Prédiction de la mortalité après chirurgie cardiaque programmée : nouvelles approches par Machine Learning et Decision Curve Analysis

Myriem Belghiti Alaoui

Université de Bordeaux
U.F.R. DES SCIENCES MEDICALES

Année 2017 N° 3041

Thèse pour l'obtention du

DIPLOME d'ETAT de DOCTEUR EN MEDECINE

Spécialité : Anesthésie Réanimation

Présentée et soutenue publiquement
le 22 mai 2017
par Myriem BELGHITI ALAOUI

Née le 25/11/1988 à Metz

Prédiction de la mortalité après chirurgie cardiaque programmée : nouvelles approches par *Machine Learning* et *Decision Curve Analysis*

Directeur de thèse
Monsieur le Docteur Jérôme ALLYN

Rapporteur
Monsieur le Docteur Bernard-Alex GAÜZERE

Jury

Monsieur le Professeur Alexandre OUATTARA	**Président**
Monsieur le Professeur Roger SALAMON	Membre
Monsieur le Professeur Eric BRAUNBERGER	Membre
Monsieur le Docteur Nicolas ALLOU	Membre
Monsieur le Docteur Jérôme ALLYN	Membre

U.F.R. DES SCIENCES MEDICALES

Année 2017 N° 3041

Thèse pour l'obtention du

DIPLOME d'ETAT de DOCTEUR EN MEDECINE

Spécialité : Anesthésie Réanimation

Présentée et soutenue publiquement

le 22 mai 2017

par Myriem BELGHITI ALAOUI

Née le 25/11/1988 à Metz

Prédiction de la mortalité après chirurgie cardiaque programmée : nouvelles approches par *Machine Learning et Decision Curve Analysis*

Directeur de thèse
Monsieur le Docteur Jérôme ALLYN

Rapporteur
Monsieur le Docteur Bernard-Alex GAÜZERE

Jury

Monsieur le Professeur Alexandre OUATTARA	**Président**
Monsieur le Professeur Roger SALAMON	Membre
Monsieur le Professeur Eric BRAUNBERGER	Membre
Monsieur le Docteur Nicolas ALLOU	Membre
Monsieur le Docteur Jérôme ALLYN	Membre

Remerciements

A Monsieur le Professeur OUATTARA,
Vous me faites l'honneur de présider ce jury de thèse. En tant qu'anesthésiste réanimateur en chirurgie cardiaque et du fait de votre implication auprès des internes d'anesthésie réanimation, je n'envisageais pas de soutenir ma thèse sans votre présence. Je vous remercie pour l'enseignement que vous assurez auprès des internes de Bordeaux et d'Outre-Mer.

A Monsieur le Docteur Jérôme ALLYN,
Je n'ai pas assez de ces quelques lignes pour t'exprimer ma reconnaissance. J'ai découvert un sujet éminemment actuel et original. Tu m'as accompagnée avec obstination, de façon pertinente avec la même rigueur que celle que tu nous enseignes en réanimation. J'espère avoir été à la hauteur de la confiance que tu m'as accordée.

A Monsieur le Docteur Bernard-Alex GAÜZERE,
Je vous remercie pour vos précieux conseils, votre rapport détaillé ainsi que pour votre réactivité. Merci également de votre investissement auprès des internes et de partager avec nous votre expérience ainsi que votre sens clinique tant au lit du malade que sur les bancs de l'université.

A Monsieur le Professeur Roger SALAMON,
Je suis honorée que vous vous rendiez disponible pour juger mon travail. Vous apportez une expertise en termes de santé publique ainsi qu'un point de vue différent du clinicien pour la critique de ma thèse. Merci également pour l'enseignement que vous délivrez aux étudiants en médecine dont j'ai pu profiter puisque je suis issue de la formation bordelaise.

A Monsieur le Professeur Eric BRAUNBERGER,
Je vous remercie d'avoir accepté de faire partie du jury de ma thèse et de représenter le point de vue chirurgical incontournable pour la critique de ce travail. Je me souviendrai de votre engagement auprès des patients et en salle d'intervention.

A Monsieur le Docteur Nicolas ALLOU,
Je te remercie de te rendre disponible pour juger mon travail. Merci également pour ta bonne humeur, ton enthousiasme et ton enseignement précieux en réanimation !

Merci à Monsieur le Docteur Cyril FERDYNUS pour l'analyse statistique.

Un remerciement particulier au Docteur Julien JABOT, tu suscites la vocation et l'entrain pour notre spécialité. Merci pour ton investissement et ton enseignement auprès des internes, j'espère que j'aurai le plaisir d'en profiter à nouveau.

A Isa, Clairette et Armellou pour votre soutien durant ces longues semaines, je pense à vous de l'autre côté de l'Équateur. A Mimi et Flo quelle joie de vous retrouver !

Aux copines bordelaises avec qui j'ai affronté la tempête : Stef, Laëti et Laure.

A mon Elsa depuis dix ans et son Lionel, merci de tout.

A mes parents, merci pour vos encouragements et votre soutien depuis le début de mon cursus il y a onze ans.

Et évidemment, à Sabria, Adam et Nour, deux sœurs et un frère pour d'infinies farandoles.

Abréviations

APACHE : Acute Physiology And Chronic Health Evaluation
ARA 2 : Antagoniste des Récepteurs de l'Angiotensine 2
AUC : Area Under the Curve
AVC : Accident Vasculaire Cérébral
BPCO : Broncho Pneumopathie Chronique Obstructive
CART : Classification And Regression Trees
CEC : Circulation Extra Corporelle
CHU : Centre Hospitalier Universitaire
CNIL : Commission Nationale de l'Informatique et des Libertés
DCA : Decision Curve Analysis
FEVG : Fraction d'Ejection du Ventricule Gauche
GBM : Gradient Boosting Machine
HAS : Haute Autorité de Santé
HTA : Hyper Tension Artérielle
HTAP : Hyper Tension Artérielle Pulmonaire
HR : Hazard Ratio
IA : Intelligence Artificielle
IEC : Inhibiteur de l'Enzyme de Conversion
IMC : Indice de Masse Corporelle
MDRD : Modification of Diet in Renal Disease
ML : Machine Learning
NB : Net Benefit
NRI : Net Reclassification Improvement
NYHA : New York Heart Association
OR : Odd Ratio
PAC : Pontage Aorto Coronarien
PAPS : Pression Artérielle Pulmonaire Systolique
RF : Random Forests
ROC : Receiver Operating Charateristic
SAPS : Simplified Acute Physiology Score
SOFA : Sequential Organ Failure Assessment
SVM : Support Vector Machine
STS : Society of Thoracic Surgeons

Table des matières

Index des tableaux

Index des figures

I – INTRODUCTION

La morbi-mortalité est importante après chirurgie cardiaque et les facteurs intervenant dans la décision d'opérer ou non sont variés (1,2). L'utilisation de scores de mortalité est recommandée pour stratifier le risque opératoire de ce type d'intervention et aider au processus décisionnel (3).

Les deux scores les plus connus sont l'EuroSCORE II et le STS score (4-7). En Europe, l'EuroSCORE II est le plus largement utilisé. Toutefois, il présente des limites d'application notamment chez les sujets à haut risque et chez les patients à opérer d'une chirurgie valvulaire (8). Ces scores de mortalité sont établis à l'aide d'une régression logistique.

Le développement exponentiel de bases de données de patients multi institutionnelles à grande échelle favorise le développement de l'Intelligence Artificielle en médecine (9,10). Cette stratégie permet l'analyse prédictive de données brutes en quantité massive (apprentissage artificiel ou en anglais *Machine Learning*). Dans ce contexte, il est légitime de proposer l'IA pour établir un nouveau score de prédiction de la mortalité d'une chirurgie cardiaque.

De plus, les méthodes habituelles d'évaluation et de comparaison de modèles prédictifs, consistant en l'analyse des courbes ROC, jugent de la validité interne d'un score *via* sa sensibilité et sa spécificité, mais n'évaluent pas sa pertinence clinique (11). Des méthodes récentes comme la *Decision Curve Analysis* ont été développées afin d'impliquer les patients dans la prise de décision les concernant et c'est ce qui est désormais recommandé par la HAS (12).

L'objectif premier de notre étude a été d'améliorer la sélection des patients en amont d'une chirurgie cardiaque programmée en optimisant la prédiction de la mortalité hospitalière post opératoire. Pour ce faire, nous avons créé un modèle de prédiction de *Machine Learning* que nous avons comparé au score recommandé : l'EuroSCORE II.

L'objectif second de notre étude a été de fournir un modèle d'évaluation des modèles prédictifs complémentaires des courbes ROC, la DCA. Nous avons donc comparé nos modèles via ces deux méthodes.

II - CONTEXTE

1. Morbi-mortalité après chirurgie cardiaque

La morbi-mortalité reste élevée au décours d'une intervention de chirurgie cardiaque. En effet, malgré les progrès des techniques chirurgicales et des techniques anesthésiques, les études montrent toujours une mortalité post opératoire voisine de 3% en France (1). Les facteurs évoqués pour expliquer cette mortalité sont d'une part une modification du profil des patients qui sont plus âgés et présentent plus de comorbidités, et d'autre part une modification du type de chirurgie cardiaque avec une proportion de chirurgie valvulaire ou combinée qui augmente aux dépens de la chirurgie coronarienne. La revascularisation percutanée est de plus en plus privilégiée (13)(8). Or, le taux de mortalité des chirurgies valvulaires ou combinées est supérieur à celui de la revascularisation coronaire (13-15).

L'évaluation du risque opératoire chez ces patients est primordiale et recommandée par les sociétés savantes afin d'adapter au mieux la prise en charge et *in fine* d'améliorer le pronostic et de diminuer les coûts de santé (3). Ceci en renonçant, dans le cadre de la discussion bénéfice-risque à des interventions chirurgicales dont l'issue est trop incertaine pour le patient. En pratique, il existe de grandes disparités dans le choix des critères d'opérabilité chez les différents médecins, intervenants et décideurs en chirurgie cardiaque (2)(16).

L'une des voies d'amélioration du pronostic repose donc sur la sélection pluridisciplinaire des patients en amont de l'intervention. En chirurgie cardiaque, les recommandations européennes préconisent l'utilisation de scores afin d'estimer la mortalité post opératoire et d'aider à la prise de décision chirurgicale. Parmi les scores proposés, les deux les plus étudiés sont le STS score (*Society of Thoracic Surgeon*) et l'EuroSCORE II avec des grades de recommandation respectifs I B et IIa B (3).

2. Les principaux modèles de prédiction

2.1 EuroSCORE I et II

La première version de l'EuroSCORE a été développée en 1999 par une équipe française (17,18). Ce score était simple à calculer au lit du malade mais surestimait le risque de mortalité des patients à haut risque (19). Il a donc été proposé en 2003 un EuroSCORE logistique dont le calcul, plus complexe, nécessitait le recours à un logiciel (20).

Néanmoins, ce modèle logistique continuait de surestimer le risque de mortalité de toutes les catégories de patients, tout particulièrement chez les patients valvulaires ou à haut risque chirurgical (21).

C'est ainsi qu'un nouveau modèle prédictif a été proposé en 2012 : l'EuroSCORE II, établi à partir d'une base de données de 22 181 patients, répartis dans 154 hôpitaux de 43 pays. La période prospective de recueil s'était étalée sur 12 semaines en 2010 (4).

La comparaison de la cohorte de l'EuroSCORE II à celle ayant permis l'élaboration de l'EuroSCORE I a permis de mettre en évidence une proportion plus importante de patients présentant un statut NYHA IV, une insuffisance rénale, une broncho-pneumopathie obstructive chronique (BPCO) ou encore une artériopathie périphérique. Le nombre de patients de plus de 90 ans s'élevait à 21 (0,09 %).

Le taux de chirurgie coronaire isolée était quasiment égal à celui de chirurgie valvulaire, soit respectivement 46,7 % et 46,3 %. Bien que les patients aient présenté plus de comorbidités, la mortalité globale à 30 jours a été de 3,9 % (4,6 % dans la cohorte de l'EuroSCORE original).

Après analyse bi puis multivariée par régression logistique, 18 variables se sont avérées significativement et indépendamment associées à la mortalité à 30 jours, telles que représentées dans le tableau I.

Tableau I : Variables de l'EuroSCORE II

Age	Angor stade IV
Genre féminin	Infarctus du myocarde récent < 90 j
Clairance de la créatinine : >85 mL/min 51-85 mL/min < 50 mL/min Dialyse	FEVG : > 50 % 31 – 50 % 21- 30 % < 20 %
Chirurgie cardiaque antérieure	NYHA I, II, III ou IV
Troubles neurologiques, mobilité diminuée	HTAP : PAPS 30 – 55 mmHg PAPS > 55 mmHg
BPCO	Urgence
Endocardite active	Poids de l'intervention : PAC isolé Chirurgie simple non coronaire 2 interventions 3 interventions
Diabète insulino-requérant	Chirurgie aorte thoracique
Artériopathie périphérique	Etat critique pré opératoire

Les modifications par rapport à l'EuroSCORE I sont portées dans les cellules grisées

Au sein de la cohorte test, l'AUC de la courbe ROC a montré une amélioration modérée du pouvoir discriminant (représenté par l'aire sous la courbe ROC) de 2 % en faveur de l'EuroSCORE II (AUC 0,81 *versus* 0,79 pour EuroSCORE). La valeur de l'EuroSCORE II au sein de la cohorte de validation était de 3,95 %. La calibration et donc la précision de ce modèle, étudiée grâce au rapport de la mortalité observée sur la mortalité prédite était de 1,058 (légère sous-estimation) *versus* 0,67 et 0,53 pour l'EuroSCORE I additif (calculable au lit du malade) et logistique respectivement (surestimation).

Notons qu'il persiste des limitations à l'application de l'EuroSCORE II. Plusieurs auteurs ont montré un défaut de calibration et/ou de discrimination pour les populations à haut risque et notamment chez les sujets âgés (14)(22-24). Une autre étude a retrouvé un défaut de calibration chez les patients à faible risque opératoire (25). Par ailleurs, le faible nombre d'études de validation de l'EuroSCORE II pour la chirurgie valvulaire limite sa généralisation au sein de cette population. De plus, un défaut de discrimination est retrouvé pour le remplacement valvulaire aortique isolé (26) et la calibration n'est pas satisfaisante lorsqu'il s'agit d'interventions valvulaires percutanées (*Transcatheter Aortic Valve Implementation, Mitral Clip*) (27-29).

2.2 STS score

Le STS score a été établi par régression logistique. Trois modèles ont été développés, en fonction du type de chirurgie : pontage aorto-coronarien, chirurgie valvulaire et chirurgie combinées. Trois bases de données américaines ont été utilisées, comprenant :

- 774 881 patients pour la chirurgie coronarienne
- 109 759 patients pour la chirurgie valvulaire
- 101 661 patients pour la chirurgie combinée

Ce score estime la mortalité mais aussi la morbidité post opératoire (AVC, dysfonction rénale, durée du séjour, médiastinite, ventilation prolongée, ré intervention) et présente l'avantage d'être adapté spécifiquement aux différents types de chirurgie cardiaque (5-7).

2.3 Méthode statistique traditionnelle pour établir un score prédictif : l'analyse multi variée par régression logistique

La régression logistique consiste à expliquer une variable qualitative (dans notre cas, le décès post opératoire) par des variables explicatives (30,31).

Le choix des variables explicatives est basé sur un travail obligatoire de revue de la littérature en préalable à l'analyse. Une fois les variables sélectionnées, une analyse bivariée est réalisée afin d'étudier séparément la force de l'association entre chaque variable explicative et la variable à expliquer. Les variables les plus intéressantes et/ou celles qui ont un p bas, sont incluses dans l'analyse multivariée. Il existe des pré-requis obligatoires avant analyse par régression logistique, comme la normalité des distributions des variables explicatives et un effectif suffisant pour le nombre de variables sélectionnées dans le modèle.

L'analyse multivariée permet d'établir un modèle apportant le maximum d'informations avec le minimum de variables explicatives : c'est le principe de parcimonie. Plusieurs modèles (avec différentes variables) peuvent ainsi être élaborés puis analysés par régression logistique, afin de déterminer s'il existe un lien significatif avec la variable à expliquer ou si les variables explicatives sont liées entre elles.

L'analyse bivariée puis multivariée permet de définir des *odd ratio* bruts puis ajustés (en fonction de la présence ou non des autres variables) et donc de pondérer chaque variable explicative.

La fonction logistique permet de linéariser la relation entre la variable à expliquer (le décès) et les différentes variables explicatives. Puis une seconde transformation mathématique complexe est réalisée pour revenir à un modèle linéaire classique et ainsi interpréter les différents *odd ratio* obtenus.

3. Alternatives à la régression logistique ou nouveaux outils : Machine Learning et processus décisionnel

3.1 Présentation

Il n'y a pas de consensus pour définir l'intelligence et encore moins pour définir l'Intelligence Artificielle (IA). Le dictionnaire Larousse définit l'intelligence comme la *« capacité de comprendre, de saisir une chose par la pensée ou encore l'aptitude à s'adapter à une situation, à choisir des moyens d'action en fonction des circonstances ».*

Le terme d'IA été créé, en 1956, en fonction des différentes définitions retrouvées dans la littérature, l'IA peut être classée selon quatre catégories, représentées dans le tableau II (32).

Tableau II : Définitions de l'IA en quatre catégories

	Standard de l'être humain	**Standard de rationalité plus général**
Focalisation sur le fonctionnement du système (raisonnement)	Capacité à penser comme un humain	Capacité à penser rationnellement selon les lois de la logique mais certaines capacités comme la perception ne font pas appel à un raisonnement logique
Focalisation sur le comportement du système	Capacité à agir comme un humain	Capacité à se comporter rationnellement : développement d'agents intelligents qui agissent pour satisfaire au mieux leurs objectifs

Le champ de l'IA va au-delà de la compréhension des mécanismes de pensée ou de comportements humains. L'objectif est de construire des entités « intelligentes », capables de réaliser des tâches « intelligentes » qui étaient antérieurement réalisées exclusivement par l'homme.

Pour Turing, cette étape passe par l'apprentissage, car une machine intelligente doit reproduire le comportement humain. Turing fait l'analogie avec le développement humain, c'est à dire entre apprentissage naturel et apprentissage artificiel (33).

Ainsi un ordinateur, tout comme un enfant, va observer, intégrer des données et à partir de ces dernières extraire des règles pour interpréter son environnement. Certaines connaissances médicales sont empiriques, issues d'observations, ne reposant pas sur un substrat explicatif théorique. Dans le domaine de l'apprentissage automatique, ce processus est appelé induction. Il permet de générer du sens en passant des faits à la loi, du particulier au général. Afin d'aboutir à un modèle performant, le programme doit analyser plusieurs milliers de données. C'est l'apprentissage par l'exemple (34).

Le « *Machine Learning* » - littéralement apprentissage automatique - est l'un des pans de l'IA le plus développé. La création de bases de données contenant plusieurs milliers de patients en chirurgie cardiaque est propice au développement de ces nouvelles technologies (9). Ainsi, un ordinateur analyse ces données massives (« *big data* ») et émet une probabilité de survenue de l'évènement d'intérêt, dans notre cas, le décès. Arthur Samuel définit cette discipline comme la capacité d'un ordinateur *« à apprendre sans y être explicitement programmé »*.

Parallèlement au développement du *Machine Learning*, on observe un bouleversement des pratiques scientifiques. Ces « *big data* » sont analysées dans leur ensemble à partir d'algorithmes qui permettent l'élaboration de modèles prédictifs performants. Les relations entre toutes les variables sont analysées, y compris s'il n'existe pas de linéarité. Une corrélation est mise en évidence entre différentes variables et le décès. En revanche, il n'y a pas d'analyse fine entre chaque facteur de risque présenté par le patient et la survenue du décès. L'objectif étant de repérer les patients à risque de décès afin d'influencer le processus décisionnel et non de comprendre les raisons de ces corrélations. C'est pour cela que certains auteurs parlent de « boîte noire » (35,36).

Il existe deux grandes familles d'algorithmes de *Machine Learning* selon qu'ils fournissent ou non la variable de sortie : les algorithmes supervisés et les algorithmes non supervisés (37).

3.2 Applications hors de la médecine

Même si le concept est ancien, le développement et les applications de l'IA sont récents. En effet, c'est la production de données massives et les capacités de traitement informatique qui ont fourni la puissance de calcul et le volume nécessaires au développement de l'IA. Cette dernière fait partie intégrante de notre quotidien personnel.

Les grandes entreprises exploitent ces bases de données en partie grâce aux algorithmes de *Machine Learning*, afin :

- d'améliorer leurs services,
- de personnaliser leur relation avec les clients,
- de développer des modèles de prédiction du comportement humain et par conséquent de mieux comprendre les évènements engendrés par ce dernier,
- d'améliorer l'expertise humaine *via* des « assistants virtuels » (38).

Le *Machine Learning* fait donc déjà partie intégrante des différentes stratégies de développement de ces entreprises. Citons Google® et sa traduction automatique, Amazon®, Facebook® (reconnaissance des visages, suggestion d'amis en fonction de notre profil, etc.), ou encore Netflix®. L'IA se développe essentiellement dans les branches recherche et développement ainsi que production de ces entreprises. Son champ d'application ne cesse de croître : finance, commerce, météorologie, communication art, marketing, éducation, culture…

En France, la transition est plus progressive et les entreprises communiquent moins à ce sujet. Mais les banques et assurances utilisent d'ores et déjà des algorithmes de *Machine Learning* afin de déterminer qui sont les « bons » emprunteurs, d'évaluer les risques financiers ou encore le risque de fraude bancaire. Sur les marchés financiers, le *trading* à haute fréquence en est un autre exemple.

3.3 Applications en médecine

3.3.1 Exemples d'application

Dès les années 60, des scientifiques ont réfléchi à l'introduction de techniques d'IA en médecine mais les outils informatiques n'avaient pas encore atteint la puissance de calcul nécessaire et les données n'étaient pas suffisamment disponibles (39). Initialement, l'apprentissage automatique était utilisé en biologie médicale et permettait de déterminer la structure d'une protéine ou encore l'expression phénotypique d'une séquence de gènes (40,41).

Dans le domaine médical, le *Machine Learning* s'est surtout développé depuis le début des années 90 avec l'émergence des dossiers médicaux électroniques et donc de bases de données des patients. La mise à disposition de « *big data* » *via* les dossiers médicaux électroniques impose une refonte de la gestion et du traitement des données en médecine. Les milliers de variables dont nous disposons ne peuvent être analysées seulement par l'œil et la main humaines mais peuvent l'être de façon complémentaire par des algorithmes de *Machine Learning*. Certes, la régression logistique peut permettre d'analyser ces données et d'établir des liens linéaires entre différentes variables et le taux de décès, mais par essence, le nombre de variables qu'elle est capable de prendre en compte est limité et prédéterminé (42).

Pour certains chercheurs, le *Machine Learning* va être impliqué plus avant dans trois champs de la médecine, à savoir :

- Amélioration des capacités des professionnels de santé à établir un pronostic,
- Assistance virtuelle des anatomo-pathologistes et des radiologues,
- Amélioration de la précision des diagnostics médicaux (42).

Le nombre de publications et d'applications de l'IA (*Machine Learning* et *Deep Learning*) en médecine est exponentiel et il est impossible de réaliser une revue de la littérature exhaustive. En voici quelques exemples, à retrouver en tableau III et IV :

- Des méthodes de reconnaissance d'images se développent de plus en plus dans les champs de la médecine qui impliquent une analyse d'images. Certaines caractéristiques histologiques ou radiologiques étant difficiles à mettre en évidence par inspection manuelle, un support informatique peut identifier de façon précise ces éléments (42-46).

- En psychiatrie, l'utilisation d'algorithmes de *Machine Learning* est d'actualité, afin de prédire, après lecture IRM, le risque d'évolution d'un patient à risque vers une pathologie cible. Plusieurs études ont été menées en ce sens chez des patients présentant une démence type Alzheimer, un épisode dépressif majeur ou encore une schizophrénie (47,48).

D'autres études ont établi des modèles prédictifs sur la base de données cliniques notamment pour prédire le pronostic d'un cancer du sein ou encore celui d'un AVC ischémique (49,50).

Tableau III : exemples d'application du *Machine Learning* en médecine, la reconnaissance d'images

Discipline - Domaine	Objectif	Design de l'étude	Algorithmes utilisés	Résultats
Psychiatrie - Schizophrénie Veronese *et al.* 2016 (48)	Prédire la survenue d'une schizophrénie par interprétation IRM	Revue de la littérature 10 essais comportant chacun : - entre 64 et 20 sujets schizophrènes - entre 75 et 20 sujets sains. IRM cérébrale à volontaires sains et volontaires schizophrènes.	*Support Vector Machine* (SVM)	Précision des modèles entre 79 et 90% en fonction des études.
Oncologie - Cancer pulmonaire Yu *et al.* 2016 (43)	Prédire le pronostic du cancer du poumon après lecture anatomo-pathologique complètement automatisée	Etude de 2 480 planches histologiques de patients présentant un adénocarcinome pulmonaire ou un cancer du poumon à petites cellules. Mise en évidence de 9 879 caractéristiques histologiques (taille et forme du noyaux, texture du cytoplasme, couleur, etc).	*Random Forest Bagging* Classifieur naïf bayésien Modèle de Cox	Classifieurs permettant de différencier : - une cellule cancéreuse d'une cellule saine avec une AUC 0,81 - les deux types de cancers pulmonaires AUC 0,78 Modèle de Cox permettant de classer en bon ou mauvais pronostic - pour les cancers à petites cellules p = 0,035 - pour les adénocarcinome p= 0,0023
Oncologie - Lymphome Sertel *et al.* 2008 (44)	Reconnaître la nature et le grade d'un lymphome	Etude de 17 planches histologiques de lymphome	*K nearest neighbour* Classifieur naïf bayésien	Précision moyenne : 83,7 %
Oncologie - Neuroblastome Sertel *et al.* 2009 (51)	Prédire le pronostic d'enfants atteints de neuroblastome	Développement du modèle avec 500 images issues de deux planches histologiques de neuroblastome Etude de validation au sein de 43 planches histologiques de neuroblastome	*K nearest neighbour*	Précision : 88,4 %

Tableau IV : exemples d'application du *Machine Learning* en médecine

Discipline – Domaine	Objectif	Design de l'étude	Algorithmes utilisés	Résultats
Oncologie - Adénocarcinome mammaire Boughorbel *et al.* 2016 (49)	Etablir un modèle prédictif pronostique du cancer du sein	Etude rétrospective de 11 variables sélectionnées parmi 25 (taille de la tumeur, âge lors du diagnostic, ganglion(s), grade, type histologique, récepteurs à œstrogène ou à progestérone, Her2, stade TNM, curage ganglionnaire) Base de données de 1981 patientes Critère de Jugement Principal : survenue décès	*Generalized Linear Model* *Support Vector Machine* *Random Forests* Arbres boostés Réseaux neuronaux *k nearest neighbour*	Aires sous la courbe entre 0,67+/- 0,09 (*k nearest neighbors*) et 0,76 +/- 0,05 (*Random Forest*)
Neurologie - AVC ischémique antérieur Asadi *et al.* 2014 (50)	Etablir un modèle prédictif du pronostic fonctionnel dans les suites d'un AVC	Etude rétrospective monocentrique d'une base de données prospective australienne de 107 patients ayant présenté un AVC traité par voie endovasculaire. Critère de Jugement Principal : *modified Rankel Scale* à J90 variant de 0 à 6 si >2 bon pronostic si ≤2 mauvais pronostic	Réseaux neuronaux *Support Vector Machine*	- moyenne *modified Rankel Scale* 2,56 - régression linéaire à partir de 8 variables : précision de 43,5 % - *Artificial Neural Network* : coefficient de détermination de 0,79 mais AUC 0,6 - *Support Vector Machine* : précision de 51% pour les patients ayant un mauvais pronostic et 87 % pour les patients ayant un bon pronostic. Moyenne de la précision 0,79

3.3.2 *Machine Learning versus* régression logistique en médecine

Plusieurs études ont montré une supériorité du *Machine Learning* par rapport à la régression logistique pour établir un modèle prédictif. Le tableau V représente des exemples de ces études.

- En psychiatrie, pour prédire le pronostic d'un épisode dépressif majeur. L'objectif étant de mieux identifier, les patients présentant cette pathologie et ayant besoin d'un suivi au long cours (52).

- En anesthésie, pour prédire la survenue de complications dans les suites d'une chirurgie afin d'améliorer la stratification du risque opératoire (53).

- D'autres auteurs se sont intéressés au risque de dégradation clinique du patient durant une garde ou au risque de décès lié à un sepsis. L'objectif final étant de recentrer l'attention du clinicien sur les patients pouvant se dégrader et de diminuer le nombre de fausses alarmes (35)(54).

- En raison du manque de calibration des scores de mortalité habituellement utilisés en Unités de Soins Intensifs (APACHE II, SOFA), plusieurs équipes se sont intéressées à l'intérêt du *Machine Learning* pour prédire le pronostic en réanimation. Ces études ont démontré que le *Machine Learning* faisait aussi bien, si ce n'est mieux, que les scores pré existants, pour la plupart basés sur des régressions logistiques (55).

A l'ère des « *big data* » et de l'informatisation des dossiers médicaux, le *Machine Learning* représente donc une alternative légitime aux méthodes statistiques habituelles pour établir des scores de mortalité. Ce d'autant plus que le risque de décès est une variable dynamique, qui ne peut dépendre uniquement d'un nombre de variables explicatives arrêté, prédéfini, dont la relation linéaire avec le décès aura été déterminée en amont de l'analyse.

Tableau V : exemples d'application du *Machine Learning* en médecine versus la régression logistique

Discipline - Domaine	Objectif	Design de l'étude	Algorithmes utilisés	Résultats
Psychiatrie - Episode dépressif majeur Kessler *et al.* 2016 (52)	Prédire le pronostic dans les suites d'un épisode dépressif majeur	Enquête prospective au sein d'une cohorte de 1056 patients : - enquête 1 : questionnaire entre 1990 et 1992 pour développer le modèle - enquête 2 : entre 2001 et 2003 pour évaluer le pronostic à 10 ans.	10 fold cross validation Ensemble *regression trees* *Penalised regression* Régression logistique	En fonction des critères (récidive, chronicité, décès, hospitalisation, handicap fonctionnel, taux de suicide) : AUC entre 0,71 et 0,76 pour le *Machine Learning* (sauf pour chronicité 0,63) AUC entre 0,68 et 0,70 pour la régression logistique
Anesthésie-Chirurgie Thottakkara *et al* 2016 (53)	Prédire la survenue de complications post opératoires (insuffisance rénale aiguë, sepsis)	Etude rétrospective monocentrique sur 10 ans chez 50 318 patients toutes chirurgies confondues.	GAM SVM	Insuffisance rénale aigue 35 % Sepsis sévère 5 % SVM, GAM, régression logistique et modèle bayésien : - AUC entre 0,797 et 0,858 pour l'IRA - AUC entre 0,757 et 0,909 pour le sepsis sévère $p < 0,05$
Urgence - Détresse Vitale Churpek *et al.* 2016 (35)	Prédire la survenue d'une dégradation clinique lors d'une garde	Etude rétrospective multicentrique sur 5 ans chez 269 999 patients présentant des signes de détresse vitale Critère de jugement principal composite : survenue d'un décès, d'un ACR et d'un transfert en réanimation.	*Random Forest* Régression logistique	424 ACR 13 188 transferts en Réanimation 2 840 décès *Random Forest* AUC 0,81 Régression logistique AUC 0,735 $p = 0,01$

Suite tableau V : exemples d'application du *Machine Learning* en médecine versus la régression logistique

Discipline - Domaine	Objectif	Design de l'étude	Algorithmes utilisés	Résultats
Urgence - Sepsis Taylor *et al.* 2016 (54)	Prédire la mortalité hospitalière chez les patients admis aux urgences pour un sepsis	Etude rétrospective, multicentrique, pendant 1 an menée chez 5278 patients Cohorte de développement 80 % Cohorte de validation 20 % Modèle de ML construit à partir de 400 données issues du dossier médical électronique	*Random Forest* CART CURB 65 MEDS score Régression Logistique	AUC : p ≤ 0,003 *Random Forest* 0,860 Régression logistique 0,755 CURB 65 0,705 MEDS Score 0,705 Modèle CART 0,693
Réanimation - Score de mortalité Pirrachio *et al.* 2015 (55)	Prédire la mortalité en réanimation	Etude monocentrique menée pendant 7 ans chez 24 508 patients en unité de soins intensifs Cohorte de validation externe de 200 patients, Variables explicatives : celle du SAPS II et de APACHE	Modèle SICULA : combinaison de plusieurs algorithmes de ML SVM, *Gradient Boost*, *regression tree*, modèle bayesien, GAM et régression logistique	Décès à l'hôpital 3 002 = 13 % Prédiction mortalité : - SICULA : 13 % - SAPS II ajusté sur notre cohorte 13 % - APACHE ajusté 12 % - SOFA 12 % - SAPS II non ajusté : 30 % Discrimination : - SOFA : AUC 0,71 - SAPS II : AUC 0,78 - SICULA : AUC 0,85 au sein de la cohorte test et 0,94 au sein de la cohorte de validation externe

SICULA *Super Intensive Care Unit Learning Algorithm* **CURB 65** *Confusion Urea Respiratory Blood Pressure 65* **MEDS** *Mortality in Emergency Department Sepsis* **CART** *Classification and Regression Tree*

4. Comment évaluer l'intérêt pratique d'un modèle prédictif ?

Une fois le modèle développé, il faut déterminer s'il est performant ou non pour prédire la survenue de l'évènement étudié, soit pour ce qui nous concerne, le décès post-opératoire dans les suites d'une chirurgie cardiaque. Habituellement les modèles de prédiction sont évalués et comparés entre eux grâce à l'aire sous la courbe ROC qui permet de déterminer les performances de discrimination des modèles, mais ne permet pas d'évaluer leur intérêt pratique.

4.1 Courbe Receiver Operating Characteristic (ROC)

Les outils statistiques classiquement utilisés pour déterminer la performance d'un test sont représentés par ses valeurs intrinsèques : sa sensibilité et sa spécificité. Ces deux valeurs sont représentées par la courbe ROC. Le pouvoir discriminant du modèle est déterminé par l'AUC. Un test qui fait aussi bien que le hasard a une AUC de 0,5. Pour un test parfait, l'AUC vaut 1.

Cependant, cette méthode statistique habituelle de comparaison entre deux modèles présente des limitations sur lesquelles nous allons nous attarder (11).

La courbe ROC est directement basée sur les données observées chez les patients malades (décédés) et non malades (survivants), et non sur les risques prédits. Un changement dans les risques prédits peut donc ne pas avoir d'impact sur une courbe ROC.

Alors que la courbe ROC permet d'évaluer le pouvoir discriminant du test, il n'y a pas d'évaluation de la calibration (risque prédit/risque observé) et donc de la précision du test. On ne peut donc pas comparer les prédictions du modèle aux valeurs réelles observées au sein de la population.

Plusieurs études ont démontré que l'ajout d'un bio marqueur associé à un sur-risque significatif de la survenue d'un événement (OR > 1) peut avoir un impact faible ou nul sur l'AUC, et ce d'autant plus que l'AUC du modèle de prédiction est déjà élevée. Ainsi, des OR très importants sont nécessaires pour augmenter de façon significative l'AUC.

Par exemple, Pencina *et al.* se sont intéressés à l'ajout du HDL cholestérol (HDLc) au score de Framingramm pour la prédiction de la survenue d'un événement coronarien. Ils ont comparé le modèle de base sans le HDLc et le nouveau modèle incluant ce marqueur. Bien que hautement significatif (HR 0,65 et p<0,001), le HDLc ne modifiait pas de façon statistiquement significative l'AUC qui passait de 0,762 à 0,774 (56).

L'AUC de la courbe ROC est nécessaire à la validité interne d'une étude mais manque de pertinence clinique. Nous souhaitions étudier nos différents modèles prédictifs dans une réflexion de prise de décision.

4.2 Decision Curve Analysis (DCA)

La *Decision Curve Analysis* (DCA, traduit par « analyse de courbe de décision ») a été développée pour évaluer et comparer des modèles prédictifs en prenant en compte les conséquences cliniques des faux négatifs et des faux positifs, ainsi que le jugement du patient. Le but était d'intégrer le fait que certains traitements ont des conséquences plus ou moins graves, et que la volonté de s'exposer à un risque varie en fonction des personnes (57).

Vickers *et al.* ont développé pour la première fois la DCA au sein d'une cohorte de patients présentant un adénocarcinome de la prostate (12). Ils ont comparé trois modèles permettant de déterminer s'il fallait réaliser une exérèse de la vésicule séminale. L'ensemble des conséquences de cette décision a été identifié et décrit. D'un côté, les effets secondaires de l'exérèse de la vésicule séminale à savoir anéjaculation, incontinence urinaire ; ces effets sont d'autant « moins acceptables » chez les patients faux positifs (donc sur-traités). De l'autre côté, le risque de récurrence du cancer de la prostate si la vésicule séminale est laissée en place ; c'est l'effet délétère chez les faux négatifs (donc patients sous-traités).

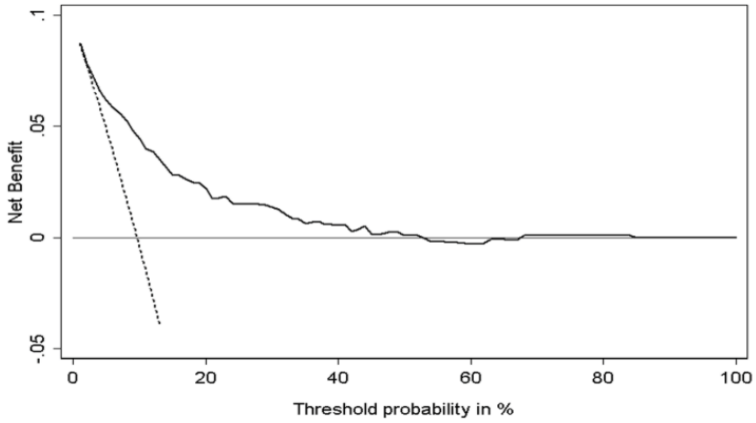
Dans ces conditions, il est légitime de se demander quel est le seuil à partir duquel un patient donné et un praticien donné jugent acceptable/nécessaire l'intervention chirurgicale ? C'est à ce moment qu'intervient le jugement du patient. En effet, toute DCA implique de déterminer un seuil de probabilité de développement de la maladie/de survenue de l'évènement d'intérêt au-delà duquel le patient optera pour le traitement. Ce seuil de probabilité *pt* permet de déterminer de quelle façon le patient pondère le préjudice lié au risque de faux négatifs ou de faux positifs.

La DCA est schématisée par un graphique comprenant :

- en ordonnée : le bénéfice net (*Net Benefit*) qui correspond à la différence entre le nombre de vrais positifs et le nombre de faux négatifs, pondéré par le potentiel préjudice lié à un résultat faussement positif ou faussement négatif,
- en abscisse : le seuil de probabilité sélectionné.

La figure 1 représente une analyse par courbe de décision permettant d'évaluer l'intérêt d'un modèle prédictif de cancer pour décider ou non d'un traitement.

Figure 1 : Exemple de DCA



En pointillé : « *treat all* »
En gris clair, trait horizontal : « *treat none* »
En noir : performance du nouveau modèle

Lors de la représentation d'une DCA, en plus de la performance du modèle, sont toujours représentées les courbes de stratégie traitement pour personne (« *treat none* ») de bénéfice net toujours nul et de stratégie traitement pour tous (« *treat all* »). Donc si le nouveau modèle prédictif n'obtient pas un bénéfice net supérieur à la courbe « *treat all* » alors le test diagnostique ou pronostique n'apporte pas d'information suffisante pour améliorer les soins.

Sur ce graphique, pour des seuils probabilité entre 0 et 5 % la stratégie « *treat all* » fait aussi bien que le nouveau modèle. Dans ce rang de *pt* faible, les patients sont plus préoccupés par le risque de manquer le diagnostic de cancer que par le risque d'un « sur traitement » (donc de faux positifs). Pour les seuils de probabilités plus haut (à partir de 60%), le bénéfice net du modèle se rapproche de 0, les patients sont plus préoccupés par le risque d'une chirurgie non nécessaire que par le risque de sous-diagnostiquer le cancer (donc de faux négatifs). Le modèle prédictif n'est alors pas utile, la stratégie sans traitement est retenue.

Dans cet exemple, si le seuil est de 10%, le bénéfice net est de 0,05 soit 5% ; c'est-à-dire que 50 patients sur 1 000 qui ne risquaient pas de développer la maladie d'intérêt ne seront pas traités, sans diminuer le nombre d'interventions réellement nécessaires (soit 50 faux positifs en moins pour 1000 patients sans diminuer le nombre de vrais positifs).

Pour les auteurs, cette méthode de comparaison, prenant en compte le jugement du patient (et/ou du praticien), est un outil d'aide à la décision pertinent, tout particulièrement, dans des situations au cours desquelles :

- les traitements présentent des effets secondaires importants,
- le choix de traiter ou non a des conséquences importantes sur la morbi-mortalité,
- la prise de décision est difficile.

La plupart des études comparant des modèles prédictifs ou diagnostiques grâce à DCA, ont été menées en oncologie, mais la chirurgie cardiaque remplit tout autant ces critères (58-60).

III – PATIENTS ET METHODE

1. Design de l'étude et patients

Il s'agit d'une étude rétrospective, observationnelle, monocentrique, réalisée à partir d'une base de données prospective du Centre Hospitalier Universitaire Bichat à Paris.

Les données de l'ensemble des patients admis pour chirurgie cardiaque sous CEC ont été recueillies. Toutes les interventions consécutives ayant eu lieu entre le 1er janvier 2006 et le 31 décembre 2012 ont été incluses au sein de cette base de données institutionnelle, y compris les actes chirurgicaux pratiqués en urgence et les chirurgies pour cardiopathie congénitale de l'adulte.

Pour la réalisation de notre étude, les patients admis pour transplantation cardiaque ou chirurgie en urgence ont été exclus de cette base de données.

2. Recueil de données

La mise en place de cette base de données électronique a été déclarée auprès de la CNIL. Cette étude a été approuvée par le comité éthique local : Comité d'Evaluation de l'Ethique des projets de Recherche Biomédicale (CEERB) - Paris Nord Institutional Review Board 00006477, Paris 7 University, AP-HP.

Les données ont été recueillies, de façon prospective, sur formulaire informatique complété en amont de l'intervention puis au fur et à mesure de l'hospitalisation par les différents intervenants.

Recueil des données relatives à la chirurgie :

- Caractère urgent ou non de la chirurgie,
- Antécédents de chirurgie cardiaque,

– Le type de geste réalisé : pontage aorto-coronarien, chirurgie valvulaire aortique mitrale ou tricuspide, chirurgie de l'aorte thoracique,

– Poids de l'intervention : nombre de gestes dans le même temps opératoire / chirurgie combinée ou non.

Recueil des données relatives aux patients, consultables en annexe 1 : caractéristiques épidémiologiques, antécédents, état clinique, traitements, et enfin résultats d'examens complémentaires.

Le critère de jugement principal a été la mortalité intra hospitalière. Tous les décès ont été pris en compte quelle que soit leur cause et quel que soit leur délai de survenue en cours d'hospitalisation.

L'EuroSCORE I et II n'ont pas été saisis par les investigateurs mais directement calculés à partir de la base de données. Pour tous les patients, les soins péri-opératoires étaient standardisés, incluant l'anesthésie, les techniques de monitorage et le maintien en normothermie pour les pontages aorto-coronariens.

3. Développement et validation du modèle

Comme lors de toutes les études de développement de scores prédictifs, nous avons divisé notre échantillon en deux cohortes : une cohorte d'apprentissage et une cohorte de validation du modèle, représentant respectivement 70 et 30 % de la population.

Afin de minimiser les écarts à la moyenne, nous avons également réalisé une k=5 *cross fold validation*. La population de notre étude a donc été divisée en cinq échantillons de taille équivalente. L'un de ces 5 échantillons a été sélectionné comme ensemble de validation et les k-1, soit 4 échantillons restants comme ensemble d'apprentissage du modèle. Chaque modèle était donc testé k=5 fois. La répétition de cette analyse a permis d'homogénéiser la répartition des données entre apprentissage et test.

Pour chaque modèle testé, le processus a été répété dix fois, de telle sorte que nous obtenions 10 probabilités individuelles. Nous avons réalisé la moyenne de ces dix mesures et considéré cette dernière comme la probabilité individuelle finale de chaque modèle.

4. Régression logistique

Nous avons appliqué l'EuroSCORE I et l'EuroSCORE II à notre population de patients (17)(4). Nous avons exclu de notre analyse les patients nécessitant une chirurgie cardiaque en urgence car les données de la base ne permettaient pas une cotation exacte. Un modèle complémentaire a été réalisé par régression logistique à partir des variables de l'EuroSCORE II au sein de notre population. Ce modèle a été appelé dans la suite de l'étude modèle de régression logistique.

5. Machine Learning

Nous avons utilisé plusieurs types d'algorithmes de *Machine Learning* et nous avons créé un dernier modèle dit « *Machine Learning* » fait de l'assemblage de plusieurs algorithmes de *Machine Learning* (61) (34) (37).

5.1 Random Forests (RF)

Les « *Random Forests* », littéralement forêts aléatoires, représentent un type d'algorithme d'apprentissage supervisé, permettant de résoudre des problèmes de régression (variable à étudier quantitative) et de classification (variable à étudier qualitative, comme dans notre étude). Cet algorithme est constitué d'une multitude d'arbres décisionnels indépendants construits de façon aléatoire et organisés en forêt (62).

Ainsi chaque arbre a une vision parcellaire du problème puisque les variables explicatives et les échantillons sont sélectionnés aléatoirement. Chaque nœud de l'arbre décrit la distribution de la variable à expliquer : les différents échantillons sont partitionnés en fonction des variables explicatives qui sont autant d'attributs décisionnels. Toute la difficulté réside dans le choix de la hiérarchie des variables de décision puisque les premiers attributs décisionnels doivent être les plus discriminants.

La combinaison de ces modèles d'apprentissage (les arbres décisionnels) permet de diminuer la variance du modèle final et de minimiser le risque d'erreur et de sur apprentissage (62).

5.2 Gradient Boosting Machine (GBM)

Le GBM, au même titre que le *Random Forest*, est une méthode ensembliste d'apprentissage supervisé permettant de répondre à des problèmes de classification et de régression.

Cet algorithme réunit un ensemble d'arbres décisionnels qui contrairement au RF ne sont pas indépendants les uns des autres, puisqu'ils sont reliés en série. Le principe est de réunir, *via* ces arbres, des algorithmes « faibles » c'est à dire faisant un peu mieux que le hasard. Leur combinaison va permettre d'améliorer l'algorithme afin d'aboutir à un modèle robuste.

Ainsi chaque algorithme faible va apprendre des erreurs faites par l'algorithme (et donc l'arbre décisionnel) précédent afin de ne pas les reproduire et d'améliorer la prédiction du modèle. Les modèles suivants peuvent donc prédire les erreurs résiduelles des précédents algorithmes. La décision finale pour les classes à attribuer au décès post opératoire sera une somme pondérée des différents algorithmes (ceux réalisant le moins d'erreur ayant les coefficients les plus importants) ; contrairement au *Random Forest* où chaque arbre décisionnel a le même poids.

5.3 Support Vector Machine (SVM)

Le SVM (*Support Vector Machine* ou encore Séparateur à Vaste Marge) est un algorithme d'apprentissage supervisé permettant de répondre à des problèmes de classification ou de régression, linéaire ou non. L'objectif du SVM est de tracer un hyperplan séparateur optimal, une frontière de décision permettant de définir des groupes homogènes pour une variable étudiée, au sein d'une population.

Considérons trois hyperplans séparant de façon linéaire une population en deux classes (dans notre cas décès ou non). Par convention, les points situés au-dessus de l'hyperplan sont positifs, ceux situés en dessous appartiennent à la classe négative. Les trois hyperplans séparateurs identifient les deux classes avec la même précision. Leur capacité de généralisabilité les différencie.

Le meilleur algorithme est celui dont la marge est la plus importante et donc dont l'hyperplan est le plus éloigné de l'ensemble des données du problème. Il s'agit de la fonction y=0 qui a la meilleure généralisation. La maximisation de la marge permet en effet de diminuer le risque de sur apprentissage.

Ce classifieur est performant mais le processus de traitement des données peut être long et complexe, surtout dans le cas de variables non linéaires en grande dimension.

5.4 Classification Naïve Bayesienne

Le *Naïve Bayes* est une méthode d'apprentissage supervisée qui repose sur une hypothèse simplificatrice forte : les descripteurs (et donc les variables explicatives) sont indépendants entre eux. Il s'agit de l'un des classifieurs linéaires les plus simples à utiliser. Pour autant ce modèle est robuste et efficace. La question sur laquelle repose ce modèle est la suivante : *considérant que l'on m'a fourni un échantillon de données, comment ces connaissances peuvent-elles modifier mes connaissances antérieures sur le monde ?*

Le théorème de Bayes permet de transformer une probabilité *a priori* d'une classe y (le décès) en une probabilité *a posteriori* en fonction des données de l'apprentissage et donc à l'aide des informations contenues dans nos variables explicatives.

Le modèle bayesien naïf est très utilisé par les chercheurs. Il s'agit en effet d'un modèle facile à programmer et la construction du modèle est rapide sur des bases de données importantes en nombre de variables et d'individus. En revanche, en pratique, les analystes ont peu recours à cet algorithme du fait de la difficulté d'interprétation des résultats. Il n'y a pas de règle d'affectation explicite du modèle prédictif (37).

5.5 Sélection des variables

Certains de ces modèles sont sensibles à la corrélation existante entre les variables explicatives d'entrée. Dans notre étude, afin de minimiser ce problème, nous avons ajusté deux fois l'ensemble des modèles de *Machine Learning* sus cités : R*andom Forest, Gradient Boosting Machine, Support Vector Machine* et *Naïve Bayes model*.

Lors du premier ajustement, nous avons conservé toutes les données explicatives pour l'analyse. Lors du second ajustement, nous avons réalisé un test du Chi 2 et utilisé uniquement les données pertinentes et significatives pour chacun des quatre modèles.

5.6 Modèle Machine Learning

Il a été démontré que le fait d'assembler les résultats de différents algorithmes de *Machine Learning* permettait d'améliorer les performances d'un modèle prédictif en comparaison à un modèle basé sur un seul algorithme (55). A cet effet, nous avons réalisé une moyenne pondérée des probabilités de décès obtenues pour chaque modèle de *Machine Learning*.

6. Performance des modèles

6.1 Courbes ROC

Les capacités de discrimination obtenues par l'EuroSCORE I, l'EuroSCORE II, le modèle « Régression Logistique » et les différents modèles de *Machine Learning* ont été comparés grâce à la mesure de l'Aire Sous la Courbe (AUC) ROC et son intervalle de confiance à 95%.

6.2 Decision Curve Analysis (DCA)

Nous avons également comparé les capacités de prédiction de l'EuroSCORE II, du modèle « régression logistique » et du modèle « *Machine Learning* » par une DCA.

7. Analyse statistique

Les variables qualitatives ont été exprimées en fréquences et pourcentages. Les variables quantitatives ont été exprimées en moyennes et écarts types ou médianes et écarts interquartiles.

Les comparaisons entre deux pourcentages ont été réalisées grâce à un test du Chi 2 ou par le test exact de Fisher. La comparaison entre deux moyennes a été réalisée par un test de t Student ou par un test de Mann and Withney. Nous avons également réalisé un test de comparaison des AUC.

Les DCA ont été réalisées en utilisant le logiciel SAS grâce aux macros fournies par Vickers *et al* (12). Une valeur de p inférieure ou égale à 0,05 était considérée comme statistiquement significative.

Toutes les analyses ont été réalisés en utilisant le logiciel SAS 9.4 (SAS Institute, Inc, Cary, NC, USA) et le logiciel R version 3.2.2 (*The R Foundation for Statistical Computing*, Vienne, Autriche) avec les extensions XGBoost, Extra Trees et e1071.

IV – RESULTATS

1. Caractéristiques de la population

De décembre 2005 à décembre 2012, 6 889 interventions de chirurgie cardiaque sous CEC ont été réalisées au CHU Bichat. Parmi ces patients, 369 soit 5,3% ont été admis pour chirurgie cardiaque en urgence et ont donc été exclus de l'analyse. Notre cohorte comportait donc 6 520 patients.

Les caractéristiques des patients sont consultables dans l'annexe 2. L'âge moyen était de 63,4 ans. Le taux de mortalité observé durant le séjour à l'hôpital était de 6,3 % soit 411 décès. Le taux de chirurgie coronarienne isolée était de 38,4%. Il y avait 61,6% de chirurgie valvulaire ou combinée.

En analyse univariée, parmi les 66 variables explicatives recueillies, 51 sont reliées de façon statistiquement significative au décès post opératoire :

- âge, poids, taille, IMC,
- 20 variables liés aux comorbidités et antécédents,
- 5 variables liées à l'état clinique pré opératoire,
- 4 variables liées aux traitements pré opératoires,
- 8 variables liées aux données paracliniques pré opératoires,
- 5 variables liées à la coronarographie pré opératoire,
- 5 variables liées aux caractéristiques de la chirurgie.

L'EuroSCORE II moyen était de 3,7 (4,8) %.

2. Analyse de la discrimination des différents modèles : comparaison courbes ROC

L'analyse des courbes ROC des modèles EuroSCORE I, EuroSCORE II, du modèle de régression logistique, des différents algorithmes de *Machine Learning* et du modèle de *Machine Learning* pour la prédiction de la mortalité post opératoire de chirurgie cardiaque est représentée dans la figure 1. Les aires sous les courbes ROC sont résumées dans le tableau VI.

Tableau VI : Aires sous la courbe ROC des modèles de prédiction

	AUC	IC 95 %
EuroSCORE I	0,719	0,674-0,763
EuroSCORE II	0,737	0,691- 0,783
Modèle de Régression Logistique (covariables de l'ES II)	0,742	0,698-0,785
Algorithmes de *Machine Learning* sans filtre		
Gradient Boosting Machine	0,786	0,748-0,826
Random Forests	0,786	0,747-0,825
Modèle Naïf Bayésien	0,734	0,689-0,779
Support Vector Machine	0,753	0,710-0,797
Algorithmes de *Machine Learning* avec le filtre du Chi2		
Gradient Boosting Machine	0,784	0,743-0,824
Random Forests	0,788	0,748-0,827
Modèle Naïf Bayésien	0,750	0,708-0,793
Support Vector Machine	0,736	0,689-0,784
Ensemble des algorithmes de ML : modèle ML	0,795[1]	0,755-0,834

[1]p< 0,001 après comparaison du ML à l'ES II et au modèle Régression Logistique

ES EuroSCORE, **ML** Machine Learning

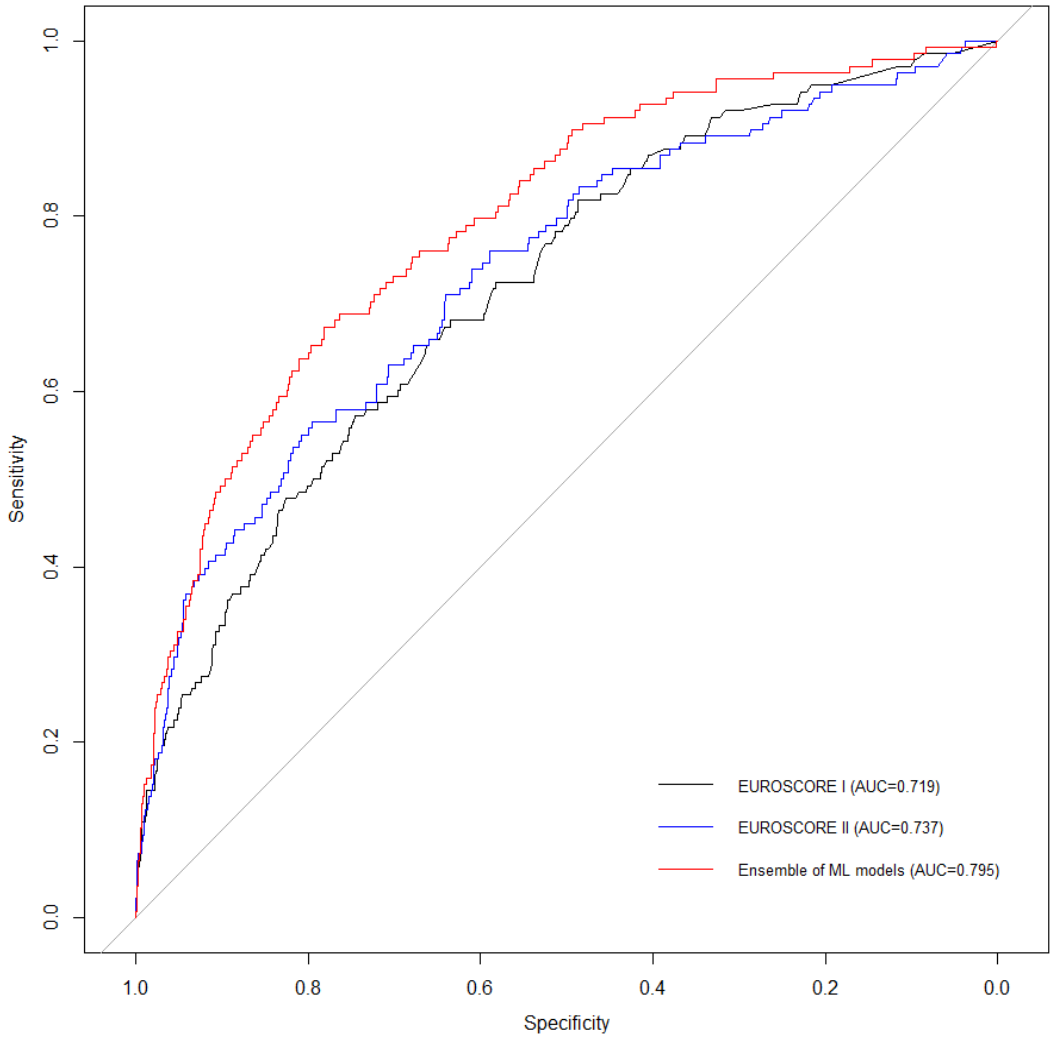
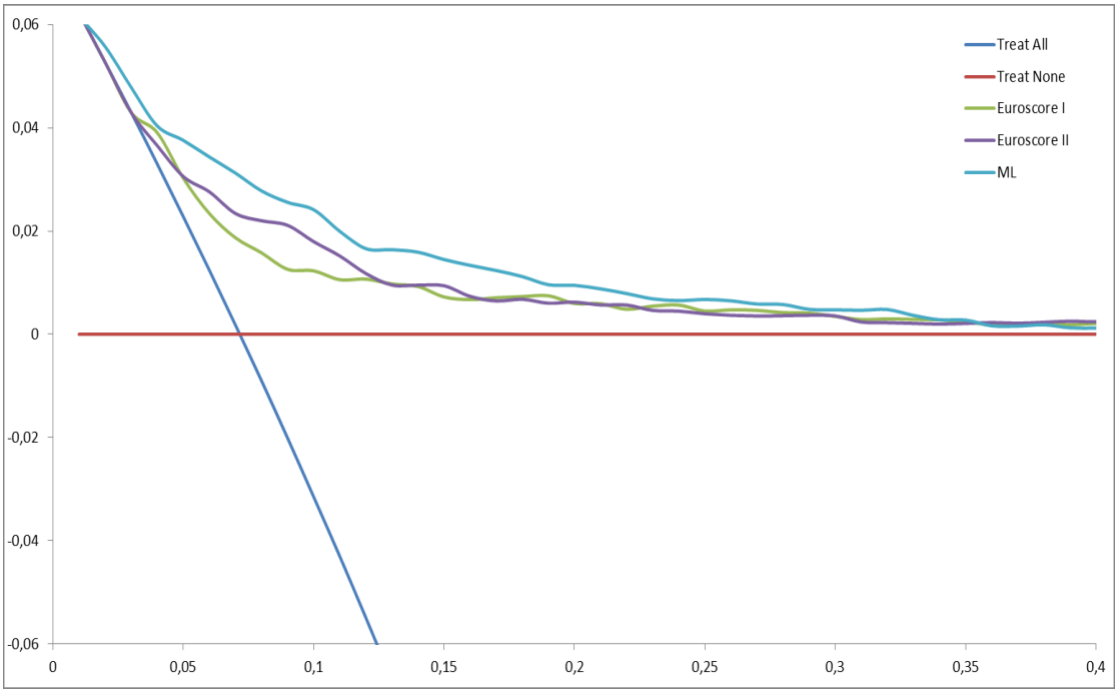


Figure 2 : Courbes ROC des modèles EuroSCORE I, EuroSCORE II et *Machine Learning* pour prédire la mortalité hospitalière après chirurgie cardiaque programmée

Le modèle *Machine Learning* possède la meilleure discrimination avec une AUC de 0,795 (0,755- 0,834). Concernant les différentes techniques de *Machine Learning*, avec ou sans l'utilisation du filtre par un test du Chi 2, le *Random Forest* a la meilleure AUC : respectivement 0,788 (0,748-0,827) et 0,786 (0,747-0,825). Par ailleurs, les modèles de *Machine Learning* étudiés séparément ont des aires sous la courbe ROC plus importantes que l'EuroSCORE II ou le modèle de régression logistique.

3. Decision Curve Analysis



En abscisse : **seuil de probabilité pt de survenue du décès** en post opératoire au-delà duquel le patient déciderait de ne pas être opéré
En ordonnée : **Bénéfice Net** du modèle évalué
En bleu : bénéfice net lié au fait de **traiter tous les patients** en supposant qu'ils survivent tous
En rouge : bénéfice net = 0 lié au fait de **ne traiter aucun patien**t en supposant qu'ils décèdent tous

Figure 3 : DCA des modèles EuroSCORE I, EuroSCORE II et *Machine Learning* pour prédire la mortalité hospitalière après une chirurgie cardiaque au sein de la cohorte de validation

L'interprétation de cette courbe montre qu'au-delà de 40 % de risque de décès post opératoire, le bénéfice net est égal à 0. A partir du seuil de probabilité de décès de 10%, le bénéfice du modèle *Machine Learning* par rapport aux autres modèles est moindre.

Par ailleurs, quel que soit le seuil de probabilité choisi, jusqu'au seuil de 35% de décès post opératoire le modèle *Machine Learning* comporte un bénéfice net plus important que l'EuroSCORE I, l'EuroSCORE II et la stratégie « *treat all* ». La différence est d'autant plus importante pour les faibles risques de décès post opératoire notamment entre 1 et 6% de seuil de probabilité pt. Au-delà de 35% de seuil de probabilité choisi, les courbes de

décision se confondent et le Bénéfice Net se rapproche de 0 (soit stratégie « *treat none* »).

Les courbes de l'EuroSCORE I et II demeurent très proches en fonction du seuil de probabilité choisi. Le Bénéfice Net lié à l'EuroSCORE II est meilleur uniquement entre 7 et 14% de *pt*.

Entre 0 et 5% de prévision de décès, la stratégie « *treat all* » fait aussi bien que l'EuroSCORE I et II. Par la suite le Bénéfice Net de cette stratégie s'effondre pour atteindre 0 dès 8% de seuil de probabilité.

V – DISCUSSION

1. Retour sur les résultats principaux et forces de l'étude

A notre connaissance, notre étude évalue pour la première fois un modèle de *Machine Learning* par *Decision Curve Analysis* en médecine et tout particulièrement dans le contexte du risque de mortalité en chirurgie cardiaque programmée.

Les caractéristiques démographiques de la population de notre étude sont comparables à celles des principaux scores en vigueur. En effet, l'âge moyen dans notre cohorte était de 64 ans, celui dans la cohorte EuroSCORE II était de 67 ans et celui dans la cohorte du STS score était de 64,7 ans. Le ratio homme/femme était comparable entre l'EuroSCORE II et notre cohorte, à environ 2,3 (4).

Au sein de notre étude, le taux de mortalité durant le séjour à l'hôpital a été de 6,7%, ce qui parait élevé par rapport au taux de mortalité à 30 jours observé dans la cohorte de l'EuroSCORE II qui était de 3,9% (4). Par ailleurs, l'étude de Barili *et al.* menée de façon prospective et multicentrique pendant 6 ans chez 12 325 patients montrait un taux de mortalité durant le séjour intra hospitalier de 2,2% (14). La cohorte rétrospective, monocentrique, de Dedda *et al* réunissant 1 090 patients montrait un taux de mortalité à J30 à 3,5 %, inférieur au nôtre, alors même que la durée de suivi de cette étude s'arrêtait à 30 jours post opératoires (22).

Cette différence peut s'expliquer par une proportion de chirurgie valvulaire plus importante au sein de notre cohorte qu'au sein de la cohorte de l'EuroSCORE II (56 *versus* 46,3%) ainsi qu'un taux de pontages aorto-coronariens moins élevé (38,4% *versus* 46,7%) (4). Concernant la cohorte de Barili, la tendance est la même puisque le taux de pontage aorto-coronarien était de 52,4% *versus* 43,5 % de chirurgie valvulaire (14). En revanche, l'étude de Dedda montre des taux similaires de chirurgie de revascularisation et de chirurgie valvulaire (respectivement 34,1% *versus* 35,4%) (22).

L'EuroSCORE II moyen était de 3,7 % au sein de notre étude donc comparable à celui de la cohorte ayant servi à l'élaboration de l'EuroSCORE II (3,9%) malgré l'exclusion des patients opérés pour une urgence.

A noter que le taux de chirurgie en urgence au sein de la cohorte de l'EuroSCORE II était de 23,3%, bien plus élevé que dans les autres cohortes. En effet, ce taux était de 3,5% dans la cohorte de Barili *et al* et de 3,9% dans la cohorte de Dedda *et al.* Il était de 5,3% dans notre cohorte (4)(14)(22).

Notre travail a montré que, quelle que soit la méthode d'évaluation que nous avons appliquée, courbe ROC ou DCA, le modèle de *Machine Learning* était plus performant que l'EuroSCORE II pour prédire la mortalité intra hospitalière après une chirurgie cardiaque programmée. Dans le souci d'éviter un biais de sélection et afin de ne pas désavantager la régression logistique par rapport au *Machine Learning*, nous avons créé un modèle de régression logistique qui définissait un risque de mortalité, en fonction des variables et coefficients de régression de l'EuroSCORE II, appliqués à notre cohorte. Mais à nouveau, le modèle *Machine Learning* avait une meilleure discrimination. Notons que les AUC de l'EuroSCORE 2 et de notre modèle par régression logistique sont très proches.

L'EuroSCORE II prend en compte un nombre de variables limité et déterminé. A chacune de ces variables est relié un coefficient de régression non modifiable. En chirurgie cardiaque, certaines situations nécessitent de développer une approche individuelle pour une stratification personnalisée du risque opératoire. Mais certaines variables associées au risque de décès ne sont pas prises en compte par l'EuroSCORE II (hypo albuminémie, dysfonction ventriculaire droite, syndrome de fragilité du sujet âgé, etc.) (63). Il s'agit en fait d'un modèle figé, non évolutif, là où le *Machine Learning* est probablement capable de flexibilité, d'adaptation, et de correction dans sa prédiction.

Cependant la régression logistique apporte certains avantages : notamment elle permet d'établir des liens entre le décès et les variables explicatives retenues dans le modèle final. Elle établit la force du lien qui associe un critère de jugement principal à différentes variables, *via* les coefficients de régression (31). Toutefois, des méthodes de *Machine Learning* se développent et permettent de hiérarchiser l'importance des variables explicatives.

1.1 Le Machine Learning en anesthésie-réanimation

L'avènement du *Machine Learning* est directement lié au développement des dossiers médicaux électroniques et des bases de données patients ainsi qu'aux avancées informatiques (9). En médecine, de plus en plus de modèles prédictifs sont établis grâce à des techniques de *Machine Learning*. Le nombre d'études disponibles à ce sujet est grandissant. Nos résultats sont cohérents avec ceux retrouvés dans la littérature.

Notre étude montre une supériorité du *Machine Learning* dans une application bien précise : la prédiction du décès dans les suites d'une chirurgie cardiaque programmée. Plusieurs études ont déjà montré la supériorité du *Machine Learning* pour la prédiction d'évènements (tels que le décès ou la survenue d'une maladie) dans le domaine de l'urgence et des soins intensifs :

- prédiction de la mortalité dans les suites d'un sepsis aux urgences, la mortalité en réanimation, le risque de dégradation clinique durant une garde ou encore le risque d'insuffisance rénale aiguë ou de sepsis post opératoires (35) (53,54).

- prédiction de la durée du séjour en unité de soins intensifs. Parmi elles, une étude a montré une AUC de 0,82 pour un algorithme de *Machine Learning* basée sur les données du score SOFA afin de prédire le risque de décès en réanimation. Ce modèle permettait de prédire la durée du séjour avec une erreur moyenne de 1,79 jours. Le *Machine Learning* permettrait donc de prévoir de façon plus précoce et précise le nombre de lits disponibles et donc d'améliorer la logistique de gestion des lits. Cette stratégie a donc un impact sur le coût et sur l'organisation du service (64).

Plusieurs équipes ont travaillé sur la création de supports décisionnels informatiques afin d'améliorer le processus décisionnel dans le contexte d'urgence, notamment traumatique. Des essais ont montré que le *Machine Learning* était supérieur aux méthodes statistiques habituelles pour prédire la nécessité d'une intervention de sauvetage chez des patients polytraumatisés aux urgences et en pré hospitalier. Il s'agissait de petits échantillons et les différences d'AUC était de l'ordre de 5% mais les auteurs abordent le traitement de données à haute fréquence, en temps réel dans des conditions d'urgence, avec pour objectif

in fine d'améliorer le triage et la prise en charge des patients polytraumatisés (65,66).

Le traitement rapide et en temps réel de ces données, afin de déterminer des conduites à tenir en situation d'urgence, est désormais une option envisageable du fait de l'évolution informatique et de la possible délocalisation des données (67).

Par ailleurs, le *Deep Learning,* qui est la branche la plus avancée du *Machine Learning*, se développe de plus en plus en médecine. Il s'agit d'un système d'apprentissage composé d'un ensemble de réseaux de neurones artificiels qui apprennent à représenter le monde, en tentant de modéliser au plus près le fonctionnement du cerveau humain (68).

Ce modèle fonctionne en couches successives qui permettent de décomposer le problème de classification. Les premières couches extraient des caractéristiques simples (des contours, des formes) que les couches suivantes utilisent pour former des concepts de plus en plus complexes (motifs puis image, etc.). Par opposition aux techniques de *Machine Learning* « classiques », nul besoin d'un travail de programmation manuelle préalable pour déterminer quelles variables vont permettre de transformer les données brutes en connaissances. Un système de *Deep Learning* étant capable de décomposer son apprentissage, il découvre lui-même ces caractéristiques d'intérêt (68,69).

Cette branche de l'IA fait partie intégrante des outils informatiques avancés à disposition des professionnels de santé afin d'aider à l'interprétation des imageries médicales. A notre connaissance, il n'y a pas d'étude publiée sur l'application du *Deep Learning* en anesthésie réanimation ou en médecine d'urgence (69-71).

Le développement d'algorithmes de *Machine Learning* non supervisés constitue une étape supplémentaire dans le développement de support informatique d'aide à la décision. Ce type d'algorithme serait alors capable d'apprendre, de façon autodidacte à partir de données brutes, sans nécessité d'étiquetage préalable des données (68).

Ainsi, les perspectives d'application de l'IA en médecine et plus particulièrement en anesthésie-réanimation sont grandes, permettant d'améliorer la prise en charge des patients et la consommation des ressources.

1.2 Decision Curve Analysis et courbe ROC

Le second objectif de notre travail était de fournir un mode d'évaluation, de notre modèle prédictif, complémentaire à l'évaluation traditionnelle basée sur les courbes ROC, et plus proche des préoccupations des cliniciens. L'utilisation de DCA était donc pertinente afin de prendre en compte l'impact clinique de la décision d'opérer ou non le patient ainsi que le jugement du patient et du clinicien pour la prise de décision (57)(12).

L'analyse de nos courbes de décision montrait un bénéfice net toujours supérieur pour le *Machine Learning*, qui toutefois restait modéré. Jusqu'au seuil de décès de 30%, le bénéfice net du modèle *Machine Learning* était de 1 à 6% : ce qui signifie que 10 à 60 patients sur 1 000 qui décéderaient après une chirurgie cardiaque ne seraient pas opérés, sans pour autant diminuer le nombre de chirurgies sans risque de décès. Ainsi même si le bénéfice en termes de pourcentage peut sembler faible, l'implication clinique rend ce bénéfice considérable. Les scores de prédiction étant recommandés par les sociétés savantes, il est logique de retenir le meilleur modèle (3).

La plupart des analyses par courbes de décision a été menée en oncologie. Seules deux études ont été menées, en dehors de cette discipline, à savoir une étude sur le pronostic du sepsis aux urgences et une autre menée sur la fin de vie en réanimation (73,74).

La technique d'analyse par DCA présente également des limitations. Il est ainsi nécessaire de disposer :

- des données individuelles de chaque patient : préférences personnelles, stade exact de la maladie, etc.
- d'une évaluation explicite des résultats de l'intervention d'intérêt sur la santé : combien de complications prévenues ? D'années de vie gagnées ? D'années de vie gagnées corrigées sur la qualité de vie ? etc (57).

Concernant les pathologies pour lesquelles plusieurs traitements sont disponibles, les seuils de probabilité peuvent changer en fonction des alternatives thérapeutiques. Il faudrait donc construire un modèle pour chacune d'entre elles. Pour certains auteurs, la vision probabiliste de ces modèles peut ajouter à la complexité du processus décisionnel. Et pour finir, la DCA ne prend pas en compte le coût des différents traitements envisagés (60)(12).

La comparaison de tests diagnostiques ou pronostiques *via* une DCA permet de prendre en compte l'impact clinique du processus décisionnel. L'objectif étant d'optimiser la stratégie thérapeutique et donc de maximiser le ratio bénéfice/risque ou encore l'utilité clinique attendue d'un modèle pour un patient. Il peut également s'agir d'un support informatique d'aide à la décision impliquant chaque patient de façon personnalisée ainsi que le médecin décideur.

Néanmoins, Il faut garder à l'esprit que l'analyse par courbe ROC reste une des garantes de la validité interne de l'étude et donc du test. Ces deux techniques sont complémentaires l'une de l'autre pour l'évaluation, la comparaison et le développement de modèles prédictifs (12)(75).

L'émergence des techniques d'analyse par courbe de décision s'inscrit dans une dynamique de précision et de personnalisation, d'individualisation de la médecine que nous retrouvons avec l'émergence du *Machine Learning* (76,77).

Le processus décisionnel est multifactoriel et inclut des facteurs humains. Aucun score de mortalité, évalué par quelque méthode que ce soit, ne saurait se substituer à la relation et à la communication entre médecin et patient. Notre modèle de *Machine Learning*, plus performant que les scores classiques, n'est donc qu'un support d'aide complémentaire au jugement clinique des principaux protagonistes (personnel soignant et patient).

2. Perspectives de l'Intelligence Artificielle en médecine

La croissance des données en santé est exponentielle : plus de 5 milliards d'objets connectés (télémédecine, glucomètre, téléphone portable, tensiomètre, etc) sont recensés. *L'American Medical Informatics Association* prédit que d'ici 2050 le volume de données disponible en santé devrait être multiplié par 50. Les supports de stockage de l'information médicale sont vastes, les données venant du monde réel sont complexes, variables et entachées de bruit. La multiplicité et l'hétérogénéité des sources et de la nature de ces « *big data* » ne facilitent pas leur exploitation (40) (78).

Ainsi, l'analyse d'un volume important de données peut favoriser la propension au sur apprentissage. Le sur-apprentissage caractérise un modèle qui décrit précisément les données disponibles mais n'est capable d'aucune généralisation (37).

Afin de diminuer le risque de sur-apprentissage, plusieurs auteurs utilisent des méthodes de réduction des données selon le principe de parcimonie. L'objectif étant de conserver le maximum d'informations avec le minimum de variables. Il s'agit de procédés de pré traitement des données mathématiques et informatiques complexes, que nous ne détaillerons pas ici (52,53)(79-81).

Sélectionner les variables pertinentes pour la prédiction d'un événement est un des enjeux majeurs du *Machine Learning*. Cette technique permet non seulement de diminuer le risque de sur-apprentissage mais également d'améliorer les performances prédictives des modèles, de diminuer le coût et d'augmenter la rapidité d'exécution des algorithmes (53). Cette stratégie permet enfin d'améliorer la compréhension du processus ayant permis de générer les données (82).

La modification de la nature des données et l'exploitation des « *big data* » expliquent en partie le défaut d'interprétabilité de certains modèles de *Machine Learning*. Désormais les explorations statistiques remontent des conséquences (observation de milliers de comportements individuels) vers les estimations des causes probables (induction) et non plus l'inverse. De plus, les algorithmes utilisés sont complexes : *Support Vector Machine, Random Forest, Deep Learning*. Or, plus un modèle est complexe, plus il est précis mais plus il est difficile pour ses programmateurs de l'expliquer (83).

Si la méthode de réduction des données permet de rendre un modèle de *Machine Learning* moins opaque, d'autres auteurs préconisent l'utilisation d'une IA hybride, associant une technique d'IA symbolique au *Deep Learning*, appelée *Deep Renforcement Learning*. L'IA symbolique consistait en l'accumulation de règles logiques qui pouvaient être généralisées (84). Le problème étant que les règles logiques étaient issues du raisonnement des programmateurs et non de l'observation du monde réel. Grâce au *Deep Learning*, les règles logiques sont inférées à partir du monde réel, à partir de données brutes qui sont ensuite interprétées par une technique d'IA symbolique. Cette stratégie permet d'améliorer les performances d'un modèle. Ce dernier serait plus transparent et moins complexe qu'une technique de *Machine Learning* avec une capacité d'apprentissage plus rapide (85).

Le *Machine Learning* est également confronté à de nombreuses questions éthiques : consentement des patients, respect de la vie privée et du secret médical, protection des données, partage de l'information médicale, etc (86-88).

La multiplicité des supports d'informations et la dématérialisation des données médicales, dans le respect de la réglementation et de la confidentialité, favorisent un partage à grande échelle des données individuelles afin d'optimiser la recherche en santé. Cependant l'exploitation des données numériques n'est pas envisageable sans éthique et l'un des défis actuels du *Machine Learning* est de faire face et d'anticiper ce risque de déviance. Il faut en permanence chercher l'équilibre entre innovation scientifique et protection individuelle des patients, entre utilité publique et intérêt personnel (89)(78).

3. Limites de l'étude

Notre étude présente plusieurs limitations. Tout d'abord, il s'agit d'une étude monocentrique, l'ensemble de la cohorte a été recruté au sein d'un centre de chirurgie cardiaque parisien. De plus, la saisie des données a été faite de façon prospective mais le calcul de l'EuroSCORE II et la comparaison des différents modèles ont été faits de façon rétrospective ; des biais, notamment de classement, peuvent donc être présents.

Par ailleurs, nous avons exclu les patients admis pour une chirurgie cardiaque en urgence. De fait, notre modèle était difficilement comparable à l'EuroSCORE II. Ce critère d'exclusion a été défini dans la mesure où n'étions pas en mesure de coter le caractère urgent de l'intervention tel qu'il est décrit (« urgence, emergency, salvage ») dans l'EuroSCORE II. Cependant, nous estimions que le processus décisionnel en situation d'urgence relevait d'enjeux différents ne pouvant être tous pris en compte par des scores de mortalité ou des supports informatiques d'aide à la décision. Du fait de la conception de notre base de données, contrairement à l'EuroSCORE II, notre critère de jugement principal était la mortalité hospitalière et non la mortalité post opératoire à 30 jours. Nous avons utilisé la mortalité hospitalière pour l'ensemble de nos modèles.

Des études ont montré que les performances prédictives du *Machine Learning* pouvaient être altérées par la rareté de la survenue de l'évènement d'intérêt. Cette altération attendue est logique puisque les algorithmes de *Machine Learning apprennent* par répétition, à partir de l'observation de milliers d'exemples. Or le taux de mortalité dans notre cohorte

n'atteint « que » 6,3 % (35)(90). Ceci explique peut-être pourquoi les différents algorithmes utilisés de façon isolée avaient des performances sensiblement similaires à celles de l'EuroSCORE II. C'est l'assemblage des algorithmes de *Machine Learning* (modèle *Machine Learning*) qui rend notre modèle plus efficace.

Nous n'avons pas comparé notre modèle de *Machine Learning* au *STS Score* alors que ce score de mortalité est largement utilisé outre atlantique. Cette limitation tient au fait que nous ne possédions pas toutes les variables nécessaires ; le calcul de ce score directement à partir des données saisies était donc impossible.

De plus, nous avons validé notre modèle au sein d'un échantillon test de notre cohorte. Mais comme la majorité des études de développement de modèle prédictif, nous n'avons pas réalisé de validation de notre score au sein d'une cohorte externe à l'étude.

Nous l'avons vu, la possibilité de généraliser notre modèle est limitée. Mais une approche locale, telle que celle que nous avons développée ou retrouvée au sein de différentes études, permet de s'assurer de l'homogénéité de la prise en charge et également de se rapprocher au plus près des problématiques d'une population donnée, à un temps donné (54). Par ailleurs, il serait intéressant d'évaluer des algorithmes de *Machine Learning* dans des situations complexes où les scores habituels sont moins performants : chez le sujet âgé de plus de 80 ans notamment ou chez le sujet candidat à une valvuloplastie aortique (voie percutanée ou chirurgicale ?) ou mitrale. Ces scores seraient des aides complémentaires au processus décisionnel dans ces situations où l'incertitude est plus grande encore.

Par ailleurs, nous avons choisi d'évaluer la pertinence clinique de nos modèles grâce à une DCA mais il existe d'autres approches telles que *le Net Reclassification Improvement* (NRI) (91).

4. Conclusion

L'objectif principal de notre travail était d'améliorer la sélection des patients en amont d'une chirurgie cardiaque programmée en optimisant la prédiction de la mortalité hospitalière post opératoire. A cet effet, nous avons créé un nouveau score de mortalité grâce à une technique d'IA : le *Machine Learning* que nous avons comparé au score de référence recommandé, l'EuroSCORE II.

L'objectif secondaire était de fournir un mode d'évaluation, de ces deux modèles prédictifs, complémentaire à l'évaluation par courbe ROC et plus proche des préoccupations des cliniciens : la DCA. Les courbes ROC montraient une meilleure discrimination du *Machine Learning*. La DCA montrait un bénéfice net de ce modèle toujours supérieur à l'EuroSCORE II. Plus précisément, jusqu'au seuil de 20 % de probabilité pt, le *Machine Learning* permettait d'identifier tous les 1000 patients 5 à 10 sujets, de plus que l'EuroSCORE II, qui décèderaient s'ils étaient opérés sans diminuer le nombre d'interventions bénéfiques.

Notre étude montre une supériorité du *Machine Learning* par rapport à l'EuroSCORE II pour la prédiction de la mortalité dans les suites d'une chirurgie cardiaque programmée.

L'utilisation de scores de mortalité permet une stratification du risque opératoire. Il ne s'agit pas d'homogénéiser ou de figer les éléments décisionnels mais bien d'apporter un support complémentaire et objectif qui contribue à l'amélioration de l'identification des patients les plus à même de bénéficier d'une chirurgie cardiaque et surtout des patients les plus à risque, potentiellement à récuser. Les variables explicatives sont intriquées, et liées au risque de décès post-opératoire de façon non linéaire, potentiellement interdépendante, évolutive et dynamique. Ces relations complexes ne peuvent être évaluées de façon précise par les techniques statistiques traditionnelles et ne peuvent être modélisées en un cours intervalle de temps par l'esprit humain. `

Les algorithmes de *Machine Learning* répondent à ces exigences. Ils sont capables d'analyser des données hétérogènes, en quantité massive, rapidement. L'exploitation de l'information médicale numérique oriente la médecine vers une pratique plus précise, personnalisée, prédictive et participative. L'objectif étant d'adapter le plus précisément possible notre prise en charge, à l'échelle individuelle, afin de prévenir ou retarder un évènement « indésirable » qui peut tout aussi bien être le décès dans les suites d'une chirurgie cardiaque ou la survenue d'une maladie chronique.

VI – REFERENCES BIBLIOGRAPHIQUES

1. Azarnoush L, Bernard A, Bori Bata AK, de Brux J, Camilleri L, François F, et al. Rapport Epicard 2015 : les bases de données de la Société Française de Chirurgie Thoracique et Cardio-Vasculaire. 2015.

2. Birzouan P. « Variabilité décisionnelle en chirurgie cardiaque et exclusion des patients âgés ou trop gravesves », *Les Cahiers du Centre Georges Canguilhem*, 1/2014 (N° 6), p. 259-279.

3. Windecker S, Kolh P, Alfonso F, Collet J-P, Cremer J, Falk V, et al. Guidelines on myocardial revascularization : The task Force on Myocardial Revascularization of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J. 2014;(35):2641‑2619.

4. Nashef S.A.M, Roques F, Sharples L, Nilsson J, Smith C, Goldstone A, et al. EuroSCORE II. Eur J Cardio Thorac Surg. 2012;41:734‑45.

5. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. Society of Thoracic Surgeons Quality Measurement Task Force. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009;(88).

6. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. Society of Thoracic Surgeons Quality Measurement Task Force. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2--isolated valve surgery. Ann Thorac Surg. 2009;(88).

7. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. Society of Thoracic Surgeons Quality Measurement Task Force. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. Ann Thorac Surg. 2009;(88).

8. Pichegru S. Evolution du profil de risque des patients en chirurgie cardiaque : performance des scores de gravité. Grenoble; 2012.

9. Litzler P., Smail H. Comment évaluer le risque chirurgical à partir des scores ? Coeur Anesth. 2013;461‑461.

10. Darcy AM, Louie AK, Weiss Robert L. Machine learning and the profession of medicine. JAMA. 2016;315:551‑2.

11. Cook NR. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. Circulation. 2007;115(7):928‑35.

12. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating

prediction models. Med Decis Mak Int J Soc Med Decis Mak. 2006;26(6):565‑74.

13. Pierri MD, Capestro F, Zingaro C, Torraca L. The changing face of cardiac surgery patients: an insight into a Mediterranean region. Eur J Cardio Thorac Surg. 2010;38:407‑13.

14. Barili F, Pacini D, Capo A, Rasovic O, Grossi C, Alamanni F, et al. Does EuroSCORE II perform better than its original versions? A multicentre validation study. Eur Heart J. 2013;34(1):22‑9.

15. Krane M, Voss B, Hiebinger A, Deutsch MA, Wottke M, Hapfelmeier A, et al. Twenty Years of Cardiac Surgery in Patients Aged 80 Years and Older: Risks and Benefits. Ann Thorac Surg.2011;91(2):506‑13.

16. Walton NA, Martin DK, Peter EH, Pringle DM, Singer PA. Priority setting and cardiac surgery: A qualitative case study. Health Policy.2007;80(3):444‑58.

17. Nashef S, Roques F, Michel P, Gauducheau E, Lemeshow S, Salomon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardio Thorac Surg. 1999;16:9‑13.

18. Roques F, Nashef S, Michel P, Gauducheau E, Vincentiis de, Baudet E, et al. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardio Thorac Surg. 1999;15:816‑23.

19. Michel P, Roques F, Nashef S. Logistic or additive EuroSCORE for high- risk patients? Eur J Cardiothorac Surg. 2003;23:684‑7.

20. Roques F, Michel P, Nashef SAM. The logistic EuroSCORE. Eur Heart J. 2003;24.

21. Siregar S, Groenwold R, de Heer F, Bots M, Van der Graaf Y, Van Herwerden L. Performance of the original EuroSCORE. Eur J Cardio Thorac Surg. 2012;41:746‑54.

22. Dedda UD, Pelissero G, Agnelli B, Vincentiis CD, Castelvecchio S, Ranucci M. Accuracy, calibration and clinical performance of the new EuroSCORE II risk stratification system. Eur J Cardiothorac Surg. 2013;43(1):27‑32.

23. Poullis M, Pullan M, Chalmers J, Mediratta N. The validity of the original EuroSCORE and EuroSCORE II in patients over the age of seventy. Interact Cardiovasc Thorac Surg. 2015;20(2):172‑7.

24. Chevalier A. Performance prédictive et limitation de l'Euroscore I et II chez les patients octogénaires ayant bénéficié d'une chirurgie cardiaque avec circulation extra corporelle. Paris; 2015.

25. Biancari F, Vasques F, Mikkola R, Martin M, Lahtinen J, Heikkinen J. Validation of EuroSCORE II in patients undergoing coronary artery bypass surgery. Ann Thorac Surg. 2012;93:1930‑5.

26. Chalmers J, Pullan M, Fabri B, McShane J, Shaw M, Mediratta N, et al. Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. Eur J Cardiothorac Surg. 2013;43(4):688‑94.

27. Haensig M, Holzhey DM, Borger MA, Schuler G, Shi W, Subramanian S, et al. Is the new EuroSCORE II a better predictor for transapical aortic valve implantation? Eur J Cardiothorac Surg. 2013;44(2):302‑8.

28. Moscarelli M, Bianchi G, Margaryan R, Cerillo A, Farneti P, Murzi M, et al. Accuracy of EuroSCORE II in patients undergoing minimally invasive mitral valve surgery. Interact Cardiovasc Thorac Surg. 2015;21(6):748‑53.

29. Biancari F, Juvonen T, Onorati F, Faggian G, Heikkinen J, Airaksinen J, et al. Meta-analysis on the Performance of the EuroSCORE II and the Society of Thoracic Surgeons Scores in Patients Undergoing Aortic Valve Replacement. J Cardiothorac Vasc Anesth. 2014;28(6):1533‑9.

30. El Sanharawi M, Naudet F. Comprendre la régression logistique - EM|consulte. J Fr Ophtalmol. 2013;36:710‑5.

31. Preux P., Odermatt P, Perna A, Marin B, Vergnenègre A. Qu'est ce qu'une regression logistique ? Rev Mal Respir. 2005;(22):159‑62.

32. Russel S, Norvig P. Artificial Intelligence : a Modern Approach. 3rd éd. 2010.

33. Turing AM. Computing Machinery and Intelligence. Mind. 1950;49:433‑60.

34. Cornuéjols A, Miclet L. Apprentissage artificiel : concepts et algorithmes. 1992.

35. Churpek M, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventionnal regression for predicting clinical deterioration on the wards. Crit Care Med. 2016;44:368‑74.

36. Castelvecchi D. Can we open the black box of AI? Nat News. 2016;538(7623):20.

37. Biernat E, LUTZ M. Data science : fondamentaux et études de cas. 2015. (EYROLLES).

38. Colloque CIGREF « Intelligence Artificielle, enjeux pour les entreprises ». In 2016
39. Beam AL, Kohane IS. Translating Artificial Intelligence Into Clinical Care. JAMA. 2016;316(22):2368‑9.

40. Wiens J, Wallace BC. Editorial: special issue on machine learning for health and medicine. Mach Learn. 2016;102(3):305‑7.

41. Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. Proc IEEE. 2016;104(1):176‑97.

42. Obermeyer Z, Emanuel EJ. Predicting the future - Big data, machine learning and clinical medicine. NEJM. 2016;375:1216‑9.

43. Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun. 2016;7:12474.

44. Sertel O, Kong J, Catalyurek UV, Lozanski G, Saltz JH, Gurcan MN. Histopathological Image Analysis Using Model-Based Intermediate Representations and Color Texture: Follicular Lymphoma Grading. J Signal Process Syst. 2009;55(1‑3):169.

45. Sabo E, Beck AH, Montgomery EA, Bhattacharya B, Meitner P, Wang JY, et al. Computerized morphometry as an aid in determining the grade of dysplasia and progression to adenocarcinoma in Barrett's esophagus. Lab Invest. t 2006;86(12):1261‑71.

46. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James J, et al. Single Reading with Computer-Aided Detection for Screening Mammography. N Engl J Med. 2008;359(16):1675‑84.

47. Amad A, Cancel A, Fovet T. L'imagerie cérébrale en psychiatrie clinique : Du diagnostic différentiel au machine learning. ResearchGate. 2016;92(92):277‑84.

48. Veronese E, Castellani U, Peruzzo D, Bellani M, Brambilla P. Machine Learning Approaches: From Theory to Application in Schizophrenia. Comput Math Methods Med. 2013;2013:e867924.

49. Boughorbel S, Al-Ali R, Elkum N. Model Comparison for Breast Cancer Prognosis Based on Clinical Data. PLOS ONE. 2016;11(1):e0146413.

50. Asadi H, Dowling R, Yan B, Mitchell P. Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy. PLOS ONE. 2014;9(2):e88225.

51. Sertel O, Kong J, Shimada H, Catalyurek UV, Saltz JH, Gurcan MN. Computer-aided Prognosis of Neuroblastoma on Whole-slide Images: Classification of Stromal Development. Pattern Recognit. 2009;42(6):1093‑103.

52. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Cai T, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol Psychiatry. 2016;21(10):1366‑71.

53. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, Rashidi P, Pardalos P, Momcilovic P, et al. Application of Machine Learning Techniques to High-Dimensional Clinical Data to Forecast Postoperative Complications. PLOS ONE. 2016;11(5):e0155705.

54. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach. Acad Emerg Med. 2016;23(3):269‑78.

55. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, Laan MJ van der. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3(1):42‑52.

56. Pencina M, D'Agostino R Sr, D'Agostino R Jr, Vasan R. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157‑72.

57. Fitzgerald M, Saville BR, Lewis RJ. Decision Curve Analysis. JAMA. 2015;313(4):409‑10.

58. Zastrow S, Brookman-May S, Cong TAP, Jurk S, von Bar I, Novotny V, et al. Decision curve analysis and external validation of the postoperative Karakiewicz nomogram for renal cell carcinoma based on a large single-center study cohort. World J Urol. 2015;33(3):381‑8.

59. Shariat SF, Savage C, Chromecki TF, Sun M, Scherr DS, Lee RK, et al. Assessing the clinical benefit of NMP22 in the surveillance of patients with non–muscle-invasive bladder cancer and negative cytology: a decision-curve analysis. Cancer. 2011;117(13):2892‑7.

60. Hernandez JM, Tsalatsanis A, Humphries LA, Miladinovic B, Djulbegovic B, Velanovich V. Defining optimum treatment of patients with pancreatic adenocarcinoma using regret based decision curve analysis. Ann Surg. 2014;259:1208‑14.

61. Hastie, Tibshirani, Friedman. The Elements of Statistical Learning - Data Mining, Inference, 2009.

62. Breiman L. Random Forests. Mach Learn. 2001;45(1):5‑32.

63. Fellahi J.L., Ouattara A, Le Manach Y. Anesthésie-réanimation en chirurgie cardiaque: 2014. (Score de risque et stratification du risque).

64. Houthooft R, Ruyssinck J, Herten J van der, Stijven S, Couckuyt I, Gadeyne B, et al. Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. Artif Intell Med. 2015;63(3):191‑207.

65. Liu NT, Holcomb JB, Wade CE, Batchinsky AI, Cancio LC, Darrah MI, et al. Development and validation of a machine learning algorithm and hybrid system to predict the need for life-saving interventions in trauma patients. Med Biol Eng Comput. 2014;52(2):193‑203.

66. Liu NT, Holcomb JB, Wade CE, Dannah MI, Salinas J. Utility of vital signs, heart rate variability and complexity, and machine learning for indentifying the need for lifesaving interventions in trauma patients. Shock. 2014;42:108‑14.

67. Batchinsky AI, Salinas J, Jones JA, Necsoiu C, Cancio LC. Predicting the Need to Perform Life-Saving Interventions in Trauma Patients by Using New Vital Signs and Artificial Neural Networks. In: Artificial Intelligence in Medicine Springer, Berlin,

Heidelberg; 2009. p. 390‑4.

68. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436‑44.

69. Wong TY, Bressler NM. Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening. JAMA. 2016;316(22):2366‑7.

70. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402‑10.

71. Ortiz A, Munilla J, Górriz JM, Ramírez J. Ensembles of Deep Learning Architectures for the Early Diagnosis of the Alzheimer's Disease. Int J Neural Syst. 2016;26(07):1650025.

72. Xu J, Xiang L, Liu Q, Gilmore H, Wu J, Tang J, et al. Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images. IEEE Trans Med Imaging. 2016;35(1):119‑30.

73. Allyn J, Ferdynus C, Bohrer M, Dalban C, Valance D, Allou N. Simplified Acute Physiology Score II as Predictor of Mortality in Intensive Care Units: A Decision Curve Analysis. PLOS ONE. 2016;11(10):e0164828.

74. Yamamoto S, Yamazaki S, Shimizu T, Takeshima T, Fukuma S, Yamamoto Y, et al. Prognostic utility of serum CRP levels in combination with CURB-65 in patients with clinically suspected sepsis: a decision curve analysis. BMJ Open. 2015;5(4):e007049. `

75. Hilden J. Evaluation of diagnostic tests - the schism. Soc Med Decis Mak Newsl. 2004;(4):5‑6.

76. Hozo I, Tsalatsanis A, Djulbegovic B. Monte Carlo decision curve analysis using aggregate data. Eur J Clin Invest. 2017;47(2):176‑83.

77. Collins FS, Varmus H. A New Initiative on Precision Medicine. N Engl J Med. 26 2015;372(9):793‑5.

78. Béranger J. Les systèmes d'information en santé et l'éthique [Internet]. ISTE éditions. 2015

79. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc. 1996;58(1):267‑88.

80. Jovanovic M, Radovanovic S, Vukicevic M, Poucke SV, Delibasic B. Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression. Artif Intell Med. 2016;72:12‑21.

81. Jolliffe IT. Principal Component Analysis. Deuxième. Springer; 2002.

82. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach

Learn Res. 2003;3:1157–1182.

83. Cardon D. A quoi rêvent les algorithmes, Dominique Cardon, Sciences humaines -. Seuil. 2015

84. Glenn D. R, Perry L. M. Artificial Intelligence in Anesthesia and Intensive Care. J Clin Monit. 1988;4:274‑89.

85. Garnelo M, Arulkumaran K, Shanahan M. Towards Deep Symbolic Reinforcement Learning. ArXiv160905518 Cs [Internet]. 18 sept 2016

86. Lipworth W, Mason PH, Kerridge I, Ioannidis JPA. Ethics and Epistemology in Big Data Research. J Bioethical Inq. 2017;1‑12.

87. Floridi L, Taddeo M. What is data ethics? Phil Trans R Soc A. 2016; 374(2083): 20160360.

88. Tattersall A, Grant MJ. Big Data – What is it and why it matters. Health Inf Libr J. 2016;33(2):89‑91.

89. Goodman KW. Ethics, Information Technology, and Public Health: New Challenges for the Clinician-Patient Relationship. J Law Med Ethics. 2010;38(1):58‑63.

90. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009;21(9):1263‑84.

91. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MSV, et al. Criteria for Evaluation of Novel Markers of Cardiovascular Risk. Circulation. 2009;119(17):2408‑16.

VII – ANNEXES

1. Données recueillies relatives aux patients

Epidémiologie	Âge, sexe, Indice de Masse Corporelle
Antécédents	FRCV : Tabagisme, HTA, Diabète +/- insulinoréquérant, Dyslipidémie Coronaropathie Artériopathie périphérique Insuffisance cardiaque Infarctus < 90 jours Endocardite infectieuse Maladie thrombo embolique AVC hémorragique ou ischémique ACFA, arythmie ventriculaire, pacemaker BPCO Insuffisance rénale chronique +/- dialysée Ulcère gastro duodénal Néoplasie, radiothérapie médiastinale Déficit immunitaire Cirrhose hépatique Faible mobilité
Etat clinique	Rythme sinusal ou non Angor et classe Stade NYHA dyspnée Insuffisance cardiaque Endocardite infectieuse ⇒ Etat pré opératoire critique ?
Traitements	AAP Bétabloquants Antiarythmique IEC / ARA 2 Statines Diurétiques Inhibiteurs calciquesDérivés nitrés
Examens complémentaires	Hémoglobine TP, plaquettes Créatininémie pré opératoire et clairance Cockroft MDRD Données échographiques : - FEVG et PAPS - valvulopathie - insuffisance tricuspide fonctionnelle - cardiopathie congénitale Pathologie aorte ascendante Coronarographie et nombre de sténoses

MDRD Modification of Diet in Renal Disease, **PAPS** Pression Artérielle Pulmonaire Systolique, **TP** Temps de Prothrombine

2. Caractéristiques de la population

	Données manquantes	Total (n = 6 520)	Vivants (n = 6109)	Décédés (n=411)	Valeur p
Age, années	0	63,4 (14,4)	63,1 (14,5)	68,2 (13,4)	< 0,0001
Sexe (masculin)	0	4449 (68,2%)	4178 (68,4%)	271(65,9%)	0,30
Taille, cm	19	168,5 (9,5)	168,6 (9,5)	166,7 (9,8)	< 0,0001
Poids, kg	14	75,4 (14,9)	75,6 (14,9)	72,4 (15,0)	< 0,0001
BMI, kg/m2	19	26,5 (4,6)	26,5 (4,6)	26,0 (4,9)	0,02
Comorbidités n (%)					
Ulcère gastro duodénal	0	288 (4,4%)	265 (4,3%)	23 (5,6%)	0,23
Cancer - Non - Oui, > 5 ans - Oui, < 5 ans - Non contrôlé sans métastase - Non contrôlé avec métastase	0	 6002 (92,1%) 238 (3,6%) 195 (3%) 66 (1%) 19 (0,3%)	5654 (92,5%) 210 (3,4%) 169 (2,8%) 60 (1%) 16 (0,3%)	348 (84,7%) 28 (6,8%) 26 (6,3%) 6 (1,5%) 3 (0,7%)	 <0,0001
Cirrhose - Non - Non compliquée - Compliquée	0	 6455 (99%) 40 (0,6%) 25 (0,4%)	6056 (99,1%) 32 (0,5 %) 21 (0,3%)	399 (97,1%) 8 (1,9%) 4 (1%)	0,0002
Déficit immunitaire	0	88 (1,3%)	73 (1,2%)	15 (3,7%)	< 0,0001
Pathologie pulmonaire chronique	0	375 (5,8%)	330 (5,4%)	45 (10,9%)	< 0,0001
BPCO	0	642 (9,9%)	561 (9,2%)	81 (19,7%)	< 0,0001
Diabète de type 2	2	1674 (25,7%)	1547 (25,3%)	127 (31%)	0,01
Diabète de type 2 insulinoréquérant	2	513 (7,9%)	459 (7,5%)	54 (13,2%)	< 0,0001
HTA	0	3678 (56,4%)	3427 (56,1%)	251 (61,1%)	0,05
Tabagisme actif ou ancien	25	3336 (52,4%)	3131 (51,4%)	205 (50,5%)	0,72

	Données manquantes	Total (n = 6 520)	Vivants (n = 6109)	Décédés (n=411)	Valeur p
Artériopathie périphérique	0	905 (13,9%)	829 (13,6%)	76 (18,5%)	0,005
IRC sous dialyse	0	57 (0,9%)	47 (0,8%)	10 (2,4%)	0,003
Dyslipidémie	0	3288 (50,4%)	3100 (50,7%)	188 (45,7%)	0,05
Coronaropathie	0	2496 (38,3%)	2351 (38,5%)	145 (35,3%)	0,19
Infarctus du myocarde < 90 jours	0	440 (6,7%)	411 (6,7%)	29 (7,1%)	0,80
ATCD Chirurgie cardiaque	0	627 (9,6%)	532 (8,7%)	95 (23,1%)	<0,001
ATCD Endocardite	0	192 (2,9%)	171 (2,8%)	21 (5,1%)	0,007
ATCD Insuffisance cardiaque congestive	0	1047 (16,1%)	919 (1,0%)	128 (31,1%)	< 0,0001
Classe Angor 0 1 2 3 4	0	 4919 (75,4%) 165 (2,5%) 949 (14,6%) 436 (6,7%) 51 (0,8%)	 4588 (75,1%) 159 (2,6%) 912 (14,9%) 408 (6,7%) 42 (0,7%)	 331 (80,5%) 6 (1,5%) 37 (9%) 28 (6,8%) 9 (2,2%)	 < 0,0001
Maladie thrombo embolique	0	353 (5,4%)	317 (5,2%)	36 (8,8%)	0,002
AVC ischémique	0	464 (7,1%)	414 (6,8%)	49 (11,9 %)	< 0,0001
AVC hémorragique	0	57 (0,9%)	52 (0,8%)	5 (1,2%)	0,44
Faible mobilité	0	179 (2,7%)	161 (2,6%)	18 (4,4%)	0,04
Radiothérapie médiastinale	0	122 (1,9%)	111 (1,8%)	11 (2,7%)	0,21
Arythmie SV Absence paroxystique permanente	0	 5148 (79%) 646 (9,9%) 726 (11,1%)	 4886 (80%) 570 (9,3%) 653 (10,7%)	 262 (63,7%) 76 (18,5%) 73 (17,8%)	 < 0,0001
Arythmie ventriculaire	0	103 (1,6%)	95 (1,6%)	8 (1,9%)	0,54
Pacemaker	0	222 (3,4%)	189 (3,1%)	33 (8%)	< 0,0001

	Données manquantes	Total (n = 6 520)	Vivants (n = 6109)	Décédés (n=411)	Valeur p
Etat clinique pré opératoire					
ACFA	0	776 (11,9%)	696 (11,4%)	80 (19,5%)	< 0,0001
NYHA classe 1 2 3 4	0	 1374 (21,1%) 603 (9,2%) 2140(40,1%) 1929 (29,6%)	 1307 (21,4%) 587 (9,6%) 2496 (40,9%) 1719 (28,1%)	 67 (16,3%) 16 (3,9%) 118 (28,7%) 210 (51,1%)	 < 0,0001
Insuffisance cardiaque congestive	0	255 (3,9%)	200 (3,3%)	55 (13,4%)	< 0,0001
Endocardite active	0	202 (3,1%)	172 (2,8%)	30 (7,3%)	< 0,0001
Etat critique pré opératoire	0	107 (1,6%)	71 (1,2%)	36 (8,8%)	< 0,0001
Traitement pré opératoire					
Anti agrégants plaquettaires - Aucun - Que Aspirine - Autres	0	 2716 (41,7%) 2666 (40,9%) 1138 (17,4%)	2526 (41,3%) 2519 (41,2%) 1064 (17,4%)	190 (46,2%) 147 (35,8%) 74 (18 %)	0,08
Béta Bloquants	0	3887 (59,6%)	3661 (59,9%)	226 (55%)	0,05
Anti arythmiques	0	880 (13,5%)	796 (13%)	84 (20,4%)	<0,0001
Statines	0	3837 (58,8%)	3636 (59,5%)	201 (48,9%)	<0,0001
Diurétiques	0	2359 (36,2%)	2125 (34,8%)	234 (56,9%)	<0,0001
Inhibiteurs Calciques	0	1298 (19,9%)	1217 (19,9%)	81 (19,7%)	0,92
IEC ou ARA II	0	3313 (50,8%)	3117 (51%)	196 (47,7%)	0,19
Dérivés nitrés		85 (1,3%)	79 (1,3%)	6 (7,1%)	0,77
Données pré opératoires paracliniques					
Taux Hb g/dL	20	13,3 (1,8)	13,4 (1,7)	12,4 (2,1)	< 0,0001
Plaquettes G/L	59	232,9 (78,3)	233,3 (77,3)	227,3 (91,1)	0,01
TP %	91	83,3 (12,6)	89,8 (12,1)	84,5 (17,1)	< 0,0001
Créatininémie, ug/L	27	99,5 (62,6)	97,8 (59,8)	124,8 (91,8)	< 0,0001
Clairance créatinine Cockroft mL/min	37	77,7 (32,7)	78,9 (32,5)	60,3 (29,5)	< 0,0001
Cl créatinine MDRD mL/min	38	75,7 (25,3)	76,5 (25)	63,9 (27,7)	< 0,0001

	Données manquantes	Total (n = 6 520)	Vivants (n = 6109)	Décédés (n=411)	Valeur p
PAPS mmHg	0	31,9 (12,6)	31,6 (12.3)	36,7 (15)	< 0,0001
FEVG %	2	57,9 (12,2)	58,1 (12)	54,7 (13.7)	< 0,0001
Coronarographie pré opératoire, n (%)					
Coronaropathie	0	3253 (4,.9%)	3069 (50,2%)	184 (44,8%)	0,03
Nombre de sténoses 0 1 2 3 4	0	 3211 (49,2%) 487 (7,5%) 624 (9,6%) 2079 (31,9%) 119 (1,8%)	 3002 (49,1%) 445 (7,3%) 588 (9,6%) 1959 (32,1%) 115 (1,9%)	 209 (50,8%) 42 (10,2%) 36 (8,8 %) 120 (29,2%) 4 (1%)	0,10
Valvulopathie	0	3810 (58,4%)	3526 (57,7%)	284 (69,1%)	<0,0001
IT fonctionnelle	0	693 (10,6%)	630 (10,3%)	63 (15,3%)	0,001
Aorte ascendante : - Pas de pathologie - Anévrysme - Dissection	0	 6026 (92,4%) 435 (6,7%) 59 (0,9%)	 5652 (92,5%) 410 (6,7%) 47 (0,8%)	 374 (91%) 25 (6,1%) 12 (2,9%)	<0,0001
Cardiopathie congénitale	0	88 (1,3%)	87 (1,4%)	1 (0,2%)	0,05
Caractéristiques de la chirurgie, n (%)					
Nombre procédure(s) chirurgicale(s) - isolée PAC - Chirurgie isolée non PAC - 2 procédures - 3 procédures	0	 2504 (38,4%) 2176 (33,4%) 1511 (23,2%) 329 (5%)	 2399 (39,3%) 2038 (33,4%) 1388 (22,7%) 284 (4,6%)	 105 (25,5 %) 138 (33,6%) 123 (29,9%) 45 (10,9%)	< 0,0001
Chirurgie coronaire	0	3127 (52%)	2953 (48,3%)	174 (42,3%)	0,02
Chirurgie valvulaire	0	3649 (56%)	3376 (55,3%)	273 (66,4%)	<0,0001
Chirurgie valve aortique	0	2382 (36,5%)	2214 (36,4%)	168 (40,9%)	0,06
Chirurgie valve mitrale	0	1517 (23,3%)	1387 (22,7%)	130 (31,6%)	<0,0001
Chirurgie valve tricuspide	0	765 (11,7%)	691 (11,3%)	74 (18%)	< 0,0001
Chirurgie aorte thoracique	0	569 (0,7%)	523 (8,6%)	46 (8,1%)	0,07

Les données quantitatives sont présentées en moyenne et écart type, les données qualitatives en n et %.

ATCD Antécédent**,**, **Hb** Hémoglobine, **IRC** Insuffisant Rénal Chronique, **IT** Insuffisance Tricuspide, **TP** Temps Prothrombine

Mortality prediction after elective cardiac surgery: new Machine Learning approach and Decision Curve Analysis assessment

ABSTRACT

Introduction

In cardiac surgery, morbi-mortality is important and the decision to operate is complex. Study purpose was to compare EuroSCORE II and Machine Learning to predict mortality after elective cardiac surgery *via* a Decision Curve Analysis (DCA).

Methods

We conducted a retrospective monocentric study from December 2015 to December 2016, using a prospective data base, from the cardiac surgery unit of a University Hospital in Paris. Non elective cardiac surgery patients were excluded. The different models of prediction of in hospital mortality, including EuroSCORE II, Logistic Regression and Machine Learning, were compared by Receiver Operating Characteristic (ROC) and a Decision Curve Analysis (DCA).

Results

The study was carried among 6520 patients. Mortality rate was 6.3%. Average age was 63.4 years old and the average EuroSCORE II was 3.7%. Area under the ROC curve (IC 95%) for the Machine Learning 0.795 (0.755-0.834) model was significantly higher than the one of the EuroSCORE II and the Logistic Regression models (respectively 0.737 (0.691-0.783) and 0.742 (0.698-0.785, $p<0.0001$). The DCA proved that the Machine Learning model had a greater clinical benefit.

Conclusion

According to ROC curve and DCA, Machine Learning model was more efficient than EuroSCORE II to predict in hospital mortality after elective cardiac surgery. This result confirms the growing interest of Machine Learning in establishing predictive models in medicine.

Key words: Cardiac surgery, Decision Curve Analysis, EuroSCORE II, Machine Learning, Post operative mortality, Predictive model

RÉSUMÉ

Introduction

En chirurgie cardiaque, la morbi-mortalité est importante et la décision d'opérer un patient est complexe. L'objectif de notre étude était de comparer l'EuroSCORE II au *Machine Learning* pour prédire la mortalité dans les suites d'une chirurgie cardiaque programmée *via* une *Decision Curve Analysis* (DCA).

Matériel et Méthodes

Nous avons réalisé une étude rétrospective, monocentrique, entre décembre 2005 et décembre 2012, à partir d'une base de données prospective, au sein de l'unité de chirurgie cardiaque d'un Centre Hospitalo-Universitaire parisien. Les patients admis pour chirurgie cardiaque non programmée étaient exclus. Les différents modèles de prédiction de la mortalité hospitalière, incluant l'EuroSCORE II, le modèle de régression logistique et le modèle *Machine Learning*, étaient comparés via une courbe *Receiver Operating Characteristic* (ROC) et une DCA.

Résultats

Il y avait 6520 patients. La mortalité hospitalière était de 6,3%. La moyenne d'âge était de 63,4 ans et l'EuroSCORE II moyen de 3,7 %. L'aire sous la courbe ROC (IC 95%) pour le modèle *Machine Learning* 0,795 (0,755-0,834) était significativement plus élevée que celle de l'EuroSCORE II et du modèle de Régression Logistique (respectivement 0,737 (0,691-0,783) et 0,742 (0,698-0,785, $p<0,0001$). Après DCA, le modèle de *Machine Learning* avait un plus grand bénéfice clinique.

Conclusion

D'après la courbe ROC et la DCA, le modèle *Machine Learning* était supérieur à l'EuroSCORE II pour la prédiction du risque de mortalité intra hospitalière dans les suites d'une chirurgie cardiaque programmée. Ce résultat confirme l'intérêt grandissant du *Machine Learning* pour l'établissement de modèles prédictifs en médecine.

Anesthésie Réanimation

Mots clefs : Chirurgie cardiaque, Decision Curve Analysis, Euroscore II, Machine Learning, mortalité post opératoire, score prédictif