont testés, il est donc important d'éviter la modification de notre jeu de données pour assurer la comparabilité de nos résultats.

3.3.1 Objectifs du modèle

Les objectifs de notre projet guideront le type de modèle utilisé. Un objectif de prédiction ou optimisation orientera vers des modèles d'apprentissage supervisé ou par renforcement, alors qu'une classification pourra également explorer des modèles non supervisés. Au-delà des résultats, le besoin d'interpréter les résultats, les sources de biais dans nos données, le degré de corrélation entre les variables influenceront également notre choix de modèle.

3.3.2 Entrainement du modèle

L'ensemble de données d'entraînement est utilisé par le modèle pour qu'il identifie la meilleure manière d'utiliser les variables d'entrée pour fournir le résultat souhaité. Par exemple, dans un problème de prédiction, le modèle modifiera à chaque itération les poids associés aux variables afin qu'il améliore ses performances de prédiction. Le modèle convergera s'il atteint une stabilité dans les poids attribués aux variables, minimisant son écart à la réalité. Ce minimum peut être un le minimum absolu s'il correspond à la meilleure solution possible, ou être un minimum dit local si l'algorithme converge vers une solution, mais que cette solution n'est pas celle optimale. Ce concept explique l'importance de tester plusieurs modèles, ainsi que de faire varier les hyperparamètres du modèle afin d'explorer l'espace de solution et ne pas se contenter d'un minimum local.

3.3.3 Validation du modèle

Des métriques préétablies servent à évaluer quantitativement la performance d'un modèle. Celles-ci diffèrent en fonction du type de modèle utilisé.

Métriques de performance pour algorithmes non-supervisés

Trois métriques sont utilisées dans le cadre de ce projet pour évaluer la performance statistique de la segmentation et sélectionner la meilleure modélisation pour phénotyper les patients de l'ensemble test. Le score de Silhouette permet de comparer la densité intra-groupe à la distance intergroupe. Une valeur de 1 témoigne de groupes denses et éloignés les uns des autres, et une valeur à 0 reflète l'inverse.

$$Silhouette = \frac{b-a}{\max(a,b)}$$

où a est la moyenne des distances à l'intérieur d'un sous-groupe et b est la moyenne des distances entre les différents sous-groupes. Les scores NMI (*Normalized Mutual Info*) et l'index de Rand ajusté (IRA) se distinguent du score Silhouette puisqu'ils analysent la segmentation des sous-groupes par rapport à une variable, par exemple la variable *y* que l'on souhaite prédire, et impliquent de connaître la réalité de terrain de *y*. L'index NMI provient du score MI (*Mutual Info*), qui témoigne de la quantité d'information qu'il est possible d'extraire d'une distribution en analysant une seconde distribution.

$$MI(X;Y) = \sum_{y\epsilon Y}\sum_{x\epsilon X} p_{(X,Y)}(x,y)\log\left(\frac{p_{(X,Y)}(x,y)}{p_{(X)}(x)p_{(Y)}(y)}\right)$$

En assumant que la distribution *X*, donc les sous-groupes attribués par l'algorithme de segmentation, et la distribution de variable *y* soient indépendantes et aléatoires, le logarithme écrasera le score MI à 0. Le score NMI normalise le score MI par la somme des entropies de chaque distribution, ce qui compense le débalancement des distributions, tel qu'attendu dans ce projet puisque la majorité des patients sont sains et qu'une minorité est considérée à haut risque.

$$NMI(Y,X) = \frac{2 * MI(Y;X)}{H(Y) + H(Y)}$$

L'IRA provient de l'index de Rand (IR) qui calcule le nombre de paires d'éléments bien classifiées, basé sur la réalité de terrain de la variable *y*.

$$\text{IR} = \frac{a+b}{a+b+c+d}$$

Où :

a est le nombre de paires d'éléments *x* partageant le même sous-groupe *x* et la même valeur *y*;
b est le nombre de paires d'éléments *x* ne partageant ni le même sous-groupe, ni la variable *y*;
c est le nombre de paires d'éléments *x* partageants la variable *x* avec une variable *y* différente;
d est le nombre de paires d'éléments *x* ne partageant pas le même sous-groupe, mais partageant *y*.

Le problème majeur associé à l'index de Rand est la variation du résultat lorsque des sous-groupes sont aléatoires et indépendants, plutôt que résulter en une valeur constante et faible. Avec un grand nombre de sous-groupes, la valeur de *b* peut rapidement augmenter malgré la segmentation aléatoire. L'IRA compense cette lacune en introduisant une table de contingence représentant cette composante aléatoire. L'index de Rand obtenu est ainsi normalisé avec l'index de Rand attendu si la variable était aléatoire, et l'index de Rand théoriquement optimal.

$$\text{IRA} = \frac{RI\ obtenu - Valeur\ attendu}{Valeur\ optimale - valeur\ attendue}$$

Il en résulte une valeur interprétable, valant 1 si la relation est parfaite, et 0 si elle est aléatoire.

Métriques de performance pour algorithmes supervisés

Dans le cas d'apprentissage supervisé appliqué à une classification binaire telle que vue avec les cinq issues cliniques de ce projet, le modèle sera testé sur les valeurs de l'ensemble de test. Les prédictions du modèle seront comparées au résultat réel, permettant de créer une matrice de confusion.

Tableau 3.1 Matrice de confusion

	Prédit vrai	**Prédit faux**
Réel vrai	Vrais positifs (VP)	Faux négatifs (FN)
Réel faux	Faux positifs (FP)	Vrais négatifs (VN)

À partir des valeurs de cette matrice, les métriques de performance pourront être calculées. Le taux de bonnes prédictions correspond au nombre de bonnes prédictions sur le nombre total de prédictions :

$$Taux\ de\ bonnes\ prédictions = \frac{VP + VN}{VP + FP + VN + FN}$$

Ce score performe mal lorsque les classes sont fortement déséquilibrées, par exemple avec les complications postopératoires. Un modèle prédisant constamment la survie post-opératoire dans une population présentant 2% de mortalité atteindrait l'excellent score de 98%.
Les scores de rappel et de précision retirent les vrais négatifs du numérateur, offrant donc une meilleure analyse sur la capacité du modèle à identifier les complications.

$$Précision = \frac{VP}{VP + FP}$$

$$Rappel = \frac{VP}{VP + FN}$$

Le score F1 utilise les deux dernières métriques pour évaluer la quantité de signaux captés par notre modèle.

$$Score\ F1 = 2 * \frac{Précision * rappel}{Précision + rappel}$$

Chacune de ces formules offre une métrique unique pour évaluer la performance du modèle. Pour contextualiser ses performances, la courbe ROC (R*eceiver Operating Characteristics*) est utilisée, ainsi que l'aire sous la courbe, AUROC (*Area under ROC*) pour valider les performances du modèle. La courbe ROC représente le taux de vrais positifs obtenus en fonction de la valeur seuil que nous utilisons pour classifier notre prédiction comme positive. L'AUROC correspond donc à la force de notre classificateur. Elle vaut 0.5 si la classification binaire est aléatoire, et atteint 1 s'il obtient des résultats parfaits.

3.4 Analyse des résultats

L'analyse des résultats est un travail continu s'étalant de l'obtention des résultats jusqu'à la rédaction du papier scientifique. À partir des métriques de performance, les modèles seront comparés entre eux. Nous évaluerons ainsi les meilleures stratégies de découpages des variables d'entrée, les types d'architectures les plus performantes ainsi que leurs hyperparamètres. Si l'interprétabilité du modèle le permet, les variables les plus utilisées par le modèle seront également évaluées. Plusieurs itérations seront appliquées à la préparation de la base de données ainsi qu'à l'entraînement des modèles.

L'analyse des résultats doit s'aligner avec les objectifs du projet initialement définis. Il faut s'assurer de répondre à notre question de recherche et d'éviter ou minimiser toutes sources de biais pouvant influencer la qualité de résultat. Plusieurs nouvelles questions de recherche émaneront de cette analyse et pourront guider les projets futurs.

CHAPITRE 4 APPLICATION DE LA MÉTHODE AUX TRAJECTOIRES DE SOIN POSTOPÉRATOIRE

Ce chapitre présente l'application de la méthodologie décrite au chapitre précédent, dans le contexte spécifique des trajectoires de soin postopératoire des patients ayant subi une chirurgie dans le réseau hospitalier lié à l'Université de Californie à Los Angeles (UCLA).

4.1 Compréhension des données

Comme son nom l'indique, UCLA se situe à Los Angeles et offre des soins médicaux à l'ensemble de la population via le réseau UCLA Health, totalisant près de 4 millions d'habitants. Contrairement au système de santé québécois, UCLA Health compétitionne avec d'autres grandes institutions médicales. L'hôpital Ronald Reagan est le centre quaternaire de UCLA Health, offrant les soins médicaux les plus spécialisés, et plusieurs autres blocs opératoires de différentes tailles œuvrent au sein du réseau. Annuellement, 50,000 chirurgies sont pratiquées au sein du réseau UCLA.

En 2013, UCLA Health a intégré le dossier médical électronique Epic et a commencé à peupler la banque de données actuelle du réseau bâtissant ainsi peu à peu un gigantesque entrepôt de données sur près de 533,000 chirurgies et plusieurs millions de visites de patients. L'ensemble des institutions médicales du réseau utilisent Epic et donc, toutes les données de patients traités à UCLA Health y sont enregistrées en temps réel, incluant les visites en clinique externe, les hospitalisations, les laboratoires, les imageries médicales et parfois même, le génome. Afin d'encourager la recherche, UCLA rend ces données accessibles aux chercheurs en les anonymisant via l'élimination de toutes données d'identification (nom, adresse, téléphone), en modifiant les dates de visites et d'opération et en modifiant certaines valeurs lorsqu'elles deviennent trop discriminatives (l'âge d'un patient est limité à 90 à partir du moment où le patient l'atteint). Les biais intrinsèques de cette banque de données sont principalement ceux liés à la population, notamment le statut socio-économique permettant l'accès aux soins de santé aux États-Unis, la population majoritairement caucasienne et anglophone habitant le sud-ouest des États-Unis.

La période périopératoire se déroule typiquement entièrement à l'hôpital, permettant d'accumuler une grande richesse de données sur le patient et son état. Ces données sont initialement enregistrées

dans Epic, avant d'être manipulées et réorganisées pour créer la banque de données périopératoires accessibles (PDW – Perioperative Data Warehouse) comportant toutes les données administratives et médicales des chirurgies s'étant déroulées depuis mars 2013 au sein du réseau de UCLA Health. Le PDW maintient près de 4,000 variables différentes sur chaque chirurgie, allant de l'heure et durée programmées pour la chirurgie jusqu'au médicament antinauséeux reçu à la salle de récupération postopératoire.

4.1.1 Description et validation des données

Les données périopératoires sont structurées à l'aide de plusieurs tables de données. La granularité de chacune varie, mais toutes évoluent autour de trois clés primaires : le numéro d'identification du patient, le numéro d'hospitalisation et le numéro de chirurgie.

Certaines données peuvent être calculées et enregistrées automatiquement (âge du patient à partir de la date de naissance et signes vitaux durant la chirurgie par exemple) alors que d'autres sont saisies manuellement. Ces dernières peuvent provenir des processus essentiels au déroulement opératoire, par exemple l'heure de fin de la chirurgie enregistrée lorsque l'infirmière change le statut du patient en salle d'opération, ou être indépendant de ces processus, par exemple lorsque l'anesthésiste coche une consommation excessive d'alcool dans les comorbidités du patient.

Définition des issues cliniques du modèle

L'objectif principal de ce projet étant de classifier les patients selon leur risque d'évoluer vers une trajectoire compliquée de soins postopératoires, la première étape est d'en uniformiser la définition. En se basant sur les métriques retrouvées dans la littérature médicale, un groupe d'experts a établi un consensus autour de la définition utilisée dans le cadre de ce projet, laquelle comprend 5 issues cliniques défavorables :

1) Décès intra-hospitalier;
2) Décès à 30 jours;
3) Réopération;
4) Admission aux soins intensifs;
5) Hospitalisation postopératoire prolongée.

Tout au long de ce mémoire, le terme « trajectoire de soin » fait référence à ces cinq « issues cliniques », ou variables y à prédire. Les deux premières sont intrinsèquement négatives et mesurées respectivement par la présence d'une date de décès pendant une hospitalisation, ou une

date de décès dans les 30 jours suivants la chirurgie. Une réopération au même site chirurgical témoigne souvent d'une complication postopératoire. Cette variable binaire fut créée spécifiquement pour le projet. Elle vaut « 1 » si le patient subit, dans les 30 jours suivants sa chirurgie, une seconde chirurgie pratiquée par le même département. La définition initiale de « réopération » utilisait le nom de la seconde chirurgie pour assurer qu'elle soit liée à une complication de la première. Toutefois, le nom des procédures est un champ libre au contenu variable. Il devenait donc trop complexe de l'utiliser et nous avons plutôt défini une réopération par la récurrence d'une seconde chirurgie par le même département. Cette variable *y* peut donc théoriquement rater des réopérations pratiquées par un autre département, mais la confirmation manuelle de quelques chirurgies a validé cette stratégie. De son côté, l'issue clinique d'admission aux soins intensifs témoigne typiquement d'un séjour postopératoire compliqué. Cette variable binaire est positive dès que la durée de séjour postopératoire aux soins intensifs est supérieure à 1h. Finalement, l'hospitalisation postopératoire prolongée peut survenir dans le cas de plusieurs évènements négatifs et témoigne donc plus largement d'une trajectoire de soins compliquée.

Afin de standardiser la programmation des algorithmes et la présentation des résultats, la durée d'hospitalisation a également été binarisée. Après avoir analysé la distribution des durées de séjour (voir Figure 4.1), le coude se situant au 90^{e} percentile fut utilisé comme valeur seuil. La variable binaire vaut « 1 » si la durée de séjour est supérieure au 90^{e} percentile, pour les patients ayant spécifiquement subi la chirurgie concernée.

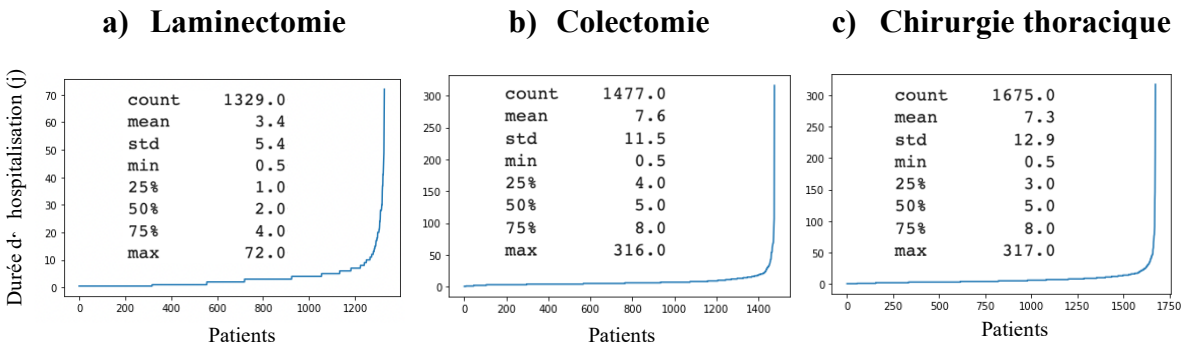


Figure 4.1 Nombre et distribution de patients selon la durée de séjour pour trois chirurgies

Première sélection des variables d'intérêt

Vu la taille de l'entrepôt de données de UCLA, la première étape était de constituer un jeu de données pertinent à la question de recherche. Puisque l'objectif est d'anticiper la trajectoire de soin avant le début de la chirurgie, il était important de rejeter toutes les variables enregistrées durant ou après la chirurgie. Les jeux de données devraient donc théoriquement contenir la même information que celle disponible à l'anesthésiste avant la chirurgie, alors qu'il stratifie le risque du patient. Cette étape a permis de réduire le nombre de variables utiles de 4,000 à 99 (voir annexe A).

En parallèle, trois types de chirurgie furent sélectionnés pour réduire le nombre de lignes analysées : laminectomie, colectomie et chirurgie thoracique. En sélectionnant trois chirurgies parmi les plus pratiquées aux États-Unis et ayant un potentiel de dangerosité variable, nous obtenons un plus grand nombre d'entrées pour entraîner le modèle, assurons la pertinence clinique de nos résultats, et couvrons une gamme de chirurgies hétérogènes pour évaluer les performances des modèles dans différents contextes [59]. Il est à la fois intuitif et démontré que le type de chirurgie pratiquée influence fortement la trajectoire de soins et le type de complications postopératoires [9, 24-26]. Le score POSPOM prédit la mortalité postopératoire par un modèle linéaire et attribue jusqu'à 22 points au type de chirurgie, alors que l'âge ajoute 16 points au-delà de 95 ans et que les comorbidités ajoutent entre 1 et 4 points [9].
Le nombre respectif de lignes retenues pour chaque type de chirurgie est de 2425, 2347 et 2978. Chacune des trois chirurgies fut traitée indépendamment pour l'étape d'analyse des variables ainsi que pour la préparation des données.

Exploration individuelle des variables

Malgré ce premier filtre sélectionnant seulement les variables préopératoires pour 3 chirurgies, une grande hétérogénéité existe toujours dans les types de variable utilisées dont voici quelques exemples:

- variable binaire : la présence d'une comorbidité spécifique, l'utilisation d'anesthésie régionale, le statut urgent de la chirurgie;
- variable quantitative (numérique) : Le nombre de chirurgies subies à ce jour, la durée prévue de la chirurgie en minutes;

- variable qualitative (catégorique) : la classe de patient (hospitalisé, chirurgie d'un jour, admission courte durée), la classe d'urgence (électif, urgent, très urgent, critique);
- variable libre : nom de la procédure effectuée, nom de l'anesthésiste en charge.

Les comorbidités du patient, essentielles à ce projet, présentaient une complexité supplémentaire. Elles pouvaient 1) être cochées par l'anesthésiste au cours de l'évaluation préopératoire et enregistrées sous forme binaire; 2) être enregistrées sous forme binaire dès qu'une preuve de la comorbidité existait au sein du dossier médical électronique; ou 3) être enregistrées sous forme de code ICD10 (et ICD9 avant 2015).

Le premier cas est simple et facile à traiter, mais il implique beaucoup de données manquantes par négligence de cocher la case, impliquant une proportion excessive de « 0 ». Le second cas capte toutes les autres informations du dossier, incluant la case à cocher du premier cas, les codes ICD du patient, l'utilisation de médication liée à la pathologie, les laboratoires diagnostiques et plus. Toutefois, beaucoup de travail a été nécessaire pour créer cet algorithme et donc seules les comorbidités les plus fréquentes ont été programmées (maladie cardiaque, respiratoire, diabète). Il demeurait donc primordial d'extraire de l'information des codes ICD10 et ICD9 pour les autres comorbidités. Ces variables, quasi-continues considérant les 70 000 différents codes ICD10 existants, se retrouvaient dans trois tables différentes à combiner: la table de comorbidités, la table de l'histoire médicale et la table de facturation. Finalement, pour l'ensemble de ces comorbidités, il a fallu restreindre par date d'enregistrement, s'assurant ainsi que la comorbidité était présente au moment de la chirurgie et non qu'elle soit apparue dans les années subséquentes.

L'exploration des variables catégoriques a révélé de l'ambiguïté et du chevauchement entre les classes d'une même variable. Par exemple, la variable PAT_CLASS décrivant la classe de patient comportait le domaine de valeurs {Hospitalisé, Chirurgie d'un jour, Court séjour, Urgence}. Or, une chirurgie peut être à la fois urgente et faite sur un patient hospitalisé. Similairement, la variable BOOKING_CASE_TYPE pouvait prendre les valeurs {Électif, Urgent, Très Urgent, Critique}. Une ambiguïté existe donc entre les trois classes non électives.

Exploration de l'interdépendance des variables

Quelques erreurs de concordance étaient présentes entre les colonnes. Trois colonnes correspondaient au degré d'urgence, mais le taux d'urgence variait significativement de 2.3% (*ASA_EMERGENT*) à 10.1% (*BOOKING_CASE_TYPE*). La première valeur a été abandonnée

puisqu'elle était notée manuellement par l'anesthésiste et donc, susceptible à la négligence, alors que la seconde contribuait à la gestion administrative.

La corrélation de Pearson a été mesurée entre chaque variable d'entrée et issues cliniques mesurées, ainsi qu'entre chaque variable d'entrée du modèle. La figure 4.2 facilite l'identification des variables fortement corrélées, qui seront traitées lors de la préparation des données.

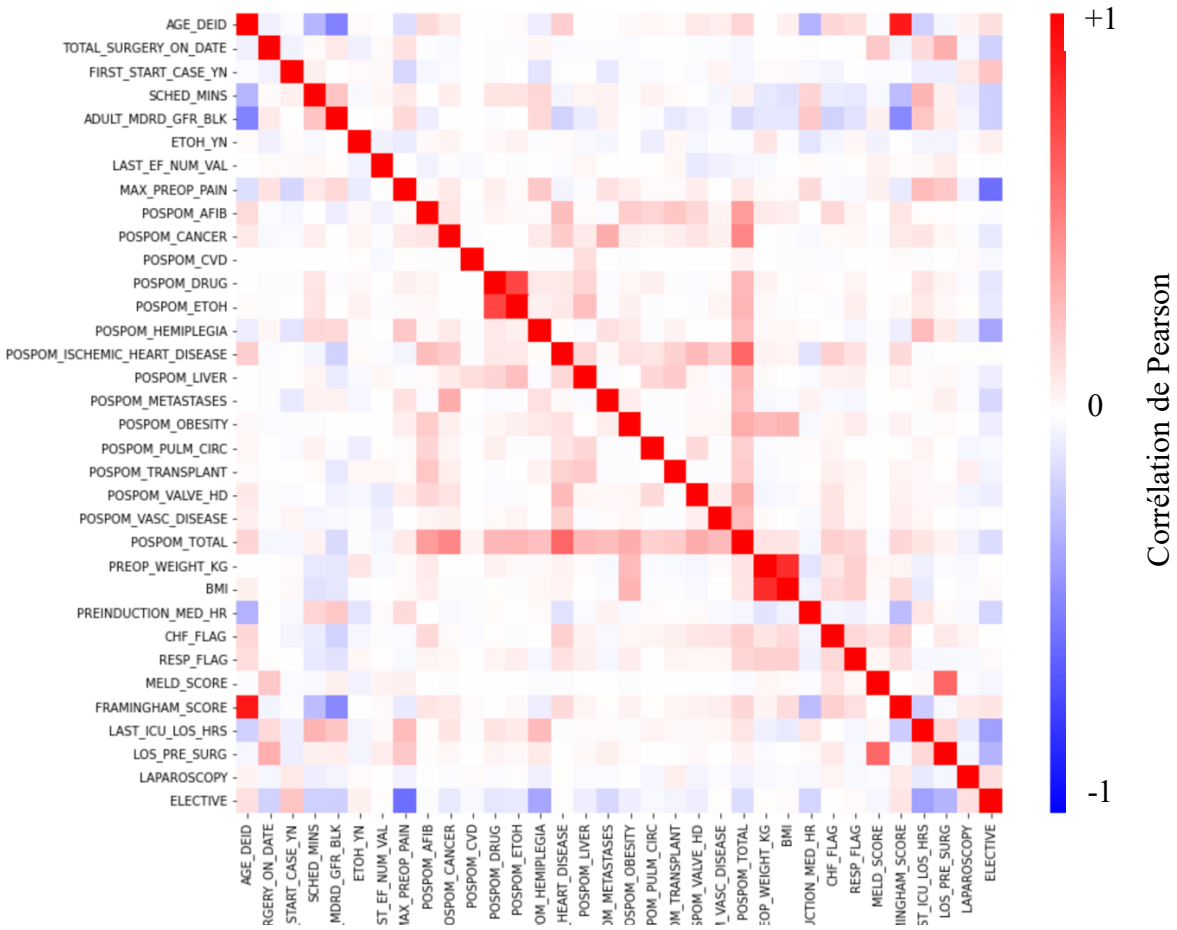


Figure 4.2 Carte de chaleur de corrélation

Représentation graphique des corrélations de Pearson entre les variables explorées du modèle.

4.2 Préparation des données

Sélection des variables

Une première sélection des variables ayant été effectuée préalablement, cette seconde sélection élimine les variables jusqu'à maintenant conservées, mais peu pertinentes pour notre objectif de prédire la trajectoire de soins des patients. Deux stratégies ont été séquentiellement utilisées.

La première exploitait l'expertise clinique et une méthodologie de Delphi modifié [60]. Trois médecins ont indépendamment analysé le jeu de données et évalué la pertinence de chacune des variables proposées. La variable était éliminée si aucun des trois experts cliniques ne la jugeaient pertinente à la trajectoire de soin postopératoire. La Figure 4.3 présente le concept général de l'approche. Chaque observateur (CC, IS, PLL) ne voyait que sa propre colonne, et inscrivait la valeur 1 lorsqu'il jugeait la variable pertinente.

CC	IS	PLL			
1	1	1	3	CL.[ANES_CASE_YN],	Presence of anesthesiologist for case
1	1	1	3	CL.[AGE],	age
1		1	2	CL.[SEX],	sex
1	1	1	3	CL.[PROC_NAME],	Name of procedure
			0	CL.[DATE_OF_SERVICE],	Date of surgery
1			1	CL.[CASE_START],	Hour of start of surgery
		1	1	CL.[GROUPED_PROC_NAME],	Procedure in broader category
		1	1	CL.[SCHED_MINS],	Scheduled OR time

Figure 4.3 Approche Delphi modifié pour la sélection de variables par trois experts

Le score κ de Cohen a été utilisé pour mesurer la concordance inter juge (CC et IS : 0.32; CC et PLL : 0.27; IS et PLL : 0.38). Étant donné l'approche statistique faite dans un second temps, la variable était conservée dans le jeu de données dès qu'un observateur la considérait pertinente. Cette première stratégie a permis de créer le jeu de données final utilisé, comportant au total 77 variables. Le détail des variables utilisées et leur description sont présentés aux annexes A et B.

La seconde stratégie explorait l'analyse des corrélations de Pearson entre les variables et les issues cliniques, telles que mesurées durant l'analyse d'interdépendance des variables. Pour chacune des variables et chacune des issues, la valeur *p,* témoignant de la probabilité que la variable d'entrée exerce une influence sur la variable de sortie, était calculée. Dès que la variable comportait une valeur *p* inférieure au seuil préétabli de 0.05 avec l'une des issues cliniques, la variable était conservée dans le jeu de données. Chaque chirurgie était analysée séparément, créant un jeu de données spécifiques pour chaque chirurgie. Par ces méthodes, deux jeux de données furent créés pour chaque type de chirurgie avec respectivement 34, 36 et 33 variables pour la laminectomie, la colectomie et la chirurgie thoracique. L'annexe A présente les variables utilisées dans chaque modélisation spécifique et l'annexe B présente les variables explorées et non-retenues à cette étape.

À partir de cette étape, chaque chirurgie a une base de données qui lui est propre, avec des variables qui lui sont propres, ce qui explique pourquoi le phénotype est spécifique à chaque chirurgie. Par exemple, il est intuitif de penser que les variables liées aux pathologies respiratoires sont incluses dans la base de données de chirurgie thoracique, alors qu'elles ne sont peut-être pas influentes pour décrire la trajectoire postopératoire de la laminectomie.

Sélection des lignes

En accord avec l'approche standard dans la littérature, seule la première chirurgie d'un patient fut retenue dans la base de données. Un patient ayant une complication chirurgicale est plus à risque de réopération et d'avoir une trajectoire de soins compliquée. En ne retenant que la première opération, nous nous assurons qu'aucune contamination du jeu de test ne survienne, et qu'un unique patient ne compte pas pour plusieurs décès ou plusieurs admissions en soins intensifs, provoquant ainsi une hypercaptation des caractéristiques de ce patient dans l'entraînement du modèle. Le Tableau 4.1 décrit le nombre de chirurgies retirées :

Tableau 4.1 Nombre de chirurgies retirées par type de chirurgie

	Nb total de chirurgies avant le retrait	Nb de chirurgie faites sur d'anciens patients	Nb de chirurgie maximale sur un seul patient
Laminectomie	2425	97	3
Colectomie	2347	102	3
Chirurgie thoracique	2978	300	5

4.2.1 Ingénieries de variables

Une première action a transformé les variables catégoriques en variables indicatrices binaires avec n-1 nouvelles variables, où n est le nombre de valeurs différentes que peut prendre la variable catégorique initiale. Si une ambiguïté ou un chevauchement existait au sein des classes possibles, l'expert était consulté sur la meilleure approche à prendre. En exemple, la valeur BOOKING_CASE_TYPE mentionnée ci-haut fut binarisée à « ELECTIVE », réunissant ainsi tous les degrés d'urgence en un seul.

Différentes approches furent explorées pour optimiser la gestion des comorbidités et des codes ICD10. L'approche initiale était de compter l'occurrence des codes ICD10 et ICD9 dans la population, et ensuite créer une nouvelle variable binaire pour chacune des pathologies les plus fréquemment obtenues. Un expert du domaine révisait ensuite la liste pour confirmer la présence des pathologies les plus pertinentes pour prédire la trajectoire de soins des patients. Cette approche introduisait un biais dans les groupes de comorbidités que nous devions créer. En exemple, le code I42 réfère aux cardiomyopathies. Tous les codes I42 peuvent être réunis sous ce titre, ou ils peuvent être subdivisés en I42.1 à I42.9 selon le type de cardiomyopathie. Certaines classifications regroupent I42.1 à 142.3 avec I42.6 et I42.9. Ces choix auraient dû être faits pour toutes les pathologies incluses dans notre modèle. L'approche finalement utilisée fut d'utiliser la classification du score POSPOM, prédisant la mortalité postopératoire. Ce score, ayant fait l'objet de plusieurs publications et ayant été validé, présente une liste de 27 comorbidités associées à une liste de codes ICD10 et, dès qu'un de ces codes ICD10 est retrouvé au dossier, la variable binaire POSPOM vaut 1, confirmant la présence de la comorbidité associée. De nouvelles variables furent créées basées sur la recommandation des experts. L'approche chirurgicale par laparoscopie est moins invasive qu'une grande incision, nommée laparotomie. La variable binaire « LAPAROSCOPY » fut donc créée selon la présence de « LAPAROSC* » dans la technique chirurgicale. Des exemples de manipulation de variables sont présentés dans le Tableau 4.2, incluant la création des deux variables *y* POSTOP_LOS et REOP. Considérant les incertitudes de mesure, la durée d'hospitalisation postopératoire est sujette à être négative si les dates de chirurgie et de départ sont trop près l'une de l'autre. La valeur minimale a dont été établie à 0.5 jour.

Tableau 4.2 Exemple de variables crées ou transformées

Variable créée	**Type**	**Variables utilisées**	**Condition**
LAPAROSCOPY	Binaire	PROC_NAME	Présence de « LAPAROSC » dans PROC_NAME
POSTOP_LOS	Quantitative	SURGERY_DATE DISCHARGE_DATE	DISCHARGE_DATE – SURGERY_DATE
PREOP_LOS	Quantitative	SURGERY_DATE ADMISISON_DATE	SURGERY_DATE – DISCHARGE_DATE
REOP	Binaire	SURGERY_DATE CAS_SRV_NAME	Présence d'une nouvelle chirurgie, avec différence de <30 jours dans SURGERY_DATE et le même CASE_SRV_NAME

4.2.2 Nettoyage des données

Tel que décrit au chapitre précédent, le nettoyage débute avec le repérage des valeurs aberrantes de la base de données, provenant d'une des trois sources d'erreur précitées. Deux approches furent testées, la première s'appuyant sur, tel que retrouvé dans la littérature, des balises mises en place par les experts pour chacune des valeurs continues [6]. Une valeur enregistrée était alors considérée manquante si extérieure aux balises définies. La seconde approche testée et abandonnée fut d'automatiquement considérer comme aberrantes toutes valeurs à plus de 3, 4, ou 5 écarts-types. Dans un contexte de médecine, la majorité des patients en santé ont des valeurs mesurées dans un petit intervalle alors que les valeurs anormales, encore plus intéressantes pour nous, se retrouvent parfois très éloignées. Avec une très grande proportion des valeurs dans un petit intervalle, l'écart-type était donc typiquement petit. Les valeurs physiologiquement plausibles, dans un contexte de maladie, étaient alors considérées manquantes par cette seconde approche exploitant les écart-types. Dans ce contexte, la première approche fut ainsi favorisée.

Le pourcentage de valeurs manquantes fut ensuite calculé pour chaque variable. Au-delà de 40%, celle-ci était présentée à l'expert clinique pour en confirmer le retrait. Malgré le traitement indépendant de chaque base de données, les mêmes quatre variables présentaient un grand nombre de données manquantes. Trois des quatre variables furent gardées et considérées « physiologiquement normales », en accord avec l'approche clinique standard. Par exemple, la fraction d'éjection du ventricule gauche (FEVG) est mesurée uniquement chez quelques patients présentant des pathologies cardiaques. Si la FEVG est diminuée, le risque augmente significativement, et si elle n'est pas mesurée, elle est très probablement normale puisque le patient n'est pas connu pour des pathologies cardiaques.

Tableau 4.3 Gestion des variables avec données manquantes élevées.

Variable	**Description**	**Conservée?**	**Valeur par défaut**
ERAS_PATIENT_YN	Patient enrôlé dans un programme de récupération rapide	Non	-
MELD_SCORE	Score stratifiant le risque d'une pathologie hépatique	Oui	1 (aucun risque)
LAST_ICU_LOS_HRS	Durée en heure du dernier séjour en soins intensifs	Oui	0 (aucun séjour)
LAS_EF_NUM_VAL	Dernière fraction d'éjection du ventricule gauche enregistrée	Oui	60 (valeur normale)

Tous les patients n'ayant pas de date de sortie de l'hôpital, donc encore hospitalisé au moment de l'extraction des données, ont également été retirés de la liste puisque les issues cliniques de durée d'hospitalisation, de séjour aux soins intensifs et de mortalité intrahospitalière n'étaient pas disponibles. À travers toutes les chirurgies, 22 patients ont ainsi été retirés.

Également, tous les patients n'ayant pas de score ASA enregistré ont été retirés de la banque de données étant donné l'importance de cette variable pour valider la pertinence de notre modèle. Le fait d'imputer une valeur lorsqu'elle est manquante aurait rendu cette variable moins performante au moment de comparer la performance de notre modèle avec cette même variable, correspondant à l'objectif 2 de cette recherche.

Une fois ces lignes éliminées, les valeurs manquantes encore présentes furent imputées par une estimation basée sur les autres caractéristiques du patient. Le module *IterativeImputer* de la librairie *SciKit Learn*, pour l'apprentissage machine dans le langage de programmation Python, a été utilisé à cet effet. Étant donné l'interdépendance des paramètres de santé d'un individu, il est plus précis d'utiliser cette méthode d'imputation plutôt qu'une valeur comme la moyenne. Cette approche est également celle favorisée dans la littérature médicale [4, 15].

4.2.3 Gestion de la corrélation entre les variables

La corrélation entre les variables ayant été explorée individuellement pour chaque chirurgie, les variables fortement corrélées variaient à travers les populations chirurgicales. L'approche préconisée fut de progressivement retirer l'une des deux variables fortement corrélées entre elles, avec le seuil de corrélation établi à 0.75. Deux experts du domaine s'entendaient sur la variable à retirer.

4.2.4 Séparation des données : dérivation et test

Chacune des trois bases de données spécifiques fut séparée en jeu de dérivation, utilisé pour créer le modèle, et jeu de test, mis de côté pour l'analyse de performance finale. Afin de reproduire la capacité du modèle à prédire prospectivement les trajectoires de soin des patients, la séparation en jeux de dérivation et de test a été faite basée sur un critère temporel.

Pour assurer l'anonymat des patients, seule l'année était incluse dans la base de données. Il était donc impossible de cibler la date exacte à laquelle faire la division temporelle pour maintenir un

pourcentage fixe dans les jeux de test. Plutôt, le nombre de patients opérés par année a été compté et la division temporelle a été faite de manière à avoir entre 30 et 40% des patients dans le jeu de test. Le tableau 4.4 décrit cette division pour chacune des chirurgies. Tel que décrit dans ce tableau, le nombre d'évènements postopératoires a également été compté pour assurer un nombre suffisant d'évènements à la fois dans le jeu de dérivation et celui de test.

Tableau 4.4 Distribution des issues cliniques entre les jeux de données de dérivation et de test

Chirurgie	**Année de séparation**	**Patient retenu en dérivation**	**Mortalité hospitalière (Dér/Val)**	**Mortalité à 30J (Dér/Val)**	**Réopération à 30J (Dér/Val)**	**Durée (h) en USI si admis (Dér/Val)**	**Durée (j) de séjour post-op (Dér/Val)**
Laminectomie	2020	1619 (69.5%)	8/7	7/6	54/29	17.9/30.1	3.4/3.6
Colectomie	2019	1477 (65.8%)	14/7	11/4	58/18	23.0/16.3	7.7/6.0
Chirurgie thoracique	2020	1675 (62.5%)	42/20	35/21	93/53	76.9/72.0	7.4/7.2

4.2.5 Standardisation des données

Pour les variables d'entrée, la standardisation des valeurs continues a été faite avec une recalibration des variables continues en cote Z avec l'algorithme *StandardScaler* de *SciKit Learn*, de manière à obtenir une valeur moyenne à 0 et un écart-type de 1. L'algorithme *MinMax* fut également testé, mais résultait en des résultats moins intéressants. Étant donné la présence de valeurs éloignées, tel que présenté avec l'exemple de la fraction d'éjection du ventricule gauche (FEVG), la majorité des variables se faisaient écraser dans un intervalle restreint. Peu de différence existait donc entre une FEVG anormale à 45% et une valeur normale à 60%. Aucun lissage ne fut fait.

4.3 Modélisation

Le premier objectif de cette recherche est de créer des phénotypes au sein de la population hétérogène préopératoire, en émettant l'hypothèse que les patients d'un même groupe partageront des trajectoires de soins similaires. En accord avec la littérature, la première étape sera de tester des modèles de segmentation non-supervisés optimisant les métriques de performance du modèle.

Le meilleur modèle sera retenu pour créer les phénotypes, qui seront ensuite évalués quant à leur performance prédictive sur les issues postopératoires.

En accord avec les objectifs 2) et 3) de ce projet, la capacité prédictive de ces phénotypes sur la trajectoire de soins sera comparée à celle du score ASA, ainsi qu'à celles de modèles supervisés entraînés à partir de la même base de données.

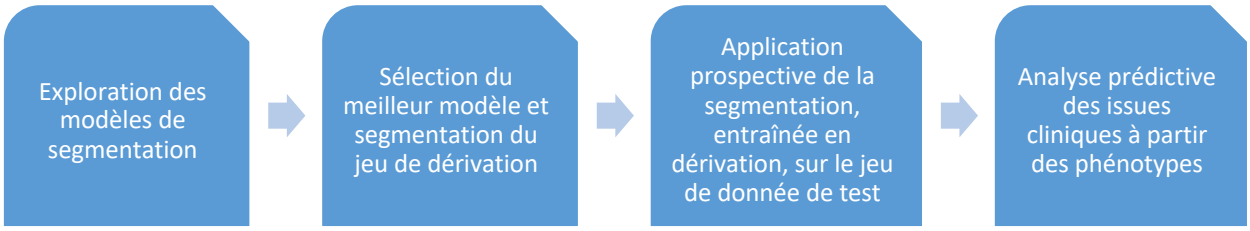


Figure 4.4 Résumé des étapes du phénotypage de cette recherche

4.3.1 Modèle non-supervisé : Phénotyper les patients

Métriques d'évaluation

Les performances du modèle seront évaluées en deux temps. La première analyse évaluera la capacité du modèle à créer des phénotypes à la fois statistiquement et cliniquement intéressants. Les experts interpréteront les caractéristiques cliniques des sous-groupes nouvellement formés et en parallèle, les métriques d'évaluation non-supervisée, décrites au chapitre 3, seront utilisées. La composante clinique sera d'abord explorée via les caractéristiques fréquemment présentes dans chacun des phénotypes, et ensuite via l'analyse descriptive des trajectoires de soins compliquées associées à chaque sous-groupe. La courbe ROC sera utilisée pour évaluer la performance du modèle et la comparer au score ASA actuellement utilisé en pratique clinique.

Données de dérivation et de test

L'entraînement de ce modèle se fera à partir du jeu de dérivation complet. Les modèles de segmentation, le nombre de sous-groupes et autres hyperparamètres, seront établis à partir de celui-ci. Une fois créé le meilleur modèle de phénotypage, les patients du groupe test se verront attribuer le phénotype le plus près de leurs caractéristiques. Seuls les caractéristiques et résultats de l'ensemble test seront présentés.

Modèles de segmentation et hyperparamètres

Quatre types de modèle de segmentation seront explorés dans ce travail. Tous exploitent les librairies de SciKit Learn.

Densité

Les algorithmes de densité réunissent entre eux les points basés sur leur distance. Un point est choisi au hasard et la distance des points avoisinants est calculée. Contrairement aux autres algorithmes décrits, un algorithme de densité peut considérer des points comme non-attribués à un sous-ensemble, et donc comme une valeur éloignée.

`Eps` *:* Correspond à la distance maximale permise pour inclure dans un sous-groupe un point n'étant préalablement pas dans un sous-groupe. Si aucun point autour d'un sous-ensemble n'est à une distance inférieure à l'eps, un nouveau sous-ensemble est trouvé.
`Min_samples` *:* Correspond au nombre minimal de points nécessaires pour considérer le groupe de points comme un sous-ensemble. Si le nombre de points est inférieur à cette valeur, tous les points seront considérés comme éloignés.

Hierarchique

Les hyperparamètres principaux entourent la manière de calculer la distance entre les points et les ensembles :

`N_clusters` **:** Nombre de sous-groupes
`Linkage` **:** Définit la manière de sélectionner les points entre lesquels la distance sera calculée. Les exemples incluent 1) les voisins les plus proches, soit la plus petite distance existante entre deux points de deux ensembles; 2) les voisins les plus éloignés, soit la plus grande distance existante entre deux points; et 3) la moyenne, soit le centre de gravité de chacun des ensembles basé sur tous les points qu'il comporte.
`Affinity` **:** Définit la manière de calculer la distance (ex : euclidienne) entre les deux points choisis dans l'hyperparamètre « Method ».

K-means

Dans le cadre d'un K-means, les hyperparamètres à spécifier sont :

`N_clusters` **:** Nombre de sous-groupes

`n_init` = Détermine le nombre d'itérations que le modèle fera avec une initialisation différente. Les résultats seront la meilleure sortie de n_init itérations consécutives en terme d'inertie.

`init` **:** Détermine la stratégie de sélection des points au moment de l'initialisation de l'algorithme. Ils peuvent être sélectionnés aléatoirement ou sélectionnés pour augmenter la vitesse de convergence.

Consensus K-means

Ce modèle utilise l'algorithme K-Means et donc tous ses hyperparamètres, auxquels il ajoute également les 2 suivants :

`N_repeats` **:** Correspond au nombre de fois que le k-means est entraîné dans le jeu de données, et donc le nombre de fois qu'un sous-groupe est attribué à chaque patient. Il s'agit donc également de la taille de la matrice d'affinité.

`Samp_frac` **:** Correspond au pourcentage des patients utilisés à chacune des itérations définies dans `N_repeats.`

Nombre de phénotypes et paramètres optimaux

Pour modifier les hyperparamètres optimaux de nos modèles et rechercher ceux optimisant nos métriques d'évaluation, deux approches sont utilisées. La recherche par grille implique de créer un ensemble de valeurs pour chaque hyperparamètre. Toutes les possibilités de l'ensemble seront successivement utilisées, nous permettant de repérer la meilleure combinaison d'hyperparamètres. À l'inverse, la recherche aléatoire implique de fournir une distribution de probabilités pouvant être prise pour chaque hyperparamètre. Après avoir spécifié un certain nombre d'itérations, des hyperparamètres sont aléatoirement tirés de cette distribution et le modèle est testé. Dans les deux types de recherche, le ou les modèles offrant les meilleures performances sont conservés.

Un phénotype est considéré acceptable s'il contient au minimum 1% des patients de la population.

Attribution prospective des phénotypes

Les modèles étant entraînés sur le jeu de données de dérivation, il faut ensuite attribuer prospectivement les patients du jeu de test à leurs sous-groupes, ou phénotype. La complexité de cette tâche est variable selon le modèle utilisé. Cette étape doit également pouvoir s'appliquer à de futurs patients, et non seulement à ceux du jeu de test. Deux approches ont été explorées dans le cadre de cette recherche. La première, plus facile mais non universelle, est d'utiliser la fonction « predict » de la librairie SciKit Learn, associée au modèle de segmentation utilisé. Certains modèles n'offrent pas cette option. Il incombe alors d'entraîner un modèle supervisé sur le jeu de dérivation ou d'utiliser la stratégie du plus proche voisin.

La régression linéaire, la forêt aléatoire et les réseaux neuronaux peuvent être entraînés sur une fraction de l'ensemble de dérivation : appelée ensemble d'entraînement, avant d'être testée sur l'autre fraction : appelée ensemble de validation. Considérant le fort débalancement des issues mesurées dans ce projet, la forêt aléatoire permet d'à la fois entraîner et valider sur l'ensemble du jeu de dérivation complet, en utilisant l'approche « out-of-bag ». Chaque fois qu'un arbre est entraîné sur une portion des variables et des patients, l'arbre est testé sur les autres patients non inclus dans l'entraînement.

La stratégie du plus proche voisin est un algorithme glouton qui attribuera aux nouveaux patients le même phénotype que le patient le plus proche. Une fois de plus, le modèle peut être validé au sein d'un sous-groupe de patients de l'ensemble de dérivation.

Le meilleur modèle peut ensuite être utilisé pour attribuer prospectivement un phénotype aux patients de l'ensemble de test.

Validation du modèle

Une fois les modèles de segmentation et d'attribution prospective entrainés, tous les patients de l'ensemble de test furent segmentés à leur phénotype, permettant d'analyser les trajectoires de soins et l'occurrence de complications dans chacun de ces phénotypes.

À partir des phénotypes et des issues cliniques, les métriques d'évaluation décrites préalablement ont également pu être utilisées, notamment la courbe ROC et le score F1.

4.3.2 Modèle supervisé prédictif

La littérature confirme déjà que les données périopératoires peuvent être analysées par l'apprentissage machine pour prédire la survenue de complications telles celles utilisées dans le cadre de ce projet. L'avantage du phénotypage est de permettre d'identifier des trajectoires de soins et des complications à partir d'un seul modèle, facilement exploitable en clinique. Afin de comparer la quantité maximale de signal extractible de la base de données utilisée *versus* la quantité de signal extraite par les phénotypes, il demeurait important d'explorer quelques modèles supervisés.

Dans le contexte de phénotypage, les patients étaient segmentés selon leurs caractéristiques et l'incidence des complications était mesurée pour chacun des phénotypes. Dans ce modèle, plutôt que de dépendre de l'hypothèse que des sous-groupes de patients présenteront des trajectoires de soins similaires, les variables d'entrée du modèle sont directement utilisées pour prédire si le patient présentera une trajectoire de soins compliquée. Des modèles indépendants furent entraînés pour chacune des issues cliniques, ainsi que pour la survenue d'au moins une des issues cliniques compliquées.

Métrique d'évaluation

Ces modèles étant également des classificateurs, les métriques de courbe ROC, de précision, de rappel, et de F1, ont été utilisées pour comparer les performances du modèle. Les données de dérivation serviront à l'entraînement et donc, les métriques d'évaluation seront uniquement appliquées à l'ensemble de validation.

Entrainement et validation

Tout comme dans le contexte du modèle de phénotypage, l'ensemble de dérivation de chaque chirurgie est utilisé pour entrainer un modèle spécifique à cette chirurgie. L'approche par replis avec 5 sous-groupes décrite au chapitre 3 a été utilisé pour explorer les hyperparamètres et sélectionner le meilleur modèle, tout en maintenant une généralisabilité. Chaque jeu d'hyperparamètres étaient donc entraînés 5 fois sur 80% des données, et testé sur le 5^{e} sous-groupe réservé pour la validation. La performance du modèle avec ces hyperparamètres correspond ainsi à la moyenne des 5 performances. Lorsque la meilleure combinaison de modèle et d'hyperparamètres est identifié, le modèle est entraîné sur le jeu de dérivation complet avant d'être utilisé dans l'ensemble test pour donner la performance finale du modèle dans un jeu vierge.

Modèles de segmentation et hyperparamètres

Tout comme pour les modèles de segmentation, l'ensemble des modèles testés et décrits dans cette section proviennent de la librairie SciKit Learn.

Régression logistique et modèle linéaire

La régression linéaire comporte principalement deux hyperparamètres à optimiser :

`C_val` **:** Correspond à un coefficient appliqué sur la pénalité, définissant son importance dans la fonction de perte. Une valeur élevée correspond à un faible poids.

`Solver` **:** Type de solver utilisé pour trouver les coefficients

`Penalty:` Correspond au type de pénalité appliquée. Peut prendre la valeur none pour appliquer aucune pénalité, la valeur `L1` pour faire un modèle *ridge*, la valeur `L2` pour faire un modèle *lasso*, ou prendre la valeur `elastic net` pour combiner les deux pénalités.

Modèle de forêt aléatoire

Les modèles de forêt aléatoire comportent les hyperparamètres suivants à optimiser :

`N_estimators` : Correspond au nombre d'arbres séquentiellement créés dans la forêt.

`Max_depth` **:** Correspond à la profondeur maximale que peut atteindre chaque arbre, soit n divisions + 1.

`Min_samples_leaf` **:** Correspond au nombre minimal de patients requis dans chaque feuille-enfant pour permettre la division

`Max_features` : Correspond au nombre de variables, ou colonnes, pouvant être utilisées par l'arbre avant la séparation d'une feuille interne.

Réseau de neurones

Les réseaux de neurones explorés sont des perceptrons multicouches comportant les hyperparamètres suivants :

`Hidden_layer_size` **:** Présenté sous formes de tuples, le nombre de valeurs dans le tuple définit le nombre de couches, et la valeur détermine le nombre de neurones dans chaque couche.

`Activation` **:** Corrrespond à la fonction d'activation appliquée à la sortie des neurones

`Learning_rate` : Correspond au taux d'apprentissage du modèle à chaque itération.

`Max_iter` : Correspond au nombre maximal d'itération ou d'époques. Si le `Learning_rate` utilisé est bas, ce paramètre devra prendre une valeur plus élevée pour permettre la convergence.
`Alpha` : Correspond à l'importance de la pénalité L2 telle que définie au chapitre 2

CHAPITRE 5 RÉSULTATS ET DISCUSSION

Les résultats sont présentés en deux sections. La première section détaillera tous les résultats ayant mené à la création des phénotypes ainsi que les performances prédictives de ces phénotypes sur la trajectoire de soins postopératoire. Plus spécifiquement, les métriques d'apprentissage non-supervisé seront présentées pour chaque modèle de segmentation et expliqueront le choix du *consensus k-means* avec trois sous-groupes comme meilleur modèle. L'algorithme d'attribution prospective sera ensuite présenté, permettant d'attribuer un phénotype à chacun des patients de l'ensemble de test.

À partir des phénotypes attribués, la performance prédictive des phénotypes sera décrite et comparée au score ASA puisqu'il demeure le score le plus utilisé par l'anesthésiste. Le fait de parvenir à remplacer ou améliorer ce score demeure donc la meilleure manière de créer un modèle implantable en milieu clinique. Pour cette même raison, la distribution des phénotypes à travers chaque score ASA sera analysée. Finalement, les caractéristiques de chaque phénotype seront explorées afin de mieux comprendre le profil de patient attribué à chaque phénotype. Cette dernière étape est cruciale si les résultats sont présentés au clinicien, puisqu'elle éclaircit la « boîte noire » souvent associée aux algorithmes d'apprentissage machine et rebutant les cliniciens souhaitant une l'interprétabilité des résultats.

La deuxième section présente la création des trois architectures de modèles supervisés pour chacune des trois chirurgies et chacune des issues postopératoires. La stratégie de sélection des hyperparamètres sera détaillée ainsi que la stratégie de sélection des meilleurs modèles. Finalement, les performances de ces meilleurs modèles au sein de l'ensemble test seront présentées.

Étant donné la quantité de résultats et le nombre de modèles, ce chapitre se concentrera sur les résultats significatifs et présentera uniquement les résultats de l'ensemble de test. Plus de détails pourront être retrouvés dans les annexes.

5.1 Phénotypage des patients

Phénotyper prospectivement les patients chirurgicaux est le premier objectif de ce projet. À cet effet, plusieurs expérimentations ont été menées pour identifier la meilleure stratégie de segmentation.

5.1.1 Exploration des algorithmes de segmentation

Au total, cinq types de modèles furent explorés. Pour chaque modèle et chaque ensemble d'hyperparamètres, les résultats de segmentation des patients étaient enregistrés, le nombre de patients par sous-groupe était mesuré et les métriques de performance du modèle étaient calculées. Les résultats étaient filtrés selon leur acceptabilité. Pour être considérée acceptable, la segmentation devait avoir un minimum de deux sous-groupes, et chacun des sous-groupes devaient compter au minimum 1% de la population. Le tableau 5.1 détaille le pourcentage de modèle avec résultats acceptables, selon le type de modèle et la chirurgie.

Tableau 5.1 Pourcentage de résultats acceptables selon le type de modèle et la chirurgie

	Laminectomie	**Colectomie**	**Chir. Thoracique**
DB scan	69.25%	60%	51.5%
Hiérarchique descendant	20%	0%	10%
K-means	63.8%	8.91%	41.1%
Consensus k-means	86.5%	93%	89%

La définition d'acceptabilité ici utilisée n'est pas décrite dans la littérature et provient d'une discussion clinique. Puisque que 12% des patients sont considérés à haut risque et expliquent la majorité des complications périopératoires significatives, il est impératif qu'un nombre minimal de patients fasse partie de chaque phénotype pour maintenir la pertinence clinique[3]. Malgré le faible pourcentage d'acceptabilité établi à 1%, le tableau 5.1 démontre que les résultats du DB Scan, des algorithmes hiérarchiques et du *k-means* génèrent fréquemment des résultats non-acceptables, alors que le *consensus k-means* est beaucoup plus constant, probablement en raison du consensus qui compense les sous-groupes créés par des points éloignés.

Un mauvais choix d'hyperparamètres aurait pu provoquer ces résultats mais ils furent explorés rigoureusement. Dans le cas du DB Scan, l'hyperparamètre de distance *eps* fut à la fois exploré aléatoirement et basé sur la distance des plus proches voisins. En créant un graphique présentant la distance moyenne des *k* plus proches voisins de chaque point, et en les affichant en ordre croissant, nous pouvons identifier le « coude » et explorer les valeurs d'hyperparamètre à proximité (Figure 5.1). Le DB Scan a également été exploré en considérant tous les points aberrants comme étant un sous-groupe, mais les résultats sont demeurés peu intéressants. Tout comme pour le DB Scan, les hyperparamètres d'algorithmes hiérarchiques ont été explorés, incluant l'affinité et le type de distance, sans générer beaucoup de résultats acceptables (annexe C).

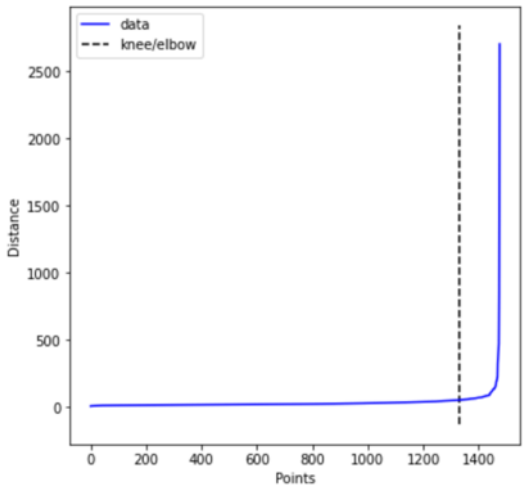
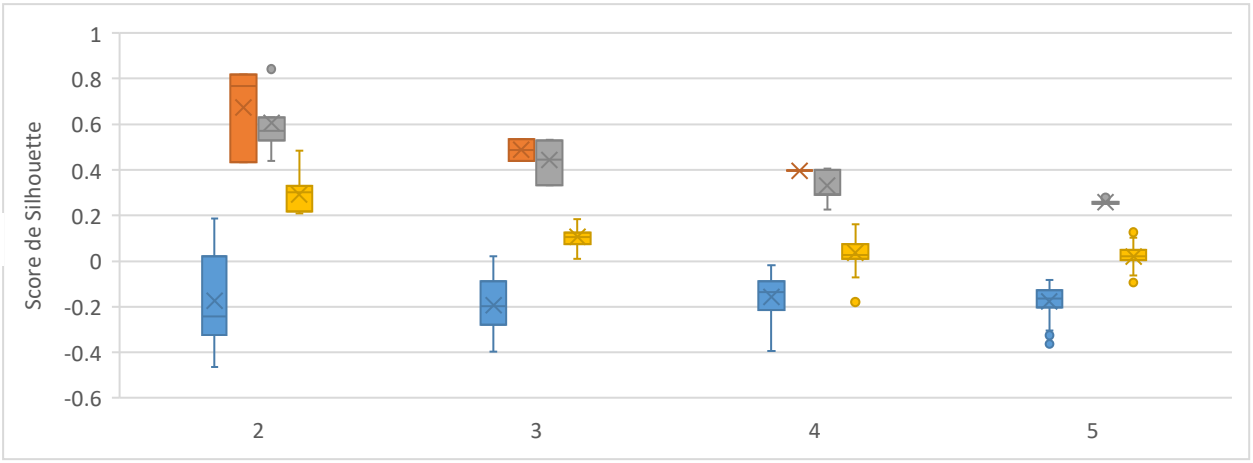


Figure 5.1 Recherche du coude définissant le paramètre *eps* optimal

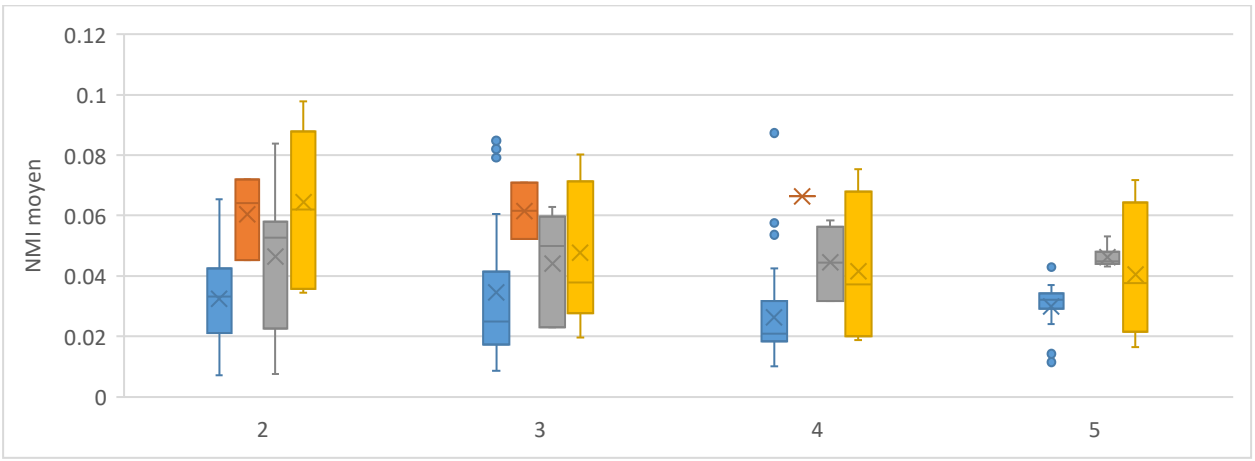
Suite à la filtration des modèles selon leur acceptabilité, la figure 5.2 présente les métriques de segmentation de tous les modèles ayant généré des résultats acceptables. Les trois chirurgies ont été combinées dans cette figure. Considérant que les score NMI et IRA nécessitent une variable de référence pour quantifier la performance de segmentation, cinq scores NMI et IRA ont été calculés pour chacune des issues cliniques et la moyenne des cinq scores a été utilisée pour synthétiser les résultats présentés dans la figure 5.2.

Nous notons d'abord que plus le nombre de sous-groupes augmente, moins grande est la performance. Nous notons avec intérêt que le score de Silhouette est plus élevé pour les modèles hiérarchiques et le *k-means*. Cette observation s'explique par la présence d'un sous-groupe ayant significativement plus de patients, et un nombre variable de petits groupes similaires. Le score de Silhouette attribue le même poids à chaque sous-groupe et donc, les petits sous-groupes tirent la valeur vers le haut. Les score NMI et IRA de ces modèles n'ont pas ces mêmes hautes valeurs et donc, ces sous-groupes ne discriminent pas spécifiquement les patients présentant des issues cliniques défavorables malgré le score de Silhouette élevé. Si on analyse spécifiquement le *consensus k-means*, soit le modèle le plus intéressant selon le tableau 5.1, on note une diminution des métriques de 50% en passant de 2 à 3 sous-groupes.

a)



b)



c)

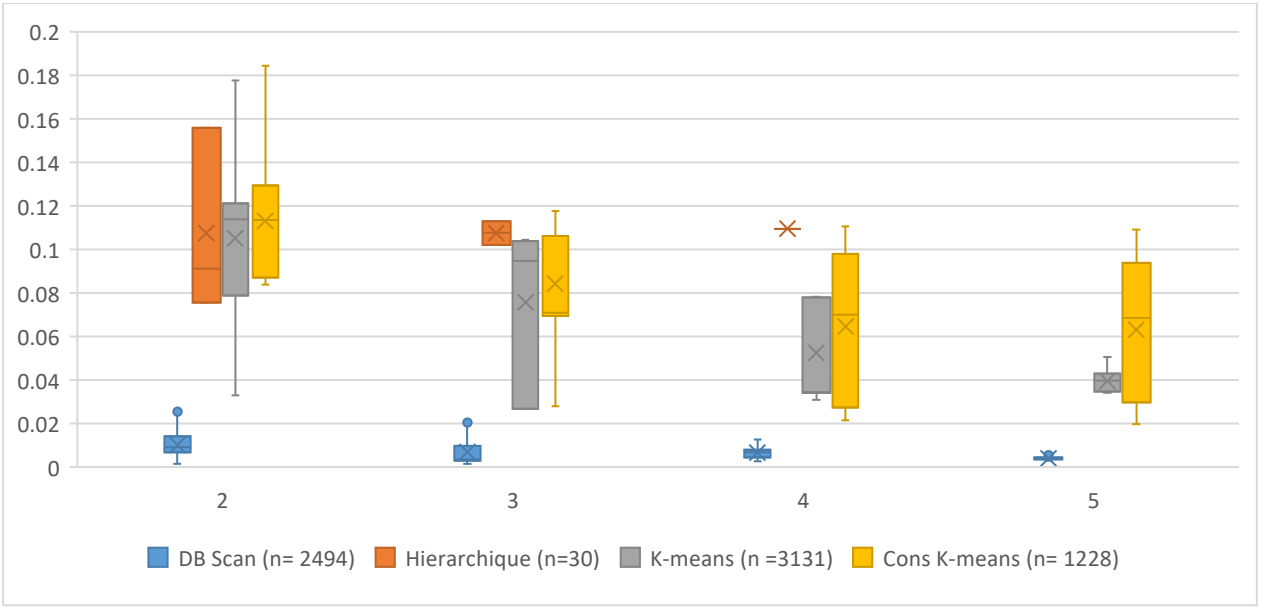


Figure 5.2 Métriques de performance des modèles supervisés - a) Score de Silhouette; b) NMI; c) index de Rand ajusté (IRA)

Au-delà de trois sous-groupes, le score Silhouette s'écrase près de 0, alors que les métriques NMI et IRA maintiennent leur valeur moyenne. En d'autres mots, au-delà de trois sous-groupes, il n'y a pas de frontière claire pour les nouveaux sous-groupes, mais la segmentation basée sur les issues cliniques demeure aussi performante. Dans le cas du *k-means*, l'index Silhouette diminue de 25% à chaque ajout de sous-groupes, le score NMI demeure similaire, et l'IRA moyen diminue de 20%

En combinant le taux de résultats acceptables (Tableau 5.1) et les métriques de segmentation (Figure 5.2), le *consensus k-means* se démarque comme supérieur, principalement dans son aptitude à générer des sous-groupes pertinents. À l'inverse, les autres algorithmes génèrent souvent des petits sous-groupes (<1% de la population). Ce manque de constance, en variant les valeurs aléatoires, rendent le modèle moins intéressant à déployer dans une population test.

Bien que les métriques supportent l'utilisation de deux sous-groupes, une discussion avec les experts cliniques a renforcé l'intérêt de plutôt subdiviser en trois sous-groupes. Cette discussion considérait l'exploration des caractéristiques individuelles des patients dans chaque sous-groupe, qui révélait un profil type intéressant pour chacun des sous-groupes.
Le tableau 5.2 présente les métriques de performance selon le type de normalisation sélectionnée, pour la chirurgie de colectomie. Les autres chirurgies sont présentées à l'annexe D. Pour toutes les métriques et nombres de sous-groupe, la normalisation standard performe davantage que la normalisation *MinMax*, et a donc été retenue pour la modélisation.

Tableau 5.2 Métriques de performance du consensus k-means en chirurgie de colectomie selon deux différentes normalisations

	Normalisation standard				**Normalisation *MinMax***			
Nb phénotypes	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
Silhouette	0.299	0.110	0.015	0.014	0.165	-0.039	-0.111	-0.010
NMI	0.049	0.023	0.015	0.018	0.030	0.015	0.013	0.012
IRA	0.072	0.034	0.010	0.014	0.041	0.008	0.003	0.001

5.1.2 Phénotypage prospectif des patients

La librairie *SciKit Learn*, n'offre pas la fonction de prédiction pour l'algorithme *k-means*, pouvant permettre de segmenter un nouveau point ne faisant initialement pas partie de la population. Un nouveau point serait simplement attribué au centre de gravité le plus près. En utilisant le *consensus k-means*, les centres de gravité varient à chaque itération et cette approche ne peut plus être exploitée. Les stratégies alternatives de segmentation prospective présentées au chapitre 3 ont été utilisées.

Le tableau 5.3 présente les quatre approches explorées dans l'ensemble de dérivation pour faire cette attribution prospective de phénotype spécifique à la chirurgie concernée. Les algorithmes de régression logistique et de perceptrons multicouches ont nécessité de subdiviser l'ensemble de dérivation en ensemble d'entraînement (60%) et ensemble de validation (40%). Le nombre de patients mal classifiés pour ces deux algorithmes provient uniquement de l'ensemble de validation.

Tableau 5.3 Nombre de phénotypes mal classifiés selon le modèle et la chirurgie.

	RF - OOB	RL	MLP	KNN
Laminectomie	0 (0%)	25 (4.7%)	14 (2.6%)	47 (3.5%)
Colectomie	0 (0%)	13 (2.2%)	15 (2.5%)	13 (0.8%)
Chirurgie Thoracique	0 (0%)	13 (1.9%)	12 (1.8%)	32 (1.9%)

RF-OOB : Random Forest, Out-of-bag Score; RL : Régression logistique; MLP : Perceptron multicouches; KNN: K plus proches voisins.

Le score *Out-of-bag* de la forêt aléatoire se démarque significativement des autres approches, avec un résultat même douteux. Bien que cette approche permette d'entraîner le modèle sur tous les patients alors que la régression logistique et le réseau de neurones est entraîné sur 60% des patients, l'obtention d'un résultat parfait pour suggérer un surapprentissage. Considérant l'absence de phénotype dans l'ensemble de test, nous ne pouvons l'utiliser pour confirmer ou infirmer cette hypothèse. Pour explorer ce surapprentissage, la répartition des patients à travers les phénotypes a été analysée.

La figure 5.3 présente le pourcentage des patients attribués à chaque phénotype lorsque l'algorithme de forêt aléatoire est utilisé, selon le type de chirurgie et selon le jeu de données. Nous pouvons donc comparer la distribution entre le jeu de dérivation et le jeu de test. Un bon algorithme d'attribution prospective devrait recréer le même type de distribution dans l'ensemble de test.

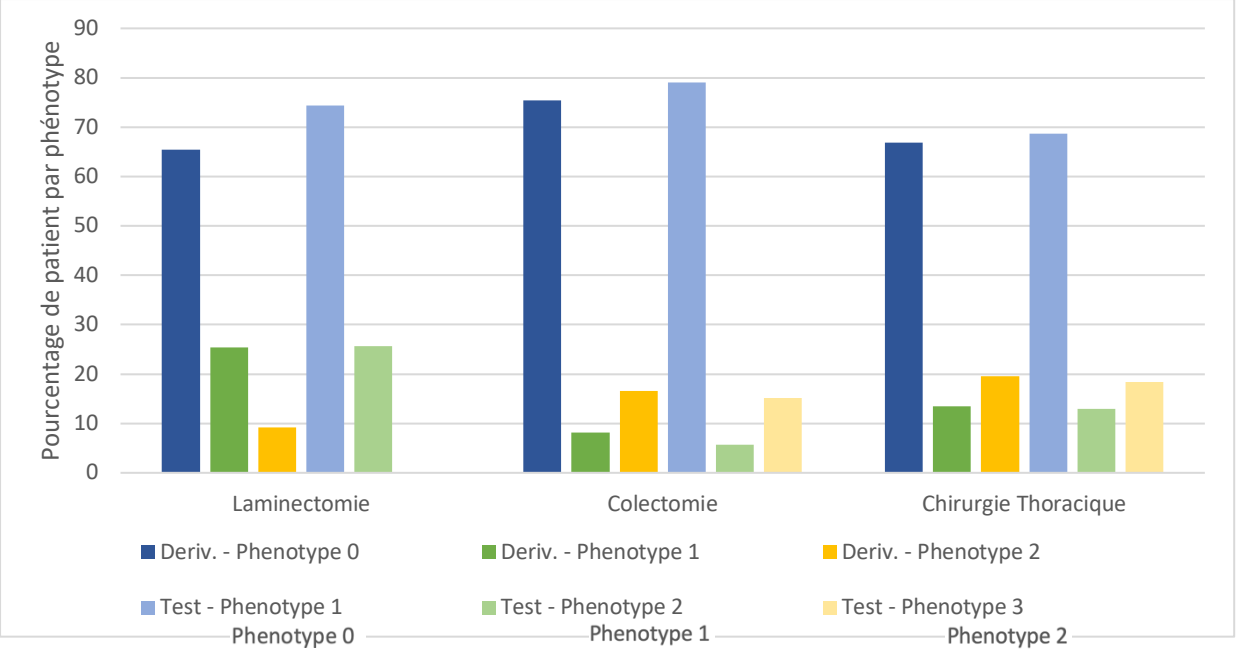


Figure 5.3 Comparaison de la distribution des phénotypes dans les ensembles de dérivation et de validation, selon le type de chirurgie

Nous notons que la répartition des phénotypes dans l'ensemble de test respecte le même patron que celui de l'ensemble de dérivation, supportant ainsi la robustesse de la méthode. L'unique exception à cette validation est le phénotype 2 de la laminectomie, qui n'est attribué à aucun patient dans l'ensemble de test. Les trois autres méthodes (RL, MLP, KNN) du Tableau 5.3 ont mené au même résultat. De façon intéressante, l'année de séparation entre l'ensemble de dérivation et celui de test est de 2020 pour la laminectomie, soit le début de la pandémie mondiale de COVID-19. En effet, l'exploration de l'ensemble test de laminectomie révèle qu'aucun patient n'avait ni même une comorbidité, alors que le phénotype 2 représentait des patients typiquement plus vieux (71 ans *versus* 42 et 67 ans) et plus comorbides (en moyenne 3.5 comorbidités *versus* 0.4 et 0.3). Puisque la laminectomie est fréquemment pratiquée chez des patients avec douleur chronique, la population chirurgicale a significativement changé durant la pandémie et le modèle a capté ce changement.

Basée sur le Tableau 5.3 et la Figure 5.3, l'approche RF-OOB a donc été retenue pour attribuer prospectivement un phénotype aux patients de l'ensemble de test.

5.1.3 Performance du phénotypage

Une fois la segmentation et l'attribution prospective de phénotypes faites pour chaque chirurgie, un quatrième jeu de données a été créé par la combinaison des trois chirurgies. À travers ces quatre jeux de données, le phénotype 0 est celui de bas risque et le phénotype 2 est celui du plus haut risque. Les résultats seront donc parfois présentés combinés pour les trois chirurgies, ou divisés par chirurgie.

La figure 5.4 présente les courbes ROC et scores AUROC du phénotype digital comparativement au score ASA, calculés dans l'ensemble test. Afin de conserver l'information du type de chirurgie dans chaque phénotype, les courbes ROC des trois types de chirurgie ont été obtenues individuellement, puis combinées en une seule courbe par le calcul de la moyenne. Si le jeu de données avec chirurgie combinée avait été utilisé, un phénotype 2 issu d'une laminectomie aurait été considéré identique à un phénotype 2 issu d'une colectomie, perdant ainsi la spécificité de la chirurgie dans le phénotype. L'annexe E présente les performances individuellement pour chaque chirurgie.

En nous basant sur l'AUROC, nous constatons que le phénotype digital préopératoire performe de façon similaire ou supérieure au score de l'ASA pour l'ensemble des 5 issues postopératoires utilisées. Par exemple, l'AUROC du phénotype digital pour l'admission en USI est 0.76, alors qu'elle est de 0.71 avec l'ASA, signifiant que, pour un même nombre de patients identifiés comme ayant la complication, le phénotype digital aura identifié plus de patients subissant réellement la complication que le score ASA. Les AUROC obtenus s'approchent de ceux rapportés dans la littérature lorsque les mêmes issues cliniques sont mesurées, tel que présenté dans la revue de littérature [4, 6, 7, 31].

De façon intéressante, le patron des courbes du phénotype et du score ASA sont significativement différents. Nous observons que le score ASA est supérieur aux extrêmes : un patient ASA 1 aura très souvent une issue favorable, alors qu'un patient ASA 5 compliquera fréquemment. Bien que prédictifs, les scores 1 et 5 sont les moins utilisées en clinique et la majorité des patients obtiennent

le score ASA 3 [20, 32]. À l'inverse, la performance du phénotype digital surpasse le score ASA au plateau de la courbe de l'ASA, correspondant à ce score ASA 3. Le phénotypage devient donc complémentaire en permettant de rediviser ce groupe hétérogène.

Une troisième courbe a donc été créée à partir d'une combinaison linéaire des deux autres scores. L'annexe F présente le calcul de cette combinaison linéaire. Tel qu'attendu, cette courbe obtient la performance du score ASA aux extrémités, et celle du phénotype digital au centre de la courbe. Elles offrent la meilleure AUROC pour toutes les issues cliniques. Des modèles plus raffinés, telles une régression logistique ou forêt d'arbre, ont été explorés, mais le besoin de scinder la population en groupes d'entraînement/validation, en plus du faible nombre de patients et d'évènements postopératoire dans certains groupes (aucun patient n'était ASA 5 et phénotype 0), menaient à des résultats moins intéressants.

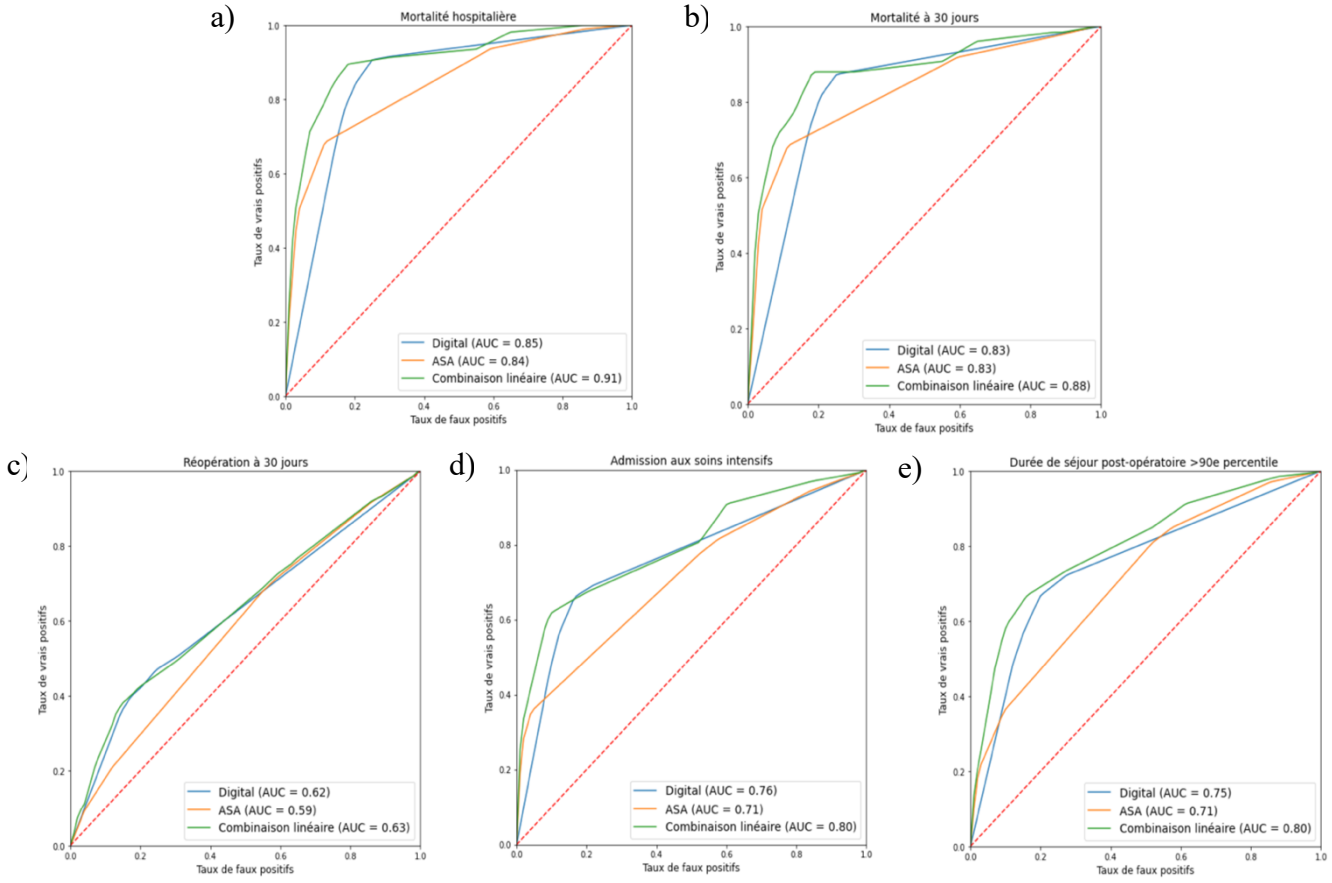


Figure 5.4 Courbes ROC selon l'issue postopératoire mesurée

La figure 5.5 présente l'incidence d'issues postopératoires défavorables selon le phénotype attribué, lorsque les trois chirurgies sont combinées. En d'autres mots, elle offre une représentation graphique d'un tableau décrivant le nombre d'évènements enregistrés au sein de chaque phénotype. Par exemple, nous constatons que 300 évènements se sont produits pour le phénotype 0, dont 60 séjours prolongés au-delà du 90^{e} percentile et 180 admissions en USI. Nous notons qu'environ 300 évènements se sont également produits pour les phénotypes 1 et 2, malgré leur nombre total de patients de 429 et 302, respectivement. À la figure 5.5 b), nous constatons que les phénotypes 0, 1 et 2 sont respectivement associés à une mortalité hospitalière de 0.2%, 2.3% et 7.3%, une réopération de 2.8%, 5.4% et 9.3%, et une admission en USI de 8%, 36.1% et 48%. La surpondération des patients ayant le phénotype 0 est alignée avec l'évolution clinique favorable de la majorité des patients [3].

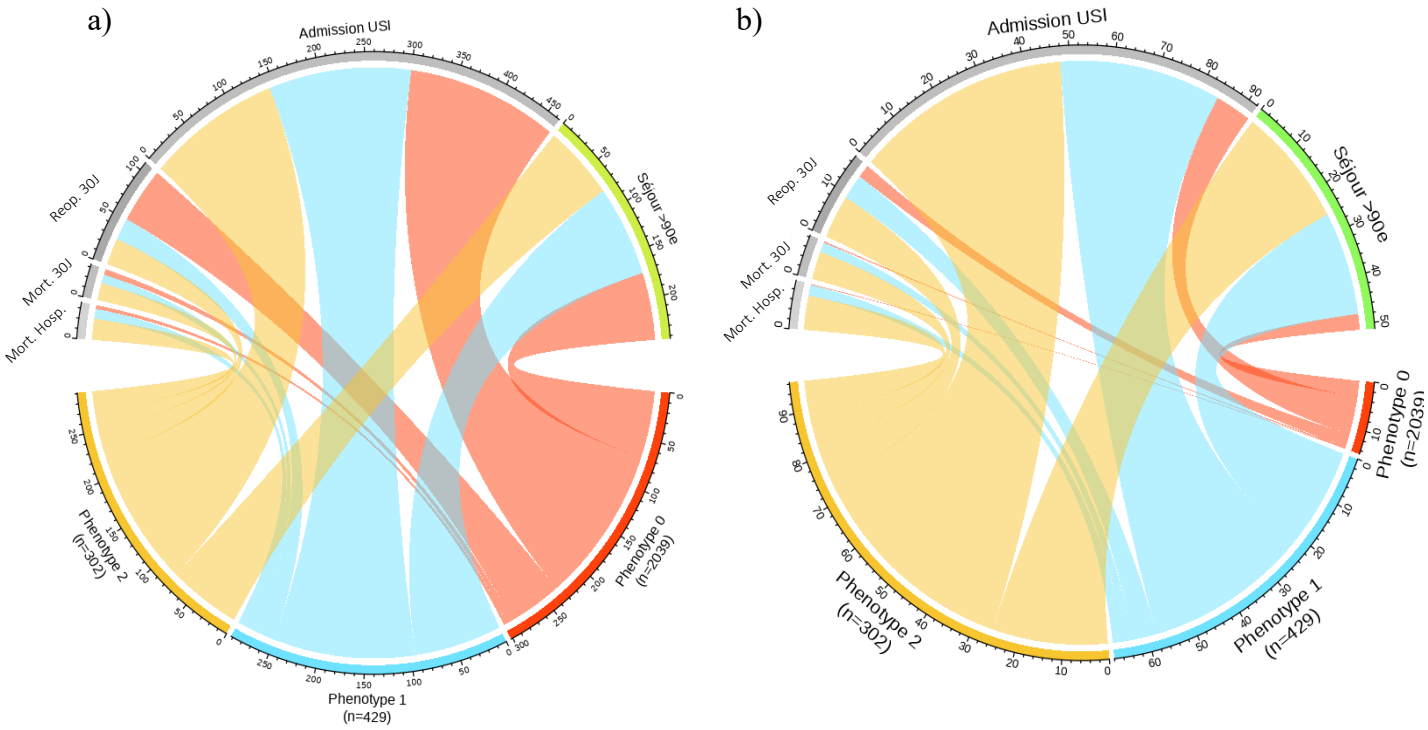


Figure 5.5 Nombres absolus (a) et relatifs (b) de complications postopératoires selon le phénotype

5.1.4 Comparaison des phénotypes avec le score ASA

Considérant la validité établie du score ASA, cette section compare la distribution des patients entre les phénotypes et le score ASA. Le Tableau 5.4 et la Figure 5.6 présentent la même information regroupant les trois chirurgies, sous forme de tableau et de représentation graphique. Nous pouvons constater, à la figure 5.6, que le phénotype 0 et l'ASA 3 sont les plus fréquents de leur score respectif. Nous pouvons également apprécier que la majorité des patients avec phénotype 0 ont un score ASA 2 ou 3, et qu'à l'inverse, les patients ASA 3 représentent la majorité des patients phénotypés à 1 ou 2.

Puisque le phénotypage et le score ASA témoignent tous deux à leur manière du risque chirurgical, la progression synchrone des deux scores est attendue. En effet, nous notons qu'aucun patient ASA 5 n'a obtenu le phénotype 0, alors que 8 des 12 patients ASA 5 ont obtenu le phénotype 2. Ce constat supporte également la validité de la modélisation de phénotypage.

Tableau 5.4 Distribution des phénotypes selon le score ASA

Phénotype \ ASA	**1**	**2**	**3**	**4**	**5**	**Total**
0	20	703	1267	49	0	**2039**
1	12	103	265	45	4	**429**
2	2	49	152	91	8	**302**
Total	**34**	**855**	**1684**	**185**	**12**	**2770**

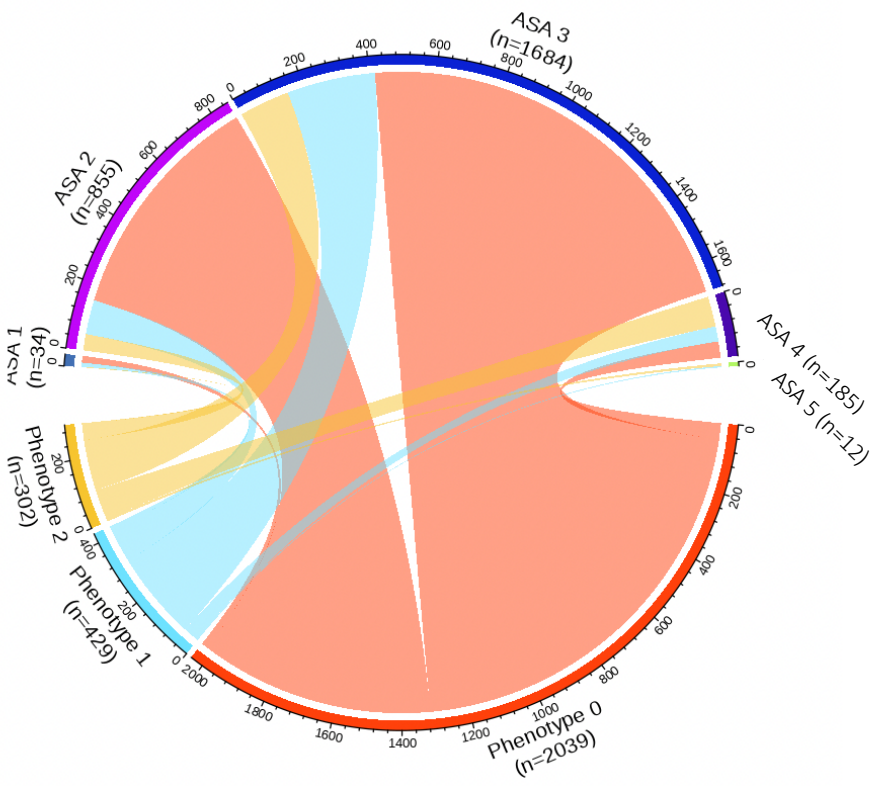


Figure 5.6 Répartition des phénotypes selon le score ASA

5.1.5 Influence des variables explicatives

Deux approches ont été utilisées pour explorer les différences existantes entre les phénotypes de chaque chirurgie. Ces deux approches sont présentées dans la Figure 5.7. La figure de gauche (a) relate les caractéristiques importantes extraites par la forêt aléatoire au moment d'attribuer prospectivement un phénotype aux patients du groupe test. La figure de droite (b) présente l'analyse descriptive des caractéristiques retrouvées au sein du phénotype, pour chaque chirurgie. Le pourcentage d'incidence des variables binaires est présenté, alors que les variables continues ont été rééchelonnées pour faciliter la lecture. L'écart type de chaque variable a été calculé et utilisé de façon décroissante pour présenter les variables sur l'axe des abscisses. Étant donné l'absence d'attribution de phénotype 2 dans le groupe laminectomie, les caractéristiques de ce groupe n'ont pu être extraites.

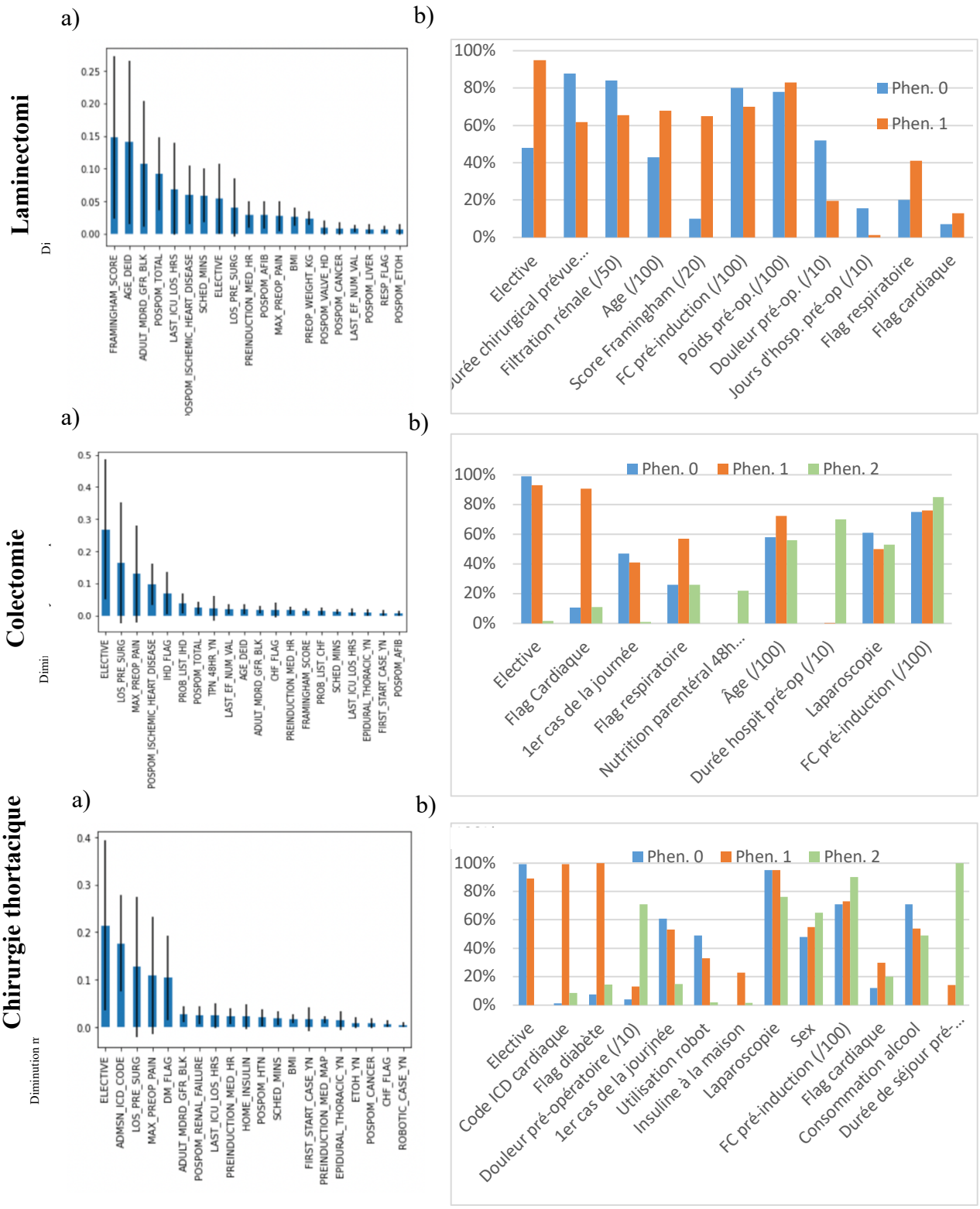


Figure 5.7 Caractéristiques importantes dans l'attribution des phénotypes selon l'algorithme d'attribution prospective (a) et l'analyse descriptive (b)

L'analyse globale montre que la variable « *Elective* » semble la plus discriminante, surtout pour la laminectomie et la chirurgie thoracique. Le phénotype 2 semble entièrement défini par cette variable, à laquelle peuvent être jumelées d'autres caractéristiques typiques de la chirurgie urgente telles la douleur préopératoire, la fréquence cardiaque élevée liée à la douleur et au stress, ainsi que la fréquence moindre de premier cas de la journée puisque réalisée à toute heure. La principale différence entre les phénotypes 0 et 1 semble liée au nombre de comorbidités et l'âge des patients. Le phénotype 1 est typiquement plus vieux et plus malade, ce qui colle une fois de plus à la pratique clinique. Ces caractéristiques semblent particulièrement appropriées pour la colectomie et la chirurgie thoracique, alors que dans le cas de la laminectomie, les différences entre sous-groupes semblent plus variables. Il est intéressant de constater que certaines variables explicatrices changent selon le type de chirurgie, ce qui confirme l'importance de créer des phénotypes spécifiques à chaque population chirurgicale.

Cette double approche des variables explicatives est intéressante puisqu'elle reflète le fonctionnement de l'algorithme sous-jacent. La déviation standard utilise la notion de distance et la différence entre le 0 et 1 d'une variable binaire est plus grande qu'entre le 0.2 et 0.9 d'une variable continue normalisée, par exemple l'âge. À l'inverse, une forêt aléatoire peut diviser une branche à toute valeur d'une variable continue, voire l'utiliser plus d'une fois dans son arbre, alors qu'elle ne peut diviser qu'une fois une variable binaire. Particulièrement avec la colectomie et la chirurgie thoracique, nous pouvons constater la prépondérance des variables continues utilisées par la forêt aléatoire, et la prépondérance de variables binaires lorsque nous décrivons les variables importantes par l'écart-type entre les phénotypes. Les variables reconnues importantes par les deux approches, par exemple « *Elective* » en colectomie, sont probablement parmi les plus influentes. L'exploration de l'attribution prospective par un autre modèle interprétable, telle la régression logistique, aurait pu confirmer par une nouvelle approche. Finalement, analyser les chutes de performances des mêmes modèles avec le retrait successif des variables importantes auraient également permis d'approfondir l'analyse d'interprétabilité.

5.1.6 Exploration des valeurs éloignées

À la lumière des résultats et des caractéristiques explicatives, il est intéressant d'explorer plus spécifiquement les deux patients ayant obtenus un phénotype 2 et donc jugés à haut risque, avec le

score ASA 1 typiquement donné à des patients en bonne santé. Le premier patient était un homme de 36 ans, a eu une thoracotomie d'urgence cédulée pour 5h, avait une douleur à 6/10 et n'a passé qu'une journée postopératoire à l'hôpital en soins intensifs, possiblement dans le contexte d'un transfert entre hôpitaux. Le second est un homme de 26 ans ayant eu une résection de nodule par thoracoscopie après 3 jours d'hospitalisation et avait une douleur préopératoire à 8/10. La notion d'urgence est contradictoire selon deux variables présentes dans la banque de données initiale, mais dans le processus d'élimination, nous avons conservé uniquement celle confirmant l'urgence. Aucune complication n'a eu lieu.

À la lumière de l'analyse, il devient évident que ces patients furent considérés à risque dans le contexte d'urgence et de douleur préopératoire, malgré la classification ASA 1 en l'absence de comorbidités. Cette analyse supporte la complémentarité de l'information donnée par le phénotype.

5.1.7 Résumé des résultats de phénotypage

En bref, la meilleure stratégie de segmentation a été identifiée à partir de l'ensemble d'entraînement. En se basant sur les métriques de performance et une discussion avec les cliniciens, l'algorithme *consensus k-means* avec trois phénotypes a été retenu. La segmentation finale a été faite avec cet algorithme dans l'ensemble d'entraînement, et une forêt aléatoire a été entraînée à prédire le phénotype des patients en se basant sur leurs caractéristiques. Cette forêt aléatoire a été utilisée pour attribuer des phénotypes aux patients de l'ensemble test, sur lesquels les performances ont été mesurées. Basé sur les courbes ROC, le phénotype digital performait légèrement mieux que le score ASA, ce qui signifie que le modèle proposé parvient à extraire, sans apport d'un clinicien et uniquement à partir des données disponibles dans un dossier médical électronique, autant d'information voire plus que l'outil actuellement utilisé en pratique clinique.

La combinaison linéaire des deux scores performait encore mieux. Le phénotype est donc complémentaire au score ASA et permettrait en clinique d'automatiser l'attribution d'un phénotype au patient en attendant l'évaluation par l'anesthésiste. Le phénotype pourrait ainsi être utilisé par d'autres membres du personnel médical lorsqu'ils interagissent avec le patient et, lorsque l'anesthésiste fait son évaluation, il pourrait inscrire le score ASA qui automatiserait l'utilisation de la combinaison linéaire pour améliorer les performances.

Le phénotype 0 est le plus fréquent, est caractérisé par des patients plus jeunes et moins malades, et ils ont le moins de complications post-opératoires. Le phénotype 1 est plus âgé et malade et

présente un risque intermédiaire. Le phénotype 2 est surtout associé à des chirurgies urgentes et il est associé aux trajectoires de soin les plus difficiles. La prochaine section traite de l'entraînement de modèles supervisés à partir de la même base de données.

5.2 Modèles supervisés

Les résultats seront présentés en deux étapes. La première consiste en l'exploration des hyperparamètres pour chaque architecture. La seconde est la sélection des meilleurs modèles, basée sur les métriques de performance de l'ensemble de validation, pour qu'il soit testé dans l'ensemble de test.

5.2.1 Exploration des hyperparamètres

À cette étape, les trois types de modèles : régression logistique, forêt aléatoire et perceptron multicouches, furent explorer en faisant varier les hyperparamètres. Le détail des hyperparamètres explorés est présenté dans l'annexe G. Le jeu de données de dérivation était aléatoirement divisé en 5 groupes. Séquentiellement, les modèles étaient entraînés sur 4 groupes (80% des données) avec un jeu d'hyperparamètres et testés sur le 5^{e} groupe (20%) des données. Puis, l'entraînement était répété avec le même jeu d'hyperparamètres en alternant le groupe utilisé en test, pour un total de 5 itérations par jeu d'hyperparamètres (algorithme *k-fold*). Les résultats de F1 et de AUROC étaient enregistrés à chaque itération et la moyenne de chacun était retenue comme performance finale de ce jeu d'hyperparamètres. Toutes les combinaisons d'hyperparamètres présentés à l'annexe G étaient séquentiellement évaluées et la figure 5.8 présente la distribution des métriques de performance, selon la chirurgie et l'issue opératoire. Nous pouvons donc y observer les 3 chirurgies, les 3 modèles et les 5 issues postopératoires.

La première observation notable est la discordance entre le score F1 et AUROC pour les issues cliniques 1, 2 et 3, soient mortalité hospitalière, mortalité à 30 jours et la réopération. Bien que les modèles de forêt aléatoire obtiennent des valeurs AUROC significativement élevées, le score F1 demeure très faible. Une observation similaire est faite pour la régression logistique et pour le modèle MLP. Cette discordance provient probablement du débalancement extrême des classes, auxquels sont sensibles les modèles, puisque les scores F1 s'améliorent pour l'issue opératoire 3, soit l'admission en USI, et 4, le séjour opératoire prolongé. Globalement, la figure 5.8 démontre

que les meilleurs modèles sont variables selon les issues, mais que le patron de performance de chaque modèle, pour chaque issue, a tendance à être similaire pour chaque chirurgie.

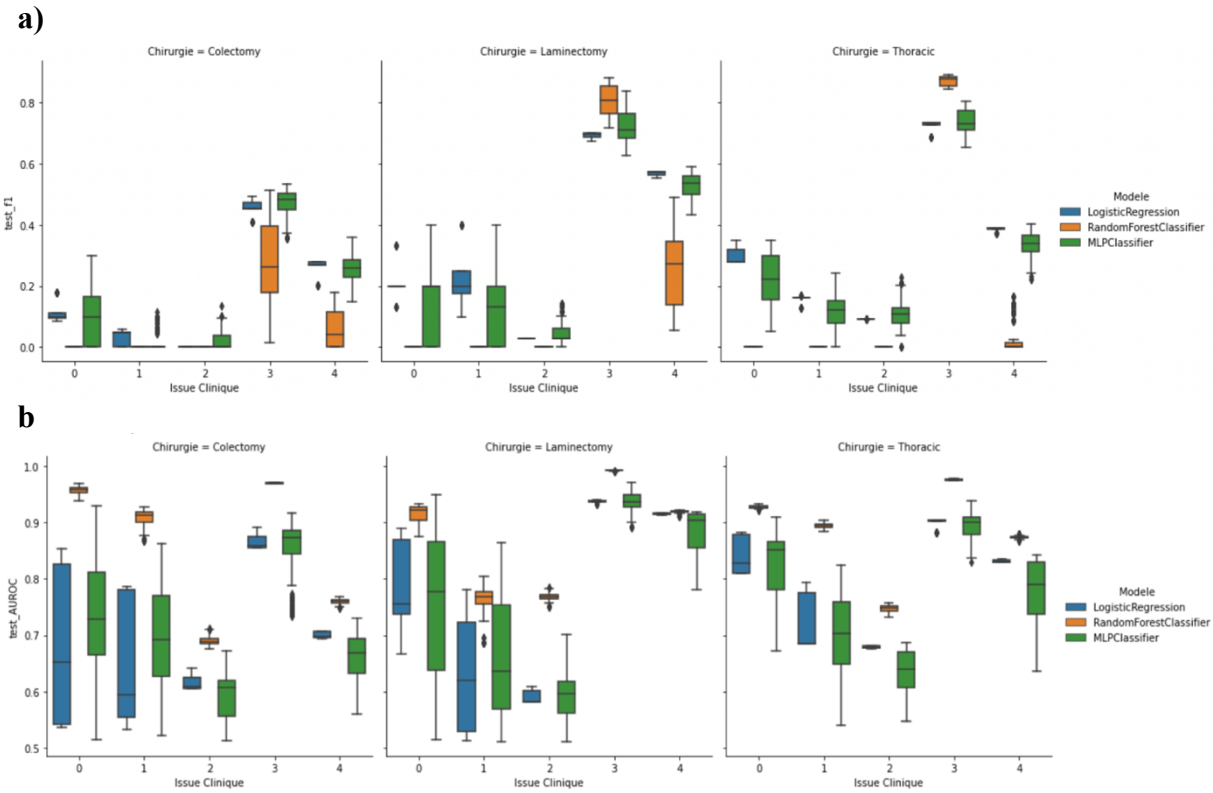


Figure 5.8 Distribution des scores f1 (a) et AUROC (b) par chirurgie, selon le type de modèle et les issues postopératoires

Bien que la figure 5.8 ne présente que les résultats avec une normalisation de données standard, la normalisation *MinMax* fut également explorée. Les performances étaient toutefois inférieures, tel qu'également noté avec le modèle de phénotypage. En effet, la normalisation standard a généré 11 fois les meilleurs résultats, sur un total de 15 (3 chirurgies et 5 issues cliniques). La normalisation standard a donc été retenue.

5.2.2 Meilleurs modèles supervisés

À partir des résultats obtenus dans l'exploration des hyperparamètres, la meilleure combinaison de modèle et de jeu d'hyperparamètres a été sélectionnée, pour chaque chirurgie et chaque issue clinique. Les modèles étaient sélectionnés en se basant sur le meilleur score F1. Le modèle était

ensuite entraîné avec la totalité du jeu de dérivation, et était testé dans le jeu de données test, pour obtenir le score F1 et l'AUROC final du modèle.

Les métriques de performance des meilleurs modèles sont présentées dans le Tableau 5.7 par chirurgie et par issue postopératoire. Nous notons que globalement, les modèles performent bien en entraînement et en validation, mais que les performances diminuent drastiquement dans l'ensemble de test. La mortalité hospitalière passe d'un score F1de 0.3-0.4 à 0.03-0.06, soit une diminution de facteur 10. Cette observation est présente pour toutes les chirurgies, et toutes les issues cliniques. Le débalancement des classes est la raison probable de cette discordance, sachant que tous ces algorithmes y sont sensibles. Une stratégie de balancement aurait pu être explorée avec du sur- et sous-échantillonnage, tel que présenté au chapitre 6, qui discutera de la méthode. Il est à noter qu'une attention particulière a été appliquée dans la révision de l'implémentation pour toute erreur de programmation. Ces résultats sont à la fois concordants et discordants avec ceux obtenus dans la littérature. En effet, nous notons que la performance des meilleurs modèles avoisine une AUROC de 0.90 dans la prédiction de la mortalité hospitalière, ce qui correspond aux valeurs publiées par les auteurs étudiant cette même issue clinique [4, 6, 31]. Ces publications n'étudiaient pas la performance du modèle dans un groupe test, il est donc impossible de comparer la perte de performance obtenue dans ces données. Nous notons également que 11 des 15 meilleurs modèles utilisent un algorithme MLP, suggérant la non-linéarité des variables.

Tableau 5.5 Performance des meilleurs modèles, en entraînement, validation et test

Chirurgie	Issue mesurée	Modèle	F1 train	AUROC train	F1 valid	AUROC valid	F1 test	AUROC test
Laminectomie	Mortalité hosp.	MLP	0.90	0.95	0.40	0.90	0.06	0.74
	Mortalité à 30j	LR	0.79	0.99	0.4	0.70	0.02	0.52
	Réopération	MLP	0.99	0.99	0.20	0.51	0.09	0.59
	Admisson en USI	RF	0.95	0.99	0.88	0.99	0.0	0.50
	Séjour >90^{e} perc	MLP	0.81	0.98	0.59	0.90	0.41	0.65
Colectomie	Mortalité hosp.	MLP	0.91	0.99	0.3	0.91	0.03	0.65
	Mortalité à 30j	MLP	1.0	1.0	0.11	0.56	0.0	0.5
	Réopération	MLP	0.97	0.99	0.13	0.61	0.04	0.5
	Admisson en USI	MLP	1.0	1.0	0.53	0.89	0.17	0.67
	Séjour >90^{e} perc	MLP	0.975	1.0	0.36	0.70	0.20	0.64
Chirurgie thoracique	Mortalité hosp.	LR	0.48	0.96	0.35	0.81	0. 04	0.51
	Mortalité à 30j	MLP	0.96	0.99	0.24	0.73	0.06	0.62
	Réopération	MLP	0.99	1.0	0.22	0.64	0.09	0.50
	Admisson en USI	RF	0.89	0.98	0.95	0.99	0.10	0.52
	Séjour >90^{e} perc	MLP	0.41	0.89	0.40	0.83	0.25	0.65

Le seul modèle permettant l'explicabilité des variables et ayant atteint la valeur F1 de 0.1 est la forêt aléatoire prédisant l'admission en USI pour la chirurgie thoracique. La figure 5.9 utilise la fonction intégrée de *Scikit Learn* pour extraire les variables explicatrices, et nous y constatons que la majorité du signal provient d'une variable spécifique, soit la dernière durée d'hospitalisation en soins intensifs répertoriée dans le DME. Il est intéressant de constater que cette variable n'est pas

ciblée comme explicatrice au moment de phénotyper les patients, confirmant que l'approche par phénotypage permet de capter un signal différent de celui du modèle supervisé.

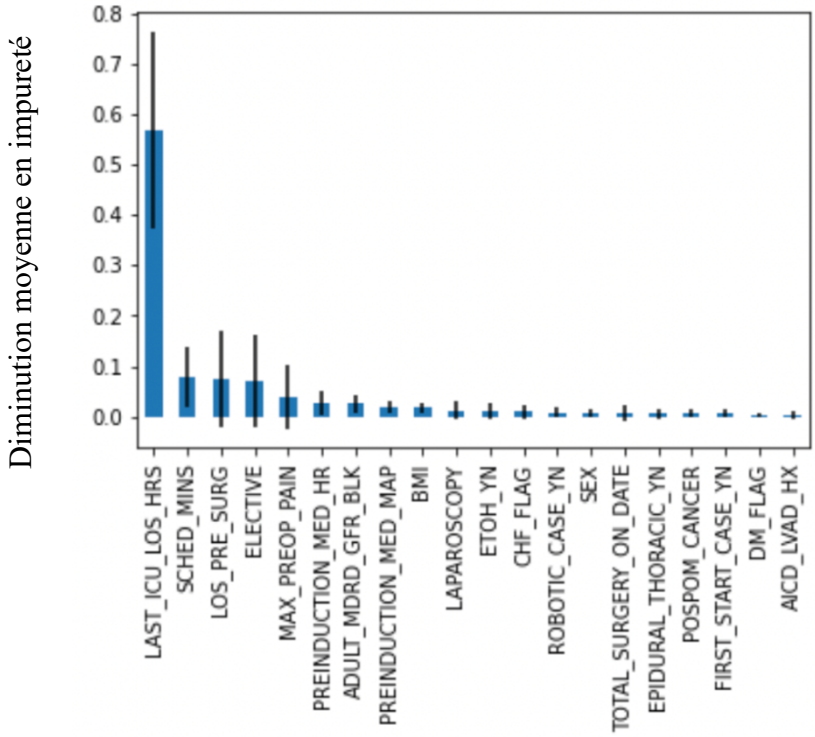


Figure 5.9 Variables explicatrices de la forêt aléatoire classifiant l'admission en soins intensifs pour la chirurgie thoracique

À la lumière de ces résultats, nous avons montré que l'approche de phénotypage par *consensus k-means*, suivie de l'attribution prospective par forêt aléatoire, permettait de créer des sous-groupes partageant des caractéristiques clés, telle l'urgence de la chirurgie, et permettrait de prédire l'évolution post-opératoire des patients. Ces stratégies avaient une performance similaire au score ASA et la combinaison linéaire de ces deux scores offrait une performance supérieure. Les modèles supervisés entraînés ont permis d'obtenir des bonnes performances en validation, mais leur performance déclinait significativement en test. Le débalancement des classes explique probablement cette constatation.

CHAPITRE 6 DISCUSSION SUR LA MÉTHODE

Lors de la réalisation de ce projet, la méthodologie a été rigoureuse, mais des biais et faiblesses sont inévitables. Cette section abordera successivement les étapes de la méthodologie et les hypothèses associées, alors que le tableau 6.1 résume les forces et faiblesses.

Tableau 6.1 Résumé des forces et faiblesse méthodologiques

Forces méthodologiques
• Source et ampleur de la base de données • Inclusion de trois chirurgies fréquemment performées ayant un risque opératoire variable • Utilisation du Delphi modifié par un groupe d'experts pour la sélection des variables • Validation temporelle de la modélisation • Comparaison rigoureuse entre les résultats de phénotypage et la stratification de risque par score ASA • Transférabilité de la méthode sur toutes les chirurgies fréquemment performées
Faiblesses et biais méthodologiques
• Absence des données externes au réseau informatique de UCLA • Subjectivité des variables postopératoires utilisées pour définir une trajectoire de soin compliquée • Faible nombre d'évènements postopératoires • Sensibilité variable et sur-représentation de certaines comorbidités (selon les variables) • Surapprentissage des modèles supervisés

6.1 Choix de la base de données

Un travail significatif fut mené pour sélectionner la banque de données utilisées dans le cadre de ce projet. Bien que UCLA offre l'accès aux chercheurs à la banque de données périopératoires anonymisées DDR, plusieurs variables ne sont disponibles que dans la base de données

nominatives PDW. En exemple, la relation temporelle de plusieurs variables est perdue dans DDR, alors qu'il importait de savoir si un patient présentait une comorbidité au moment de la chirurgie, ou si elle s'était plutôt développée après l'opération. Une collaboration étroite avec l'équipe de chercheurs fut nécessaire pour extraire et anonymiser les données requises pour ce projet. Malgré cela, certaines données sont inévitablement incomplètes, par exemple, si le patient a eu des soins en dehors du réseau de UCLA.

6.2 Sélection des variables utilisées

La sélection des chirurgies et des variables avec la méthode Delphi modifiée fut une étape cruciale dans le contexte de l'utilisation d'un modèle non-supervisé pour créer les phénotypes. Cette première sélection par des experts du domaine assurait d'éviter d'accumuler des variables avec peu d'intérêt clinique sur la trajectoire, pouvant ainsi dégrader l'intérêt des cliniciens envers les résultats de ce travail. La combinaison d'une sélection clinique et statistique est optimale.

6.3 Choix des issues postopératoires

La trajectoire de soin demeure un terme imprécis, même pour les cliniciens. La gamme des complications pouvant être utilisées pour décrire cette trajectoire est immense, allant de simples nausées postopératoires jusqu'au décès. Dans le cadre de ce projet, les issues cliniques fréquemment rapportées dans la littérature médicale furent utilisées, incluant la mortalité hospitalière et à 30 jours ainsi que la durée de séjour en hôpital. La binarisation de la durée de séjour avait pour objectif d'unifier la modélisation et la présentation des résultats, et le 90^{e} percentile représentait le coude auquel la durée d'hospitalisation augmentait rapidement. Étant donné le peu de patients admis en USI dans le contexte des chirurgies étudiées, binariser par l'admission en USI était plus intéressant. Finalement, la variable « Réopération » fut créée dans le contexte d'une discussion clinique considérant la forte corrélation entre une réopération et des complications chirurgicales, en plus de son influence sur le retour à la maison du patient. Ce groupe de cinq issues cliniques, couvrant largement la trajectoire de soins, était approprié pour le premier projet de phénotypage pré-chirurgical mais aurait pu être défini spécifiquement pour chaque chirurgie en ciblant les complications, par exemple la fuite anastomotique d'une colectomie.

6.4 Équilibre des classes

En sélectionnant des issues cliniques graves et moins fréquentes, le débalancement des classes était inévitable. Le faible nombre d'évènement fut exacerbé par la séparation des données en ensemble de dérivation et test, et selon les modèles, parfois également en ensemble d'entraînement et de validation. Ce débalancement fut moins notable au moment de créer les phénotypes, mais est devenu plus significatif à l'étude des modèles supervisés, contribuant à la forte différence entre la performance des modèles supervisés entre l'ensemble de validation et de test. Des stratégies de sur- et sous-échantillonnage auraient pu être explorées. Toutefois, l'objectif de ce projet était d'analyser la performance du modèle de phénotypage, et de la comparer au score ASA et à des modèles supervisés entraînés sur la même base de données. Si des approches de balancement de classe avait été appliquée dans le contexte des modèles supervisés, la même modification aurait dû être appliqué à la base de données avant de segmenter par *consensus k-means*, altérant probablement les résultats présentés. Finalement, étant donné le très faible nombre d'évènements comme la mortalité, le risque de sur-apprentissage aurait été présent si les patients avaient été dupliqués quelques fois. Néanmoins, des stratégies d'équilibre des classes pourraient être explorées dans un futur projet.

6.5 Manipulation des variables

La manipulation des variables influence fortement les résultats de segmentation. Alors qu'un algorithme supervisé peut écraser le poids de la variable pour éliminer son effet, l'influence d'une variable peu pertinente, tout comme la combinaison de plusieurs variables corrélées, influencera la distance entre deux points, et donc les résultats de segmentation. Le filtrage séquentiel par l'expert et la corrélation statistique visaient à éliminer ces variables créant une distance sans affecter la trajectoire de soins, alors que l'analyse de corrélation et le retrait des variables corrélées visaient à éviter la surpondération. De façon similaire, l'approche de normalisation utilisée influencera significativement les résultats. Dans le contexte de ce projet, la variable LAST_EF_VAL_NUM était souvent manquante, mais fut conservée vu son importance clinique. La valeur imputée de 60 est celle considérée par défaut par un anesthésiste n'ayant aucune preuve qu'elle est diminuée. La déviation standard était de 5 et la valeur minimale de 10, la valeur normalisée de 10 devient donc 10 fois plus influente qu'une variable binaire ayant une distance maximale de 1. Une normalisation

MinMax provoquait une autre sorte de biais, par exemple en écrasant toutes les valeurs légèrement anormales vers la normalité lorsqu'une valeur très anormale est présente.

6.6 Validation temporelle et attribution prospective de phénotypes.

La séparation de l'ensemble test par l'année de chirurgie est une force significative de ce projet. En entraînant le modèle avec les années 2013 à 2019/2020, la pertinence future de la modélisation était simultanément testée. Heureusement, le nombre d'évènements postopératoires était similaire pour toutes les chirurgies, permettant d'utiliser cette approche (Tableau 4.4).

Toutefois, en absence de fonction préprogrammée pour attribuer prospectivement un phénotype à la suite d'une segmentation *k-means*, une stratégie alternative d'apprentissage supervisé a été appliquée pour bien attribuer un phénotype aux patients du groupe test. La stratégie de forêt aléatoire avec utilisation du score « *Out-of-bag* » s'est démarquée par sa capacité à bien attribué les phénotypes au sein de l'ensemble de dérivation. Considérant qu'aucun patient ne fut mal classifié, il est possible qu'un surapprentissage ait été fait, mais aucun groupe test n'existe pour confirmer cette hypothèse. À l'inverse, le profil similaire de distribution des phénotypes au sein de la population test, ainsi que la progression simultanée des scores de phénotype et score ASA, supportent tous deux l'efficacité de l'algorithme de forêt aléatoire.

Toutefois, considérant qu'un algorithme supervisé est finalement utilisé pour stratifier le risque des patients, la question se pose de savoir si le phénotype n'est qu'un résultat intermédiaire et qu'il serait plus simple de directement faire un apprentissage machine supervisé pour prédire l'évolution des patients. En regardant les variables explicatives de cette forêt d'attribution prospective, ainsi que les forêts prédisant directement les issues cliniques, nous constatons que les variables sont distinctes, notamment avec la dernière durée d'hospitalisation en USI, absente du modèle d'attribution de phénotype alors qu'elle est la plus influente pour prédire directement les issues cliniques. Il est donc clair que le phénotype capte un signal distinct de ces modèles. Ceci est sans aborder l'utilisation clinique plus facile du phénotype, offrant des informations sur plusieurs issues cliniques, comparativement à un modèle supervisé ne prédisant qu'une seule issue préprogrammée.

6.7 Transférabilité et généralisabilité de la méthode

Un des points notables associés à la méthodologie appliquée, et confirmée par les résultats intéressants obtenus pour trois chirurgies très différentes, est la transférabilité de la méthode à

différents types de chirurgie. Étant donné l'éventail de chirurgies fréquemment pratiquées, il est primordial que la méthode de phénotypage puisse être appliquée à toute base de données regroupant suffisamment d'évènements pour un même type de chirurgie.

Également, une fois les phénotypes créés et l'arbre d'attribution prospective créé, l'algorithme peut même être exporté en dehors d'un dossier médical électronique, tel au Québec. Sachant par exemple qu'en chirurgie thoracique, un phénotype 2 correspond à une chirurgie urgente avec douleur préopératoire et quelques autres caractéristiques, il est alors possible de manuellement phénotyper un patient.

Néanmoins, cette dernière approche dépend de la généralisabilité des résultats, soit la validité externe. L'ensemble de cette recherche a été faite au sein des données du réseau de UCLA. Bien qu'une stratégie de validation temporelle ait été utilisée, il demeure importante de confirmer les résultats obtenus dans des jeux de données provenant de d'autres institutions et d'autres pays, et analyser la performance prédictive des phénotypes déjà créés avec la laminectomie, colectomie et chirurgie thoracique.

CHAPITRE 7 CONCLUSION ET RECOMMANDATIONS

Les objectifs de ce projet étaient de développer le premier phénotype digital du patient préchirurgical pour stratifier son risque opératoire, de comparer la performance prédictive de ces phénotypes sur la trajectoire de soins à la performance prédictive du score ASA actuellement utilisé en pratique clinique, ainsi qu'à celle de modèles d'apprentissage machine supervisé entraînés sur les mêmes données. Parvenir à créer ces phénotypes est intéressant à plusieurs égards puisqu'il peut en résulter 1) une automatisation de la stratification de risque préopératoire sans intervention humaine préalable, 2) un consentement plus éclairé par la disponibilité de données personnalisées sur le risque opératoire et 3) puisqu'il s'agit d'une étape vers la médecine préopératoire de précision, dans laquelle les traitements individuels sont personnalisées à des caractéristiques de plus en plus complète de chaque patient. Pour cette recherche, la définition de trajectoire de soins fut divisée en cinq issues cliniques : la mortalité hospitalière, la mortalité à 30 jours, la réopération, l'admission en soins intensifs, et l'hospitalisation postopératoire prolongée.

Afin de créer des phénotypes digitaux, nous avons extrait tous les dossiers médicaux électroniques des patients ayant été opérés au sein du réseau hospitalier de l'Université de Californie à Los Angeles (UCLA) pour l'une des trois chirurgies explorées dans le cadre de ce projet : la laminectomie, la colectomie et la chirurgie thoracique. En appliquant un algorithme de segmentation non supervisé, trois phénotypes spécifiques aux chirurgies ont été identifiés, nommés 0, 1 et 2. Les phénotypes furent entraînés dans une population plus ancienne, et testés dans une population plus jeune afin de reproduire l'utilisation prospective de phénotype sur de nouveaux, ou futurs, patients. Dans l'ensemble de test, les 3 phénotypes avaient respectivement une mortalité hospitalière de 0.2%, 2.3% et 7.3%. La réopération était respectivement de 2.8%, 5.4% et 9.3%, alors que l'admission en soins intensifs était respectivement de 8%, 36.1% et 48%. La capacité prédictive des phénotypes était similaire à celle du score ASA, mais la combinaison du phénotype avec le score ASA attribué par le clinicien est systématiquement supérieure à chaque score individuel. Le phénotype capte donc un signal distinct et complémentaire du score ASA. Bien que les variables explicatives du phénotype varient en fonction de la chirurgie, le phénotype 0 était le plus fréquent (74%) et généralement en bonne santé. Le phénotype 1 (15%) était typiquement plus âgé et avait plus de comorbidités, alors que le phénotype 2 (11%) englobait les patients subissant une chirurgie urgente. Les meilleurs modèles supervisés entraînés pour prédire chacune des issues

postopératoires ont généralement bien performé sur l'ensemble de validation, mais leur performance était significativement moindre dans l'ensemble de test, témoignant ainsi d'un surapprentissage. Le débalancement des classes prédites contribue à ces résultats, et constitue une des limites de cette recherche Davantage d'expérimentations seraient nécessaires pour conclure sur l'efficacité du phénotypage par rapport à celle des modèles supervisés utilisant les mêmes données en entrée.

En somme, ce projet propose le premier phénotypage digital de patients préchirurgicaux comme outils de stratification de risque opératoire. À partir du dossier médical électronique d'un patient, les données du patient sont automatiquement extraites pour permettre à l'algorithme de phénotypage d'attribuer l'un des trois phénotypes spécifiques à la chirurgie. Ce phénotype permet d'anticiper la trajectoire de soin aussi bien que le score ASA actuellement utilisé, mais ne nécessite pas d'expertise clinique. La combinaison du phénotype et du score ASA offre la meilleure performance, supportant l'idée d'allier la compétence clinique du médecin aux algorithmes d'apprentissage machine, pour améliorer la qualité des soins offerts aux patients. Ce travail demeure le premier de son style pour phénotyper des patients chirurgicaux et encadrer rigoureusement la sélection des variables et l'attribution prospective. Néanmoins, plusieurs améliorations possibles peuvent faire l'objet de recherches supplémentaires. Il est démontré que l'utilisation des résultats de laboratoire améliore la performance de modèle supervisé prédictif de complications postopératoires[6]. Il serait également intéressant d'utiliser les variables peropératoires, combinées aux variables préopératoires, pour phénotyper un patient en postopératoire immédiat pour ainsi évaluer si le phénotype postopératoire correspond toujours à celui préopératoire et éventuellement même identifier des évènements intraopératoires influençant le phénotype. Finalement, une autre approche intéressante serait d'utiliser l'apprentissage profond pour mieux segmenter les classes les patients. Un exemple serait de créer un réseau de neurones prédisant différentes issues postopératoires, et d'appliquer l'algorithme de segmentation sur l'une des couches cachées plutôt que sur les variables d'entrée. Cette approche permettrait potentiellement de corriger la distance, ou importance, de variables moins pertinentes à la prédiction.

RÉFÉRENCES

[1] B. Horvath *et al.*, "The Evolution, Current Value, and Future of the American Society of Anesthesiologists Physical Status Classification System," vol. 135, nº. 5, p. 904-919, Nov 1 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/34491303

[2] D. Nepogodiev *et al.*, "Global burden of postoperative death," vol. 393, nº. 10170, p. 401, Feb 2 2019. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/30722955

[3] R. M. Pearse *et al.*, "Identification and characterisation of the high-risk surgical population in the United Kingdom," vol. 10, nº. 3, p. R81, 2006. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/16749940

[4] B. L. Hill *et al.*, "An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data," vol. 123, nº. 6, p. 877-886, Dec 2019. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/31627890

[5] L. Jalilian et M. Cannesson, "Precision medicine in anesthesiology," vol. 58, nº. 4, p. 17-22, Fall 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/32991333

[6] C. K. Lee *et al.*, "Development and Validation of a Deep Neural Network Model for Prediction of Postoperative In-hospital Mortality," vol. 129, nº. 4, p. 649-662, Oct 2018. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/29664888

[7] C. K. Lee *et al.*, "Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality," vol. 4, nº. 1, p. 8, Jan 8 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/33420341

[8] D. I. Sessler *et al.*, "Broadly applicable risk stratification system for predicting duration of hospitalization and mortality," vol. 113, nº. 5, p. 1026-37, Nov 2010. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/20966661

[9] Y. Le Manach *et al.*, "Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and Validation," vol. 124, nº. 3, p. 570-9, Mar 2016. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/26655494

[10] K. Stavem *et al.*, "Charlson comorbidity index derived from chart review or administrative data: agreement and prediction of mortality in intensive care patients," vol. 9, p. 311-320, 2017. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/28652813

[11] S. H. Jain *et al.*, "The digital phenotype," vol. 33, nº. 5, p. 462-3, May 2015. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/25965751

[12] A. Oellrich *et al.*, "The digital revolution in phenotyping," vol. 17, nº. 5, p. 819-30, Sep 2016. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/26420780

[13] G. Data Science Collaborative, "Differences in clinical deterioration among three sub-phenotypes of COVID-19 patients at the time of first positive test: results from a clustering analysis," vol. 47, nº. 1, p. 113-115, Jan 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/33074342

[14] M. D. Wilkerson et D. N. Hayes, "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking," *Bioinformatics,* vol. 26, nº. 12, p. 1572-1573, Jun 15 2010. [En ligne]. Disponible: <Go to ISI>://WOS:000278689000067

[15] C. W. Seymour *et al.*, "Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis," vol. 321, nº. 20, p. 2003-2017, May 28 2019. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/31104070

[16] J. C. Ferreira et C. M. Patino, "Types of outcomes in clinical research," vol. 43, nº. 1, p. 5, Jan-Feb 2017. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/28380183

[17] A. R. Tait, M. K. Teig et T. Voepel-Lewis, "Informed consent for anesthesia: a review of practice and strategies for optimizing the consent process," vol. 61, nº. 9, p. 832-42, Sep 2014. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/24898765

[18] K. S. Braghiroli *et al.*, "Perioperative mortality in older patients: a systematic review with a meta-regression analysis and meta-analysis of observational studies," vol. 69, p. 110160, May 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/33338975

[19] J. R. Reilly *et al.*, "Systematic review of perioperative mortality risk prediction models for adults undergoing inpatient non-cardiac surgery," vol. 91, nº. 5, p. 860-870, May 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/32935458

[20] N. J. Hackett *et al.*, "ASA class is a reliable independent predictor of medical complications and mortality following surgery," vol. 18, p. 184-90, Jun 2015. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/25937154

[21] M. Froehner *et al.*, "Validation of the Preoperative Score to Predict Postoperative Mortality in Patients Undergoing Radical Cystectomy," vol. 5, nº. 2, p. 197-200, Mar 2019. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/28753894

[22] S. Juul *et al.*, "Validation of the preoperative score to predict postoperative mortality (POSPOM) in patients undergoing major emergency abdominal surgery," vol. 47, nº. 6, p. 1721-1727, Dec 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/31161251

[23] W. H. Organization. (2022) International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). [En ligne]. Disponible: https://www.cdc.gov/nchs/icd/icd-10-cm.htm

[24] L. G. Glance *et al.*, "The Surgical Mortality Probability Model: derivation and validation of a simple risk prediction rule for noncardiac surgery," vol. 255, nº. 4, p. 696-702, Apr 2012. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/22418007

[25] K. L. Protopapa *et al.*, "Development and validation of the Surgical Outcome Risk Tool (SORT)," vol. 101, nº. 13, p. 1774-83, Dec 2014. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/25388883

[26] D. Campbell *et al.*, "National risk prediction model for perioperative mortality in non-cardiac surgery," vol. 106, nº. 11, p. 1549-1557, Oct 2019. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/31386174

[27] C. Bortz *et al.*, "The Patient-Reported Outcome Measurement Information System (PROMIS) Better Reflects the Impact of Length of Stay and the Occurrence of Complications Within 90 Days Than Legacy Outcome Measures for Lumbar Degenerative Surgery," vol. 15, nº. 1, p. 82-86, Feb 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/33900960

[28] P. Y. Tseng *et al.*, "Prediction of the development of acute kidney injury following cardiac surgery by machine learning," vol. 24, nº. 1, p. 478, Jul 31 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/32736589

[29] V. V. Misic *et al.*, "Machine Learning Prediction of Postoperative Emergency Department Hospital Readmission," vol. 132, nº. 5, p. 968-980, May 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/32011336

[30] B. Xue *et al.*, "Use of Machine Learning to Develop and Evaluate Models Using Preoperative and Intraoperative Data to Identify Risks of Postoperative Complications,"

vol. 4, nº. 3, p. e212240, Mar 1 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/33783520

[31] I. S. Hofer *et al.*, "Development and validation of a deep neural network model to predict postoperative mortality, acute kidney injury, and reintubation using a single feature set," vol. 3, p. 58, 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/32352036

[32] E. H. D. M. El Amine Lazouni M, Settouti N, Chikh MA, Mahmoudi S. , *Studies in Computational Intelligence. Chapter : Machine learning tool for automatic ASA detection*: Springer International Publishing, 2013.

[33] O. Sobrie *et al.*, "A new decision support model for preanesthetic evaluation," vol. 133, p. 183-193, Sep 2016. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/27393809

[34] L. A. Marsch, "Opportunities and needs in digital phenotyping," vol. 43, nº. 8, p. 1637-1638, Jul 2018. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/29703995

[35] C. Molina et B. Prados-Suarez, "Digital Phenotypes for Personalized Medicine," vol. 285, p. 141-146, Oct 27 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/34734865

[36] N. C. Jacobson, H. Weingarden et S. Wilhelm, "Using Digital Phenotyping to Accurately Detect Depression Severity," vol. 207, nº. 10, p. 893-896, Oct 2019. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/31596769

[37] E. M. Kleiman *et al.*, "Digital phenotyping of suicidal thoughts," vol. 35, nº. 7, p. 601-608, Jul 2018. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/29637663

[38] I. Barnett *et al.*, "Relapse prediction in schizophrenia through digital phenotyping: a pilot study," vol. 43, nº. 8, p. 1660-1666, Jul 2018. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/29511333

[39] J. Benoit *et al.*, "Systematic Review of Digital Phenotyping and Machine Learning in Psychosis Spectrum Illnesses," vol. 28, nº. 5, p. 296-304, Sep/Oct 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/32796192

[40] F. Hatib *et al.*, "Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis," vol. 129, nº. 4, p. 663-674, Oct 2018. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/29894315

[41] C. Canales, C. Lee et M. Cannesson, "Science Without Conscience Is but the Ruin of the Soul: The Ethics of Big Data and Artificial Intelligence in Perioperative Medicine," vol. 130, nº. 5, p. 1234-1243, May 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/32287130

[42] D. M. Bierle *et al.*, "Preoperative Evaluation Before Noncardiac Surgery," vol. 95, nº. 4, p. 807-822, Apr 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/31753535

[43] C. M. Quinn *et al.*, "Creating Individual Surgeon Performance Assessments in a Statewide Hospital Surgical Quality Improvement Collaborative," vol. 227, nº. 3, p. 303-312 e3, Sep 2018. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/29940332

[44] Y. Tsugawa *et al.*, "Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians," vol. 177, nº. 2, p. 206-213, Feb 1 2017. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/27992617

[45] R. Li *et al.*, "Weekday of Surgery Affects Postoperative Complications and Long-Term Survival of Chinese Gastric Cancer Patients after Curative Gastrectomy," vol. 2017, p. 5090534, 2017. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/28484712

[46] D. P. Fudulu *et al.*, "Weekday and outcomes of elective cardiac surgery in the UK: a large retrospective database analysis," vol. 61, nº. 6, p. 1381-1388, May 27 2022. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/35092280

[47] A. Jerath *et al.*, "Socioeconomic Status and Days Alive and Out of Hospital after Major Elective Noncardiac Surgery: A Population-based Cohort Study," vol. 132, nº. 4, p. 713-722, Apr 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/31972656

[48] N. Kaur et S. K. Sood, "Efficient Resource Management System Based on 4Vs of Big Data Streams," *Big Data Research,* vol. 9, p. 98-106, Sep 2017. [En ligne]. Disponible: <Go to ISI>://WOS:000413205200009

[49] I. H. Theodora Wingert, Tristan Grogan, Mlissa McCabe, Eilon Gabel, Richard Shemin, Aman Mahajan, Maxime Cannesson, "Are Patient Risk Factors Consistent Across Data Sources: A Comparison of EMR, Billing, and Clinician-Abstracted Data," 2016.

[50] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, p. 2825-2830, Oct 2011. [En ligne]. Disponible: <Go to ISI>://WOS:000298103200003

[51] S. Learn. Cross-validation: evaluating estimator performance. [En ligne]. Disponible: https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation

[52] L. Hubert et P. Arabie, "Comparing Partitions," *Journal of Classification,* vol. 2, nº. 2-3, p. 193-218, 1985. [En ligne]. Disponible: <Go to ISI>://WOS:A1985AVF9100003

[53] S. Learn. Clustering. [En ligne]. Disponible: https://scikit-learn.org/stable/modules/clustering.html#k-means

[54] V. Cherkassky et Y. Q. Ma, "Comparison of model selection for regression," *Neural Computation,* vol. 15, nº. 7, p. 1691-1714, Jul 2003. [En ligne]. Disponible: <Go to ISI>://WOS:000183421400012

[55] L. Breiman, "Random forests," *Machine Learning,* vol. 45, nº. 1, p. 5-32, Oct 2001. [En ligne]. Disponible: <Go to ISI>://WOS:000170489900001

[56] S. Cavalieri, A. Distefano et O. Mirabella, "Assessment of the Communication in a Multi-Dsp Architecture for a Multilayer Perceptron Simulation," p. 313-316, 1991. [En ligne]. Disponible: <Go to ISI>://WOS:A1991BV14Z00041

[57] Y. Shao *et al.*, "Shedding Light on the Black Box: Explaining Deep Neural Network Prediction of Clinical Outcomes," vol. 45, nº. 1, p. 5, Jan 4 2021. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/33404886

[58] W. Luo *et al.*, "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View," vol. 18, nº. 12, p. e323, Dec 16 2016. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/27986644

[59] C. S. Kathryn R. Fingar, Audrey J. Weiss, Claudia A. Steiner. (2014) Most Frequent Operating Room Procedures Performed in U.S. Hospitals, 2003-2012. [En ligne]. Disponible: https://www.hcup-us.ahrq.gov/reports/statbriefs/sb186-Operating-Room-Procedures-United-States-2012.jsp

[60] R. Jandhyala, "Delphi, non-RAND modified Delphi, RAND/UCLA appropriateness method and a novel group awareness and consensus methodology for consensus measurement: a systematic literature review," vol. 36, nº. 11, p. 1873-1887, Nov 2020. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pubmed/32866051

ANNEXE A LISTE DES VARIABLES UTILISÉES PAR CHIRURGIE

Tableau Supplémentaire A.1 Liste et descriptions des variables utilisées pour chaque chirurgie

Nom de variable	Description	Lam.	Col.	Tho.
AGE_DEID	Âge (limité à 90 ans)	1	1	
TOTAL_SURGERY_ON_DATE	Nombre de chirurgies dans le passé	1	1	1
FIRST_START_CASE_YN	Premier cas de la journée	1	1	1
SCHED_MINS	Durée cédulée pour la chirurgie	1	1	1
ADULT_MDRD_GFR_BLK	Débit de filtration glomérulaire (reins)	1	1	1
ETOH_YN	Consommation d'alcool	1	1	1
LAST_EF_NUM_VAL	Fraction d'éjection du ventricule gauche	1	1	
MAX_PREOP_PAIN	Douleur préopératoire maximale	1	1	1
POSPOM_AFIB	Fibrillation auriculaire	1	1	
POSPOM_CANCER	Présence d'un cancer	1		1
POSPOM_CVD	Maladie cardiovasculaire	1		1
POSPOM_DRUG	Consommation de drogues	1		1
POSPOM_ETOH	Consommation d'alcool	1		1
POSPOM_HEMIPLEGIA	Hémiplégie	1		
POSPOM_ISCHEMIC_HEART_D	Maladie cardiaque ischémique	1	1	
POSPOM_LIVER	Pathologie liée à la fonction hépatique	1		1
POSPOM_METASTASES	Présence de métastases sur un cancer	1		
POSPOM_OBESITY	Présence d'obésité	1		
POSPOM_PULM_CIRC	Maladie de la circulation pulmonaire	1	1	1
POSPOM_TRANSPLANT	Présence d'un organe transplanté	1		
POSPOM_VALVE_HD	Maladie cardiaque valvulaire	1	1	
POSPOM_VASC_DISEASE	Maladie vasculaire périphérique	1	1	
POSPOM_TOTAL	Nombre total de comorbidités	1	1	
PREOP_WEIGHT_KG	Poids (kg) préopératoire	1		
BMI	Indice de masse corporelle	1		1
PREINDUCTION_MED_HR	Fréquence cardiaque préopératoire	1	1	1
CHF_FLAG	Drapeau pour insuffisance cardiaque	1	1	1
RESP_FLAG	Drapeau d'une maladie respiratoire	1	1	
MELD_SCORE	Score MELD (lié à la fonction du foie)	1		1
FRAMINGHAM_SCORE	Score Framingham (risque cardiaque)	1		
LAST_ICU_LOS_HRS	Dernière durée d'hospitalisation en USI	1		1
LOS_PRE_SURG	Durée d'hospitalisation préopératoire	1	1	1

Tableau Supplémentaire A.1 - Liste et descriptions des variables utilisées pour chaque chirurgie (suite et fin)

LAPAROSCOPY	« Laparoscopie » utilisée dans le nom de chirurgie	1	1	1
ELECTIVE	Chirurgie élective (non-urgente)	1	1	1
SEX	Sexe		1	1
SMOKING_YN	Fumeur		1	
TPN_48HR_YN	Nutrition parentérale totale dans les 48h précédent la chirurgie		1	1
POSPOM_COPD	Maladie pulmonaire obstructive		1	
POSPOM_DEPRESSION	Dépression majeure		1	
POSPOM_DIALYSIS	Dialyse		1	1
EPIDURAL_THORACIC_YN	Utilisation d'épidurale thoracique		1	1
ANES_PLAN_GENERAL_YN	Planification d'anesthésie générale		1	1
LV_EF	Présence d'une valeur de fonction cardiaque au dossier		1	
PROB_LIST_CHF	Nombre de points de probabilité pour une insuffisance cardiaque		1	
IHD_FLAG	Drapeau maladie cardiaque ischémique		1	
PREV_CABG	Antécédent de pontage cardiaque		1	
PROB_LIST_IHD	Nombre de points de probabilité pour une maladie cardiaque ischémique		1	
ROBOTIC_CASE_YN	Chirurgie impliquant un robot			1
POSPOM_HTN	Hypertension artérielle			1
POSPOM_RENAL_FAILURE	Insuffisance rénale			1
PREINDUCTION_MED_MAP	Tension artérielle préopératoire			1
AICD_LVAD_HX	Présence de stimulateur cardiaque			1
DM_FLAG	Drapeau de diabète			1
HOME_INSULIN	Utilisation d'insuline à la maison			1
ADMSN_ICD_CODE	Présence de codes ICD à l'admission lié à la maladie cardiaque			1

ANNEXE B LISTE DES VARAIBLES EXPLORÉES ET NON-UTILISÉES PAR MANQUE DE CORRÉLATION

Tableau Supplémentaire B.1 - Liste et descriptions des variables utilisées pour chaque chirurgie

Nom de variable	**Description**
ANES_CASE_YN	Chirurgie cédulée avec un anesthésiste
CHD_PROBLIST	Nombre de points de probabilité pour une maladie coronarienne congestive
PEPC_STATUS	Patient vu en clinique préopératoire
PONV_HR_PATIENT_YN	Patient à risque de nausée postopératoire
PONV_HR_SURGERY_YN	Chirurgie à risque de nausée postopératoire
POSPOM_ANEMIA	Anémie
POSPOM_ARRHYTHMIA	Arythmie cardiaque
POSPOM_CHF	Maladie cardiaque congestive
POSPOM_CHRON_RESP_FAILURE	Insuffisance respiratoire chronique
POSPOM_DEMENTIA	Démence
POSPOM_PSYCHOSIS	Pscyhose
IDEAL_BODY_WEIGHT_KG	Poids idéal (kg)
EPIDURAL_LUMBAR_YN	Présence d'épidurale lombaire
REGIONAL_BLOCK_YN	Présence d'un bloc nerveux périphérique
REGIONAL_CATH_YN	Présence d'un cathéter nerveux périphérique
SPINAL_YN	Utilisation d'anesthésie rachidienne
TRANSPLANT	Présence de transplantation
PREV_STENT	Présence de tuteurs coronariens
ANES_PREOP_DM	Présence de diabète à l'évaluation préopératoire
OTHER_INJECTABLE	Présence de médicaments injectables pour diabète outre l'insuline
DM_MED_HX	Présence de médication pour diabète
DM_PROB_LIST	Nombre de points de probabilité pour un diabète

ANNEXE C HYPERPARAMÈTRES UTILISÉS DANS L'EXPLORATION DES MODÈLES NON-SUPERVISÉS

Cette annexe présente la liste des hyperparamètres explorés en grille pour chaque modèle supervisé. Pour assurer la reproductibilité des résultats, le paramètre `random_state` était fixé à 1.

Tableau Supplémentaire C.1 – Liste des hyperparamètres explorés par grille dans les modèles supervisés

a) DB-Scan

Hyperepamètres	**Liste explorée par grille**

`eps`	`[20,25,30,35,40,45,50,55,60,65,70,75,80]`
	
	Puis autour du coude:
	`[42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61]`

`min_samples`	`[14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]`

b) Hiérarchique descendant

Hyperepamètres	**Liste explorée par grille**
`n_clusters`	`[2, 3, 4, 5, 6]`
`linkage`	`['ward','complete','average','single']`
`affinity`	`['euclidean']`

c) K-means

Hyperepamètres	**Liste explorée par grille**
n_clusters	[2, 3, 4, 5, 6]
N_init	[2, 4, 6, 8, 10]
init	['k-means++', 'random']

d) Consensus k-means

Hyperepamètres	**Liste explorée par grille**
n_clusters	[2, 3, 4, 5, 6]
N_init	[2, 4, 6, 8, 10]
init	['k-means++', 'random']
n_repeats	[7, 10, 13]
samp_frac	[0.5, 0.6, 0.7, 0.8, 0.9]

ANNEXE D COMPARAISON DES PERFROMANCES DE SEGMENTATION SELON LA STRATÉGIE DE NORMALISAITON DES DONNÉES

Cette annexe est complémentaire au Tableau 5.2 présentant les résultats pour la chirurgie de colectomie. Il permet d'apprécier que les performances supérieures de la normalisation standard comparativement à la normalisation *MinMax*.

Tableau D.1 - Métriques de performance du consensus k-means en chirurgie de laminectomie (a) et chirurgie thoracique (b) selon deux différentes normalisations

a)

	Normalisation standard				**Normalisation *MinMax***			
Nb phénotypes	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
Silhouette	0.326	0.127	0.070	0.060	-0.002	-0.02	-0.06	-0.08
NMI	0.002	0.007	0,008	0.020	0.007	0.012	0.006	0.006
IRA	0.005	0.009	0.010	0.010	-0,003	0.006	0.002	0.001

b)

	Normalisation standard				**Normalisation *MinMax***			
Nb phénotypes	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
Silhouette	0.211	0.072	0.021	-0.02	0.169	-0.045	-0.054	-0.106
NMI	0.050	0.028	0.026	0.028	0.056	0.027	0.019	0.016
IRA	0.079	0.043	0.041	0.042	0.055	0.014	0.010	0.003

ANNEXE E DÉTAILS DE PERFORMANCE DU PHÉNOTYPAGE PAR CHIRURGIE ET ISSUE CLINIQUE

Tableau E.1 - Détails des aires sous la courbe ROC par chirurgie et issue postopératoire

		Mortalité hospitalière	Mortalité 30J	Réopération 30J	Admission USI	Séjour >90e
Laminectomie	Digital	0.82	0.81	0.64	0.81	0.78
	ASA	0.76	0.86	0.59	0.64	0.72
	Combiné	0.81	0.90	0.65	0.76	0.82
Colectomie	Digital	0.91	0.90	0.63	0.78	0.75
	ASA	0.67	0.83	0.59	0.77	0.70
	Combiné	0.80	0.93	0.64	0.85	0.78
Chirurgie thoracique	Digital	0.73	0.69	0.54	0.65	0.69
	ASA	0.75	0.71	0.58	0.66	0.67
	Combiné	0.78	0.72	0.59	0.72	0.74

ANNEXE F COMBINAISON LINÉAIRE DU PHÉNOTYPE ET SCORE ASA

Afin de combiner le phénotype du patient au score ASA, l'échelle suivante fut utilisée :

Tableau F.1. Combinaison du phénotype et score ASA en score linéaire

Score ASA	Phénotype	Score linéaire
1	0	**1**
1	1	**2**
1	2	**3**
2	0	**4**
2	1	**5**
2	2	**6**
3	0	**7**
3	1	**8**
3	2	**9**
4	0	**10**
4	1	**11**
4	2	**12**
5	0	**13**
5	1	**14**
5	2	**15**

ANNEXE G HYPERPARAMÈTRES EXPLORÉS POUR LES MODÈLES SUPERVISÉS

Cette annexe présente la liste des hyperparamètres explorés en grille pour chaque modèle supervisé. Pour assurer la reproductibilité des résultats, le paramètre `random_state` était fixé à 1.

Tableau G.1 – Liste des hyperparamètres explorés par grille dans les modèles supervisés.

a) Régression logistique

Hyperepamètres	Liste explorée par grille
C_Val	[0.01,0.1,1.0,10.0,100.0]
Solver	['newton-cg','lbfgs','saga']
Penalty	['none','l1','l2']

b) Forêt aléatoire

Hyperepamètres	Liste explorée par grille
n_estimators	[100, 200, 300, 400, 500]
max_depth	[4, 8, 12]
min_samples_leaf	[5, 10, 15, 20]
max_features	['sqrt', 'log2']

c) Perceptron multicouche

Hyperepamètres	Liste explorée par grille
hidden_layer_sizes	[(200,100,50),(100,50,25),(100,25),(200,100)]
activation	['identity','tanh','relu']
learning_rate_init, max_iter	[[1e-2,100],[1e-3,200],[1e-4,400]]
alpha	['sqrt', 'log2']