Université de Montréal

Développement et validation d'un modèle d'apprentissage machine pour la détection de potentiels donneurs d'organes

Par

Nicolas Sauthier, M.D., B.Ing.

Faculté de Médecine

Mémoire présenté en vue de l'obtention du grade de maître ès sciences (M.Sc.)

En sciences biomédicales option médecine computationnelle

Août 2022

Université de Montréal

Faculté de Médecine

Ce mémoire intitulé

Développement et validation d'un modèle d'apprentissage machine pour la détection de potentiels donneurs d'organes

Présenté par

Nicolas Sauthier, M.D., B.Ing.

A été évalué par un jury composé des personnes suivantes

Aude Motulsky, Ph.D.

Président-rapporteur

Michaël Chassé, M.D., Ph.D.

Directeur de recherche

Guillaume Dumas, Ph.D.

Membre du jury

Résumé

Le processus du don d'organes, crucial pour la survie de nombreux patients, ne répond pas à la demande croissante. Il dépend d'une identification, par les cliniciens, des potentiels donneurs d'organes. Cette étape est imparfaite et manque entre 30% et 60% des potentiels donneurs d'organes et ce indépendamment des pays étudiés. Améliorer ce processus est un impératif à la fois moral et économique. L'objectif de ce mémoire était de développer et valider un modèle afin de détecter automatiquement les potentiels donneurs d'organes.

Pour ce faire, les données cliniques de l'ensemble des patients adultes hospitalisés aux soins intensifs du CHUM entre 2012 et 2019 ont été utilisées. 103 valeurs de laboratoires temporelles différentes et 2 valeurs statiques ont été utilisées pour développer un modèle de réseaux de neurones convolutifs entrainé à prédire les potentiels donneurs d'organes. Ce modèle a été comparé à un modèle fréquentiste linéaire non temporel. Le modèle a par la suite été validé dans une population externe cliniquement distincte. Différentes stratégies ont été comparées pour peaufiner le modèle dans cette population externe et améliorer les performances.

Un total de 19 463 patients, dont 397 donneurs potentiels, ont été utilisés pour développer le modèle et 4 669, dont 36 donneurs potentiels, ont été utilisés pour la validation externe. Le modèle démontrait une aire sous la courbe ROC (AUROC) de 0.966 (IC95% 0.949-0.981), supérieure au modèle fréquentiste linéaire (AUROC de 0.940 IC95% 0.908-0.969, p=0.014). Le modèle était aussi supérieur dans certaines sous populations d'intérêt clinique. Dans le groupe de validation externe, l'AUROC du modèle de réseaux de neurones était de 0.820 (0.682-0.948) augmentant à 0.874 (0.731-0.974) à l'aide d'un ré-entrainement.

Ce modèle prometteur a le potentiel de modifier et d'améliorer la détection des potentiels donneurs d'organes. D'autres étapes de validation prospectives et d'amélioration du modèle, notamment l'ajout de données spécifiques, sont nécessaires avant une utilisation clinique de routine.

Mots-clés : don d'organes, transplantation, modèle prédictif, apprentissage machine, autoencodeur, réseaux de neurones, transfert de connaissance.

Abstract

The organ donation process, however crucial for many patients' survival, is not enough to address the increasing demand. Its efficiency depends on potential organ donors' identification by clinicians. This imperfect step misses between 30%–60% of potential organ donor. Improving that process is a moral and economic imperative. The main goal of this work was to address that liming step by developing and validating a predictive model that could automatically detect potential organ donors.

The clinical data from all patients hospitalized, between 2012 and 2019 to the CHUM critical care units were extracted. The temporal evolution of 103 types of laboratory analysis and 2 static clinical data was used to develop and test a convolutive neural network (CNN), trained to predict potential organ donors. This model was compared to a non-temporal logistical model as a baseline. The CNN model was validated in a clinically distinct external population. To improve the performance in this external cohort, strategies to fine-tune the network were compared.

19 463 patients, including 397 potential organ donors, were used to create the model and 4 669 patients, including 36 potential organ donors, served as the external validation cohort. The CNN model performed better with an AUROC of 0.966 (IC95% 0.949-0.981), compared to the logistical model (AUROC de 0.940 IC95% 0.908-0.969, p=0.014). The CNN model was also superior in specific subpopulation of increased clinical interest. In the external validation cohort, the CNN model's AUROC was 0.820 (0.682-0.948) and could be improved to 0.874 (0.731-0.974) after fine tuning.

This promising model could change potential organ donors' detection for the better. More studies are however required to improve the model, by adding more types of data, and to validate prospectively the mode before routine clinical usage.

Keywords: organ donation, transplant, predictive model, machine learning, autoencoder, neural networks, transfer learning

Table des matières

Liste des tableaux

Liste des figures

Liste des annexes

Liste des sigles et abréviations

AUROC : Aire sous la courbe fonction d'efficacité du récepteur (de l'anglais *Area Under the Recieving Operating characteristic Curve)*

AVC : Accident Vasculaire Cérébral

CITADEL : Centre d'intégration et d'Analyse en Données Médical du CHUM

CHUM : Centre Hospitalier de l'Université de Montréal

DDC : Don après décès cardiocirculatoire

DDN : Diagnostique de décès neurologique

ECMO-VV : Oxygénation par membrane extra-corporelle veino-veineuse (de l'anglais *Extracorporal Membrane Oxygenation)*

ODO : Organisme de Don d'Organes (p.ex : Transplant Québec)

RNA : Réseaux de Neurones Artificiels

ROC : Fonction d'efficacité du récepteur (de l'anglais *recieving operating characteristic)*

SADC : Système d'aide à la décision clinique

SI : Soins intensifs

SIC : Soins intensifs coronariens

TQ : Transplant Québec

À Valérie, Clara et Juliette

Remerciements

Cette maitrise est née d'une envie qui date d'il y a plus de dix ans. Alors finissant en génie biomédical à Polytechnique Montréal et appliquant en médecine, je rêvais aux avancées potentielles que pourraient apporter la collaboration des deux disciplines. Quoi de mieux comme démonstration que l'utilisation de l'apprentissage machine en transplantation d'organes!

Mon premier remerciement va à mon superviseur, Dr Michaël Chassé. Un chercheur et intensiviste d'exception qui a partagé dès le début mon intérêt pour les données massives et leur utilisation dans le but d'améliorer la réalité clinique. Son soutien et son mentorat ont été nécessaires au développement de ce projet et à celui de ma carrière d'anesthésiste-intensiviste.

Je tiens aussi à remercier l'équipe de CITADEL sans qui rien n'aurait été possible. Leur travail acharné pour mettre de l'ordre dans le désordre, pour lier et valoriser d'innombrables bases de données est colossal. Un merci particulier à Rima Bouchakri qui m'a énormément appris sur la science des données.

Merci au Programme de Recherche en Don et Transplantation du Canada, qui a financé ce projet par une bourse de recherche en innovation.

Merci aussi à Dr François Martin Carrier pour son mentorat et nos discussions sur l'anesthésie, les statistiques, la recherche et les soins intensifs. Elles ont toutes été précieuses.

Merci particulier à mon frère, Dr Michaël Sauthier, qui a été un mentor dans l'enfance, dans la médecine et maintenant dans la recherche. Merci aussi à mes parents, beaux-parents, ma sœur et mes belles-sœurs qui ont été présents pour moi et pour ma famille et qui continuent de m'encourager et m'appuyer dans mes nombreux projets.

Finalement, merci à Valérie Reinhardt, ma conjointe depuis tant de belles années. M'entendre parler de données, de dons d'organes et d'apprentissage machine aussi souvent n'a pas dû être toujours palpitant, mais ta compréhension, ton amour et ton soutien ont été indispensables à la réussite de ce projet. Ensemble, nous entrainons les plus complexes et les plus beaux des réseaux de neurones que sont nos filles, Clara et Juliette.

Structure du mémoire

Ce mémoire est écrit par article. Il comprend deux articles qui sont ou seront soumis pour publication dans des revues révisées par les pairs. Bien que le mémoire soit rédigé en français, les deux articles sont rédigés en anglais en vue de leur publication. À noter qu'il s'agit d'un domaine émergent et que certains termes n'existent pas encore en français et sont donc utilisés en anglais pour conserver un langage précis. Afin de faciliter leur lecture, ces articles sont insérés dans ce mémoire en gardant leur propre structure et leur propre numérotation. Pour des raisons de limitations liées au logiciel de bibliographie, l'ensemble des références se trouvent dans une unique bibliographie située en fin de mémoire.

Le mémoire est séparé en 9 chapitres. Le premier chapitre introduit le projet de même que ses objectifs principaux et secondaires. Le second chapitre est une introduction à la réalité et aux enjeux du don d'organes au Canada, incluant une revue de la littérature sur l'épidémiologie des potentiels donneurs manqués. Le troisième chapitre est une revue non exhaustive des éléments méthodologiques importants de l'apprentissage machine, particulièrement pour les projets de science des données en santé. Le sujet de ce mémoire étant à la jonction d'un domaine technique et clinique, ce chapitre est de fournir, à une personne qui ne serait pas familière avec le domaine, une introduction à ce domaine très large. Il se termine par un survol des enjeux de l'utilisation de technique d'apprentissage machine en clinique. Le quatrième chapitre couvre la méthodologie de la préparation de la base de données, du développement des modèles et de leur validation. Le cinquième chapitre présente les résultats non publiés de préparation de la base de données et d'une première approche non supervisée. Le sixième chapitre présente le premier article, lequel couvre les résultats du développement et de la validation du modèle prédictif. Le septième chapitre présente les résultats du second article, qui couvre la validation externe du modèle, de même que des approches d'amélioration et de peaufinage de ce modèle. Le mémoire se termine avec un chapitre de discussion, de conclusion ainsi que les références et annexes.

1 Mise en contexte et question de recherche

Pour certaines maladies incurables en stade terminal, la transplantation d'organesreprésente l'unique option de traitement. Ces maladies sont occasionnellement foudroyantes, comme certaines hépatites fulminantes ou certaines myocardites. Plus souvent, ce sont des pathologies pulmonaires, rénales ou cardiaques chroniques, évoluant sur plusieurs années, pour lesquelles toutes les options thérapeutiques ont été essayées. Il existe quelques thérapies de support mécanique comme la dialyse, la ventilation mécanique ou encore le cœur artificiel, cependant aucune n'a la qualité d'un organe humain. À l'exception du cas particulier qu'est le don vivant, une transplantation d'organesrequiert de manière *sine qua non* qu'un autre être humain décède. Il n'y a qu'une poignée de décès permettant le prélèvement d'organes. En 2021, il y a eu 69 900 décès au Québec et seulement 724 (~1%) ont été référés à Transplant Québec (TQ)[1,2]. Des 724 références, seuls 144 (20%) ont été des donneurs effectifs[1]. Ces donneurs ont permis de greffer 208 reins, 33 cœurs, 72 poumons et 88 foies. En plus d'être une situation rare, les décès compatibles avec un don d'organes doivent se passer dans une situation contrôlée. Une détection précoce des décès compatibles permettrait un meilleur contrôle de la situation. Il s'agit de situations complexes et rares qu'il est impératif de détecter, des vies sont en jeu. Toutefois, de nombreuses données nationales et internationales suggèrent qu'un nombre non négligeable de potentiels donneurs d'organes sont manqués ou non référés aux organismes de transplantation[3–6]. Une amélioration du processus de détection et de référence est un des éléments qui a permi à l'Espagne, un des pays avec le plus haut taux de donneur par million d'habitant au monde, de doubler son taux de donneurs[7].

À l'opposé de ces situations hautement chargées en émotion se trouve la froideur rationnelle et calculée des données et de l'informatique. Ce monde a vécu une réelle révolution durant les dernières années. Le matériel informatique s'est amélioré et son coût a chuté. Cela a mené à l'explosion et la démocratisation d'algorithmes d'apprentissage machine très performants. Ils permettent d'utiliser des données complexes, de formats variés et en grandes quantités pour détecter des *patterns* subtils. Leurs réussites sont nombreuses dans des champs d'applications variés, notamment dans le domaine de la santé.

Une détection automatisée des potentiels donneurs d'organes pourrait permettre d'augmenter le nombre de références à TQ tout en surveillant en continu et à distance de nombreux sites hospitalier. Il existe actuellement un seul système remplissant cette fonction[8], mais son fonctionnement nécessite des données absentes de la quasi-totalité des dossiers patients informatisés. Le seul modèle prédictif basé sur des données cliniques usuelles utilisait des mots-clefs dans les rapports de radiologies, avec un succès intéressant, mais d'importantes limitations fonctionnelles[9].

Plusieurs éléments importants compliquent l'automatisation de ce processus, notamment la complexité du *pattern* clinique, la relative rareté des cas comparés au nombre total de patients hospitalisés et l'absence d'examen de dépistage. L'application d'algorithmes d'apprentissage machine sur des données cliniques simples et de routine pourrait permettre d'identifier ce type de *pattern* clinique. Ce projet est l'étape préliminaire d'une possible transformation dans le processus de détection et recrutement des donneurs d'organes.

1.1 Question de recherche

La question de recherche de ce mémoire était la suivante :

Est-ce qu'un modèle d'apprentissage machine basé sur un réseau de neurones, entrainé sur des données temporelles rétrospectives cliniques et administratives de routine, permet de mieux identifier les potentiels donneurs d'organes qu'une approche traditionnelle avec un modèle prédictif fréquentiste linéaire standard?

1.2 Objectifs

1.2.1 Principal

- Investiguer la capacité d'un modèle temporel de réseaux de neurones artificiels (RNA) de détecter automatiquement de potentiels donneurs d'organes (Chapitre 5, 6 et 7)

1.2.2 Secondaires

- Évaluer un modèle de RNA uniquement basé sur un autoencodeur avec erreur de reconstruction (Chapitre 5).
- Comparer la capacité prédictive d'un modèle temporel de RNA à un modèle prédictif fréquentiste linéaire non temporel (Chapitre 6).
- Valider le modèle des modèles dans un groupe de test séparé du groupe utilisé pour développer le modèle (Chapitre 6).
- Évaluer la performance des modèles dans les différents sous-groupes de potentiels donneurs d'organes (Chapitre 6).
- Évaluer la capacité prédictive précoce des modèles dans une fenêtre de 48h avant la fin des soins actifs aux soins intensifs (Chapitre 6).
- Mesurer la généralisation du modèle final dans une population externe (Chapitre 7).
- Évaluer l'ajout du transfert de connaissance pour améliorer la validité externe du modèle final (Chapitre 7).
- Améliorer la stratégie de transfert de connaissances et de peaufinage pour améliorer la validité externe du modèle (Chapitre 7).

1.3 Hypothèse

La première hypothèse était qu'un modèle prédictif temporel basé sur des RNA serait capable de détecter automatiquement de potentiels donneurs d'organes.

La seconde était que sa performance serait meilleure qu'un modèle prédictif fréquentiste linéaire standard.

La troisième hypothèse était que l'ajout de transfert de connaissance et de peaufinage améliorerait les performances du modèle prédictif temporel basé sur des RNA dans une cohorte de validation externe.

2 Introduction au décès et don d'organes

2.1 Décès et don d'organes : contexte clinique, statistique et légal

La majorité des dons d'organes proviennent de patients décédés (16.7 par million d'habitants en 2021[1] au Québec). Le don d'organes vivants est plus rare (8.7 par million d'habitants en 2021[1]) et est réservé aux greffes partielles de foie et aux greffes de rein. Bien qu'à première vue évidente, la définition du décès n'est pas si simple. Le Code civil du Québec précise que le décès doit être constaté par un médecin (art. 122 C.c.Q.). Toutefois, l'exigence dans le cas du don d'organesest plus haute et « le prélèvement ne peut être effectué avant que le décès du donneur n'ait été constaté par deux médecins » (art 45. C.c.Q.). Deux types de décès peuvent mener à un don d'organe. Le premier est constaté lorsque seul le cerveau décède alors que le reste des fonctions vitales sont maintenues artificiellement. Il s'agit du diagnostic de décès neurologique (DDN). Le second type de décès est constaté après une période de 5 minutes suivant un arrêt de la fonction cardiaque. Il s'agit du décès cardiocirculatoire, pouvant mener à un don après décès cardiocirculatoire (DDC). Il est important de préciser qu'il n'existe pas de définition légale de la mort. Un avis juridique signé conjointement par le bureau du coroner, l'association québécoise d'établissements de santé et de services sociaux et TQ suggère que le décès légal correspond au premier examen de mort neurologique tel que décrit ci-dessous et dans les protocoles de TQ[10]. De plus, la cour supérieure de l'Ontario s'est prononcée en 2018[11] sur le cas de *McKitty v Hayani* en confirmant l'existence de critères de décès neurologique, tout en statuant qu'il appartient au corps médical d'établir des lignes de conduite définissant la mort. La pratique médicale canadienne actuelle du don après DDN et DDC se base sur une consultation pancanadienne des différents intervenants ayant eu lieu en 2003[12,13].

TQ publie annuellement des statistiques détaillées[1,14,15]. En 2021 au Québec, le taux de donneurs par million d'habitants était de 16.7, ce qui est légèrement inférieur à l'Ontario (20.0) et à la moyenne canadienne (19.3). La situation québécoise se compare défavorablement à la France (19.1), aux États-Unis (41.7) et à l'Espagne (40.2), bien que les taux de mortalité

attribuables aux causes pouvant mener à un don d'organes soient relativement similaires (Fig. 1). Les raisons expliquant cette différence entre les pays sont nombreuses et complexes[16].

La réussite du modèle espagnol, grandement étudié comme une référence dans le domaine, peut en partie s'expliquer par un investissement de ressources logistiques et financières importantes, de même qu'un plan global pour l'ensemble du pays. Toutefois, l'identification précoce et proactive des potentiels donneurs est un des éléments centraux de leur réussite[7,16,17]. La différence de taux de donneurs d'organes suggère qu'une importante amélioration du processus est possible, notamment par une amélioration de la détection des donneurs potentiels manqués, non référés ou refusés.

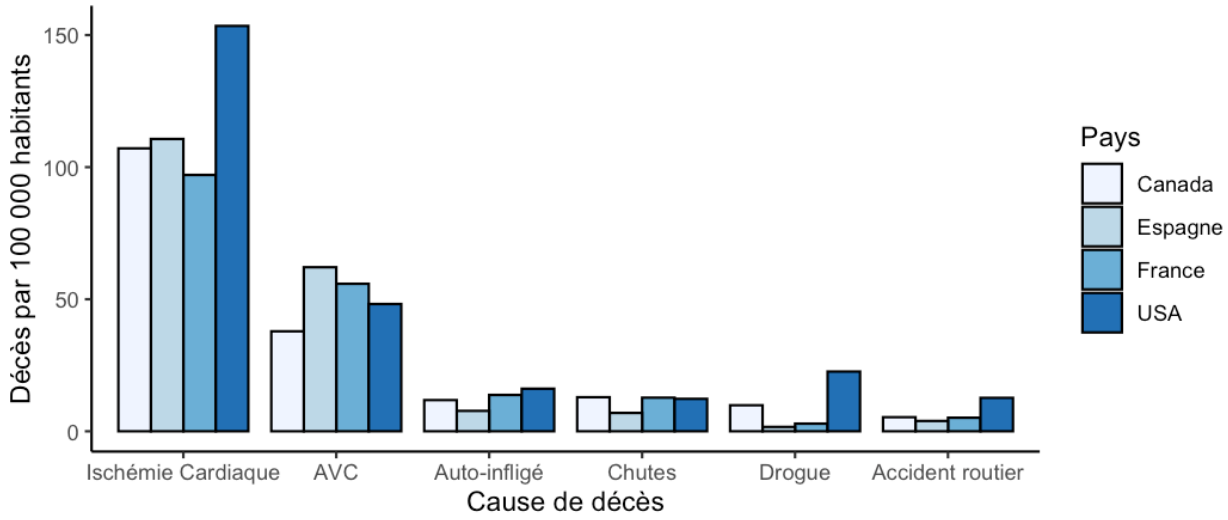


Fig. 1 Taux de décès par 100 000 habitants en 2019 de causes reliées au don d'organes, tiré de l'OMS[18]. AVC :Accident vasculaire cérébral.

Durant la seule année 2021, 724 patients ont été référés pour n'avoir, finalement, que 144 donneurs effectifs[1]. Les principales raisons de refus par TQ étaient une maladie préexistante (23%), un refus de la famille (26%) et un état neurologique ou hémodynamique ne permettant pas le don (20%). Pendant cette période, 888 patients étaient en attente d'une transplantation et 34 sont décédés durant l'attente. Sur une période de 9 ans (2012 vs 2021), la tendance de nombre de donneurs est à la hausse (120 donneurs vs 144). En conséquence, le nombre de personnes en attente a diminué (1250 vs 888). Le nombre de personnes décédées durant l'attente a lui aussi

diminué (69 vs 34), mais reste conséquent. Ces tendances sont similaires lorsqu'on regarde les données pan-canadiennes[19].

Outre l'impératif moral évident de favoriser le don d'organes, il y a aussi un argument économique. Pour l'insuffisance hépatique, la transplantation est la seule option. Pour l'insuffisance pulmonaire, les thérapies de remplacement (oxygénation par membrane extracorporelle par abord veinoveineux, ECMO-VV) coûtent en moyenne 70 000$ par utilisation[20] et son utilisation est limitée à quelques semaines. Pour l'insuffisance cardiaque, les cœurs mécaniques sont temporaires et ont aussi un rapport coût-bénéfices défavorable, avec un coût de 130 000$ pour l'appareil[21], excluant les coûts reliés à son installation, son suivi et ses complications. La thérapie de remplacement rénale est la seule qui puisse être utilisée sur le long terme. La greffe rénale a été toutefois clairement démontrée comme meilleure du point de vue de la qualité de vie et du rapport coûts-bénéfice. La greffe rénale coûte 40 000$ pour la chirurgie, et environ 8 000$ annuellement par la suite, comparée à la thérapie de remplacement rénal qui coûte 55 000$ par année par patient.

2.2 Don après diagnostic de décès neurologique

Le diagnostic de décès neurologique (DDN) est un diagnostic clinique. Il correspond à « l'expression clinique finale de la défaillance neurologique complète et irréversible »[13]. Un forum canadien de 89 experts a émis en 2003[22] des recommandations qui définissent l'examen clinique requis pour confirmer le DDN. Ces recommandations sont en accord avec les consensus internationaux les plus récents[23]. Cet examen ne peut être fait que chez des patients ayant un coma profond provenant d'une cause connue et irréversible. Il nécessite aussi d'exclure plusieurs facteurs confondants que sont une surdose de médicaments, une hypothermie, un désordre métabolique ou encore un choc persistant. L'examen, fait par deux médecins, confirme l'absence de fonction neurologique par l'absence de réflexes moteurs, l'absence de réflexes du tronc cérébral et l'absence de réflexe respiratoire. L'examen clinique est la manière de poser le diagnostic, bien que des tests auxiliaires de perfusion cérébrale puissent être utilisés si, et seulement si, une partie de l'examen ne peut être effectué. Il est donc primordial de comprendre

qu'il n'existe aucun test de laboratoire ou d'imagerie permettant de dépister ou de diagnostiquer le décès neurologique.

Les dons après DDN représentent la majorité des cas de don. En 2019, ils représentaient 79.9% (143/179) au Québec et 71% (580/820) au Canada[15,19]. Les statistiques de 2020 et 2021 de TQ ne différencient plus entre DDN et DDC, empêchant d'avoir des données plus récentes. Les étiologies principales (don après DDN et DDC confondus) sont l'accident vasculaire cérébral (AVC, 43%), l'anoxie cérébrale (26%) et le trauma crânien (18%). La tendance des 10 dernières années montre une stabilité du nombre de dons après trauma, mais une augmentation des dons après AVC (62 vs 77) et surtout des dons après anoxies (21 vs 46). Ce sont des signes d'une probable amélioration de la prise en charge précoce et de la reconnaissance de ces donneurs potentiels.

Le don d'organesaprès DDN est un processus qui dure plusieurs jours. Typiquement, le patient a un événement causal (par exemple un trauma, un AVC ou un arrêt cardiorespiratoire) qui nécessite une réanimation immédiate et une hospitalisation aux soins intensifs pour stabilisation et traitement de la cause. Cette période dure habituellement entre 24h et 48h. S'il n'y a pas d'amélioration, la famille est habituellement avisée du pronostic neurologique et vital faible. C'est à ce moment que l'équipe clinique doit reconnaitre la possibilité d'un don d'organes après DDN et en discuter avec la famille. Si le patient avait rédigé des directives anticipées à ce sujet, ou si la famille démontre une ouverture, une discussion avec TQ peut débuter au besoin pour aider l'équipe clinique et la famille au bon déroulement du processus. Lorsque le DDN est prononcé, le cas peut être officiellement référé à TQ (ou équivalent selon la province). Si la famille va de l'avant avec le don, les fonctions vitales du patient décédé sont maintenues le temps d'effectuer des tests de laboratoire et d'imagerie permettant d'exclure des pathologies infectieuses ou néoplasiques, d'évaluer la qualité des organes et de préparer les receveurs. Au terme de ce processus, le corps du patient est amené en salle d'opération pour prélever les organes. Le corps est ensuite ramené à la famille pour les rites funéraires. Occasionnellement, il arrive que le patient soit transféré en état de mort neurologique vers un autre hôpital spécialisé, si le centre qui a admis initialement le patient n'est pas un centre de prélèvement d'organes.

2.3 Don après décès cardiocirculatoire (DDC)

Bien que le décès cardiocirculatoire soit la forme la plus fréquente de décès, le don d'organesdans cette situation est relativement nouveau. La première greffe de rein suite à un DDC a eu lieu en 2007 et le premier protocole a été publié en 2011. Toutefois, c'est la forme de don qui a le plus augmenté sur 10 ans tant au Québec (7 en 2010 et 36 en 2019, +414%) qu'au Canada (42 en 2010 et 240 en 2019, +471%)[14,19].

Un forum d'expert a défini, en 2005, des lignes de conduite canadiennes encadrant cette pratique[12]. Le DDC s'applique aux situations où « le décès est appréhendé, mais n'est pas encore survenu ». Il doit s'agir de situations où il y a une atteinte sévère sans possibilités de guérison nécessitant des thérapies de maintiens des fonctions vitales (vasopresseurs, ventilation mécanique ou ECMO). Il doit être prévu d'arrêter ces thérapies et le décès doit être attendu rapidement après l'arrêt des thérapies. Contrairement au DDN qui est un diagnostic clinique, le DDC est un processus. Le patient est amené dans une salle d'opération où des soins de conforts sont administrés. Les thérapies de support sont retirées et le patient progresse éventuellement vers un arrêt cardiocirculatoire. Un délai de 5 minutes doit être observé avec une absence complète de pouls, de pression artérielle et de respiration. Ce délai est prévu pour confirmer l'irréversibilité de la situation. Une fois les 5 minutes complétées, le prélèvement des organes peut débuter. La période entre l'arrêt des thérapies de maintien et le prélèvement est appelé ischémie chaude et influe sur la qualité des organes. Il est critique de choisir des patients avec des pathologies suffisamment sévères pour que ce délai ne dépasse pas 1 à 2 heure, ce qui rendrait les organes non éligibles au prélèvement. Préalablement au processus de DDC, le processus de discussion avec la famille et de références à TQ est le même que décrit dans la section précédente.

Sur les 227 DDC entre 2007 et 2019 au Québec, 62% provenaient d'une pathologie neurologique centrale (AVC ou anoxie). Il s'agit typiquement de patients avec une atteinte neurologique sévère, mais chez qui persistent certaines fonctions vitales empêchant un DDN. 24% des cas étaient secondaire à un trauma. Finalement, le DDC après une aide médicale à mourir

arrive en troisième (10%), en augmentation marquée depuis 3 ans, moment du début de l'aide médicale à mourir.

2.4 Potentiels donneurs d'organes et référence à Transplant Québec

En 2021, il y a eu 69 900 décès au Québec[2]. 75% des décès ont lieu dans un centre hospitalier[24] et environ 40% des décès ont comme cause une néoplasie ou une infection disséminée[25], les rendant non éligibles à devenir donneurs d'organes. De l'ensemble des décès restants, environ 40 000, le nombre exact de potentiels donneurs est inconnu. Toutefois, tel que mentionné ci-dessus, seuls 724 patients (1% des décès) ont été référés à Transplant Québec (TQ)[1] et de ceux-ci seuls 144 ont pu donner leurs organes.

Pour améliorer le taux de référence aux organismes de don d'organes(ODO) tel que TQ, un groupe d'experts a émis des recommandations pour le dépistage des potentiels donneurs d'organes au Canada[26]. Se basant sur une revue des différents critères existants dans la littérature[27], ils ont défini un potentiel donneur d'organes comme un patient nécessitant un support ventilatoire mécanique (invasif ou non invasif), ayant une pathologie avec pronostic vital faible à inexistant et pour lequel un passage en soins de confort est décidé, mais non complété. Ils ont recommandé que le don d'organessoit une part intégrante des soins de fin de vie, pour les bienfaits altruistes du don d'organes, mais aussi pour le respect des possibles souhaits du patient et pour la facilitation que cela apporte au processus de deuil.

De plus, le même groupe d'experts a recommandé que tous les donneurs d'organes potentiels soient référés à un ODO pour améliorer le suivi et les chances de succès du processus. Le projet au cœur de ce mémoire vise à développer un modèle prédictif, permettant le développement d'un outil informatisé qui pourrait automatiser et faciliter cette recommandation.

Finalement, il est important de différencier deux termes qui sont utilisés dans ce mémoire : le donneur potentiel « non référé » et le donneur potentiel « manqué ». Trois raisons peuvent expliquer qu'un patient ne soit pas référé à TQ.

1. Une contrindication évidente au don d'organes(cancer récent, actif ou métastatique, infection disséminée ou dysfonction multi-organes). Ces patients ne sont pas des donneurs potentiels.
2. Un patient non reconnu par le clinicien comme potentiel donneur. Ce type de patient est un donneur potentiel manqué.
3. Un patient reconnu comme potentiel donneur, mais que le clinicien a choisi de ne pas référer, quelle que soit la raison (contrindications selon lui, refus familial, etc.). Ces donneurs potentiels sont non référés, mais pas manqués.

2.5 Épidémiologie des potentiels donneurs non référés et manqués

Malgré une amélioration de l'ensemble du processus de détection des potentiels donneurs d'organes, de nombreux potentiels donneurs d'organes sont manqués. Les études quantifiant ce nombre varient beaucoup en matière de méthodologie (audit de décès ou extrapolation statistique de données populationnelles), de population (unicentrique ou non, soins intensifs, urgence ou tous patients), de critères d'inclusion et d'exclusion (catastrophe neurologique, pathologies neurologiques ou tous décès) et d'issues (donneurs potentiels, références à un ODO ou donneurs confirmés).

Au sujet des patients non référés, deux études canadiennes et une australienne permettent un aperçu de l'épidémiologie. Tout d'abord, Krmptoc et coll.[3] ont fait un audit des décès survenus entre 2013 et 2015 dans une des unités de soins intensifs et d'urgence de l'Ontario (total de 72 hôpitaux). Sur les 1 407 décès chez des patients ventilés pour lesquels le passage en soin de confort a été fait, 760 (54%) semblaient médicalement compatibles avec une référence à l'ODO. 438 (31.1%) ont été approchés pour un don d'organes. De ceux-ci, 208 (14.8%) ont accepté. En fin de compte, seuls 119 (8.5%) sont devenus des donneurs confirmés. Des 760 potentiels donneurs d'organes, 257 n'ont pas été référés à l'ODO de manière approprié (non référé, ou référé pendant ou après l'arrêt de soins). Aucune approche aux familles en vue d'un don d'organesn'a été documentée pour 251 de ces 257 potentiels donneurs d'organes. Cela correspondrait donc à un taux de patient non référé à 33,8%. En appliquant les mêmes

proportions d'attrition, les auteurs estimaient que la référence de l'ensemble de ces donneurs potentiels aurait pu augmenter le nombre de donneurs de 119 à 217, une augmentation de 82%.

Ensuite, Opdam et coll.[28] ont procédé à un audit de décès de 12 hôpitaux de l'état de Victoria en Australie sur une période d'un an. Des 5 551 décès, ils ont isolé 112 décès avec un potentiel de DDN. Les familles de 66 (59%) d'entre eux ont été approchées pour don d'organe. Cela correspondrait à un taux de « non référés » de 41%. L'article comprenait aussi une brève revue de la littérature indiquant qu'entre 30% et 75% du total de potentiels donneurs après DDN seraient non réalisés. À noter que seuls les dons après DDN étaient considérés.

Finalement, Kutsogiannis et coll.[4] ont procédé à un audit de décès aux soins intensifs et à l'urgence en Alberta, pour la période 2010-2013. Des 2 931 décès, 227 sont définis comme donneurs potentiels « non référés », avec un taux de 7.7%. Ils n'indiquaient pas combien de références ont été faites à l'ODO durant la même période. Ils ont par la suite examiné les dossiers de l'ensemble des 227 patients. Ils estimaient de manière conservatrice que, si l'ensemble des donneurs avaient été approchés et traités de manière appropriée, cela aurait pu se traduire par une augmentation d'environ 42% des dons après DDN et de 7% des DDC.

L'institut canadien d'information sur la santé (ICIS), basé sur des données du registre canadien des insuffisances et des transplantations d'organes (RCITO), a publié en 2014 un rapport étudiant le potentiel de don d'organesau Canada[29]. Ils estimaient qu'en 2012 il y aurait eu 3 088 potentiels donneurs d'organes pour seulement 520 donneurs. À noter que leur méthodologie était propice à la surestimation. Leur estimation conservatrice était de 1 544 donneurs potentiels. Ils utilisaient le taux de conversion (ratio entre donneur confirmé et potentiel) comme méthode de comparaison des sous-groupes. Globalement, leur estimation de taux de conversion était de 17% (34% conservateur). Ils notaient de grosses disparités dans certains sous-groupes. Ainsi, le taux de conversion était meilleur en DDN (30%) qu'en DDC (5%). Il était aussi meilleur chez les moins de 50 ans (30%) par rapport aux 60-69 ans (7%). Finalement, le taux de conversion variait par province et était meilleur au Québec (21%) qu'au Manitoba (10%). Ces disparités sont des sources de potentiel amélioration du processus de don d'organes. Leur estimation finale était d'un taux de potentiels donneurs d'organes de 89-107 donneurs par

million d'habitants, un taux presque sept fois plus élevé que le taux de donneurs au Canada la même année (15.5 par million d'habitants).

Plusieurs autres études provenant d'autres régions du monde se sont penchées sur la question, mais aucune ne fonctionnait avec la manière actuelle de référence vers un ODO. L'indicateur de qualité le plus répandu était le taux de conversion (ratio entre donneur confirmé et potentiel). La définition de potentiels donneurs était variable et peu se penchaient sur la raison du manquement. Une étude des Pays-Bas d'audit de décès par catastrophe neurologique[30] rapportait un taux de conversion de 46%. Une autre étude ontarienne avec une approche similaire rapportait un taux de conversion de 35%[31]. Une troisième rassemblant la totalité des catastrophes neurologiques d'Islande rapportait un taux de conversion de 92%. Une dernière étude finlandaise, toujours basée sur un audit de décès par catastrophe neurologique, montrait un taux de conversion étonnamment bas de 13%. O'Brien et coll.[32] suggéraient un taux de conversion de 58% dans un des réseaux hospitaliers d'Australie. Procaccio et coll. suggéraient un taux de conversion global de 78% en Italie avec d'importantes variations par région[33].

Certaines de ces études se sont penchées sur l'origine des potentiels donneurs manqués. La salle d'urgence semblait être une source importante de potentiels donneurs manqués[34,35], bien que cela était variable selon l'endroit et la vitesse à laquelle les patients ventilés sont transférés vers une unité de soins intensifs[5,36]. Peu d'informations étaient disponibles sur les potentiels donneurs d'organes manqués sur les étages standards[37]. Toutefois, le recrutement de ces potentiels donneurs exigerait une intubation et un transfert aux soins intensifs de patients mourants, uniquement dans le but du don d'organes, soulevant plusieurs enjeux éthiques, d'acceptabilité et organisationnels. Finalement, il semblait y avoir une différence notable entre les centres de transplantation et les autres hôpitaux. Une étude ontarienne[31] montrait un taux de 5% de donneurs par décès dans les centres transplanteurs, par rapport à 1.4% dans les autres hôpitaux ontariens. Lorsque le bassin de décès était réduit à de potentiels donneurs d'organes, donnant un taux de conversion, la différence réduisait, mais restait notable: 30% (centre transplanteurs) par rapport à 20% (centres non-transplanteurs).

En somme, il semble clair qu'un nombre possiblement important de potentiels donneurs d'organes sont manqués ou non référés. Le taux de conversion variait beaucoup selon la

population et la méthodologie, mais pointe vers un potentiel important d'amélioration du processus. Cette amélioration peut se faire à plusieurs niveaux, notamment au niveau de la référence aux ODOs[29]. Le taux de patient non référé était plus difficile à caractériser, mais semblait se situer entre 30% et 60% dans les études basées sur des cohortes rétrospectives. Ce taux, possiblement surestimé, reste certainement substantiel. Un chiffre de cet ordre de grandeur se confirmait dans l'ensemble des pays et régions étudiés. Ces potentiels donneurs semblaient manqués en majorité aux soins intensifs et à l'urgence et de manière plus marquée dans les centres non-transplanteurs. Pour assurer le succès d'un système automatique de référence des potentiels donneurs, ce système devrait être transférable dans des hôpitaux non transplanteurs, là où il serait le plus utile.

3 Revue de l'apprentissage machine et de la science des données médicales

L'apprentissage machine est un domaine qui s'est développé énormément dans les dernières années. Ce chapitre est une brève introduction à ce sujet complexe et une revue des éléments importants qui seront utilisés dans la méthodologie du projet. Le but est de de vulgariser, pour une personne qui ne serait pas familière avec le domaine, les bases théoriques nécessaires à la compréhension des problématiques et des enjeux d'un projet de d'apprentissage machine en santé.

3.1 Définitions

3.1.1 Intelligence artificielle et apprentissage machine

Bien que les deux termes soient souvent utilisés de manière interchangeable, leur définition diffère légèrement. Le terme « intelligence artificielle » est un terme moins précis et utilisé autant pour un sujet de recherche actif, que pour de la science-fiction ou du marketing. Cette imprécision, et le fait qu'il ait été grandement galvaudé durant les dernières années, le rend moins désirable dans un mémoire de recherche. Le terme « apprentissage machine » est plus technique et fait référence à « la capacité d'un système [informatique] d'acquérir sa propre connaissance par l'extraction de *patterns* parmi les données brutes »[38]. Cela se différencie de la programmation classique, dans laquelle le programmeur définit une série de règles rigides. En apprentissage machine, le programmeur choisit l'algorithme. Chaque algorithme « apprend » de manière différente à l'aide d'un mécanisme de rétroaction et par l'ajustement automatique de ses propres poids. L'apprentissage machine, à mi-chemin entre statistique et informatique, englobe un spectre très large d'algorithmes. À un extrême se trouve une simple régression logistique et à l'autre un réseau de neurones d'une grande complexité. Cette complexité peut parfois se traduire par le nombre de paramètres utilisés, bien qu'il n'existe pas de classification bien établie de complexité[39]. À des fins de simplification, la majorité des algorithmes peuvent être classés en deux grandes familles : apprentissage supervisé et non supervisé.

3.1.2 Apprentissage supervisé

La première famille est la plus connue et la plus utilisée des deux. Le but de ces algorithmes est de s'entrainer à prédire la relation entre des données et une issue qui est connue. Cette issue, appelée *outcome* en anglais, est le résultat de la prédiction de l'algorithme. L'issue est parfois un élément catégoriel, comme dans le cas de la régression logistique, le perceptron, l'algorithme *support vector machine* (SVM), l'algorithme des « k » plus proches voisins ou encore les arbres de décisions. Cette issue peut aussi être continue, comme dans le cas de la régression linéaire ou certains réseaux de neurones. Tous ces algorithmes ont en commun de nécessiter, pour leur entrainement, un jeu de données pour lequel on connait à la fois les données à entrer dans l'algorithme et l'issue qu'on veut prédire. À travers l'apprentissage, les paramètres de l'algorithme choisis sont optimisés vers une solution idéale liant les données et l'issue.

Le modèle développé dans ce mémoire est un exemple d'apprentissage supervisé : on utilise une base de données de résultats de laboratoire pour prédire le statut de potentiel donneur d'organes.

3.1.3 Apprentissage non supervisé

La seconde famille rassemble une variété d'algorithmes qui ont comme propriété de ne pas nécessiter d'issue connue lors de leur apprentissage. Ils peuvent servir notamment à estimer des paramètres dans les données, le calcul de la moyenne et de l'écart type. Une autre application est le regroupement des données en fonction d'un critère de similitude, comme l'algorithme des k-moyennes. Une troisième application est la réduction de dimension. En effet, les projets de recherche actuels utilisent des données de grande complexité avec beaucoup de dimensions et de paramètres. La réduction du nombre de dimensions permet une visualisation humainement compréhensible de données autrement trop complexes. On peut ainsi réduire un ensemble de données qui possède de nombreuses dimensions en une image standard en deux dimensions. Pour cette utilisation, deux algorithmes populaires sont l'analyse du composant principal (PCA) et la méthode du *t-distributed stochastic neighbor embedding* (t-SNE). Finalement, une dernière application est l'extraction de paramètres. Certains algorithmes permettent de transformer les données d'une forme en une autre, pour en extraire des informations significatives. Un exemple

serait celui du résumé d'un film. Il y a transformation d'une donnée visuelle temporelle en un court texte. Cette transformation peut se faire de manière non supervisée, vise à une perte d'information minimale et le résultat permet des applications qui auraient été infaisables sur la vidéo brute.

Une partie du modèle développé dans ce mémoire utilise un algorithme non supervisé, l'autoencodeur, tel que décrit à la section 3.3.4.

3.2 Modèles prédictifs : approche classique, épidémiologique traditionnelle et apprentissage machine

Les modèles prédictifs font partie intégrante de la médecine depuis de nombreuses années. On peut notamment penser aux scores cliniques comme le MELD[40], l'APACHE[41], lesquels sont basés sur des régressions logistiques. L'apport de l'apprentissage machine à la prédiction en clinique est un phénomène nouveau et, si l'on se fie au nombre de publications répertoriées sur PubMed, il s'agit d'un phénomène exponentiel avec un temps de doublement d'environ deux ans. Pour des chercheurs cliniques habitués à des décennies de modèles linéaires, la valeur ajoutée de l'apprentissage machine dans les modèles prédictifs clinique reste à prouver.

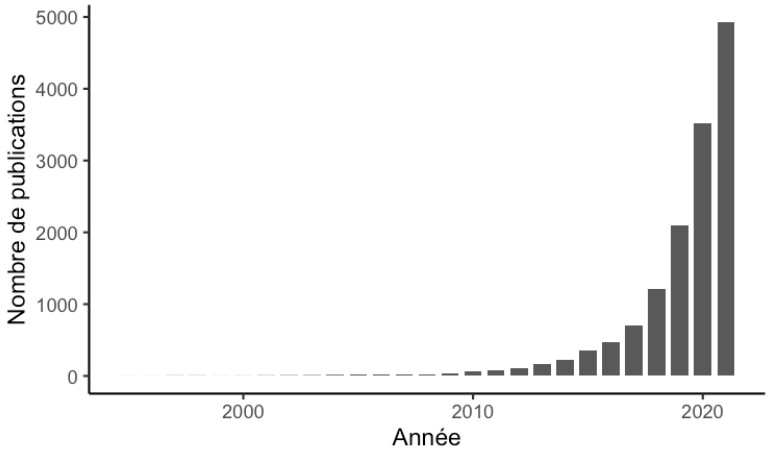


Fig. 2 Résultats PubMed pour ((machine learning) OR (deep learning)) AND (clinical prediction)

3.2.1 Comparaisons des modèles prédictifs « traditionnels » aux modèles d'apprentissage machine.

Pour créer un modèle clinique prédictif, il apparait intéressant de mettre en comparaison « approche traditionnelle » et « apprentissage machine ». Toutefois, bien que la question soit importante, obtenir une réponse claire est complexe. En effet, aucun consensus n'existe pour définir précisément la frontière entre ces deux approches conceptuelles puisqu'il s'agit d'un continuum[42]. De plus, les « approches traditionnelles » sont basées sur des régressions, qui sont, au final, une forme élémentaire d'apprentissage machine.

Christodoulou et coll. [43] se sont attardés à cette question dans une revue systématique. Ils concluaient qu'en comparant régression logistique et apprentissage machine, pour des articles de prédictions cliniques à bas risque de biais publiés entre 2016 et 2017, il n'y avait pas de différence d'aire sous la courbe ROC (AUROC). Toutefois, cette revue systématique souffre de plusieurs limitations méthodologiques qui, sans formellement invalider leur conclusion, limitent fortement la puissance et l'utilité de leur étude.

Tout d'abord, la question posée est très critiquable. Ils mettent en opposition deux approches qu'ils présentent comme opposées, sans expliquer ce qui les opposent. De plus, leur définition de « régression logistique » est arbitraire. La méthodologie requise pour développer une « vraie » régression logistique épidémiologique nécessite un travail considérable[44] (paramètres, analyse d'interaction, analyse de résidus, etc.). Le nombre d'études qui suivaient cette méthodologie rigoureuse est inconnu. Par ailleurs, ils ont choisi d'inclure les régressions logistiques avec des éléments de régularisation (lasso, *ridge*, etc.) alors qu'il s'agit de techniques qui s'apparentent à l'apprentissage machine. Ils y opposent les algorithmes d'apprentissage machine comme s'il s'agissait d'un tout, mettant en commun des techniques radicalement différentes en ce qui concerne leur approche et leur complexité (arbre de décisions, réseaux de neurones, etc.).

De plus, même si leur dichotomisation des algorithmes était fondée, le choix d'utiliser une régression logistique comme concurrent aux algorithmes d'apprentissage machine met en lumière une divergence conceptuelle dans l'utilisation des algorithmes prédictifs. Le statisticien

émérite américain Leo Breiman, dans une publication d'importance[45], revient sur cette différence de culture. Bien que l'article date de 2001, son contenu et ses conclusions sont d'autant plus d'actualité aujourd'hui. Il met en opposition l'approche dominante de modélisation des données, avec l'approche de modélisation par algorithme. La première a pour but d'expliquer la relation entre des variables fixes et une issue, en modélisant un bruit comme stochastique. Le but de cette approche, plus épidémiologique, est de démontrer un lien statistique d'association entre exposition et issue. L'interprétabilité du modèle, de même que la rigueur mathématique sous-jacente, prévaut sur la précision de la prédiction. C'est le courant dominant actuel en sciences cliniques. L'approche de modélisation par algorithme, quant à elle, ne cherche pas nécessairement (quoique cela soit un domaine de recherche en effervescence) à modéliser le lien entre exposition et issue ni à inférer sur la causalité, mais plutôt à prédire avec précision l'issue. Dans cette approche, la précision de la prédiction prévaut sur l'interprétabilité du modèle. C'est le champ d'application dans lequel s'est développé l'apprentissage machine. Mettre en opposition une approche logistique traditionnelle avec une approche plus complexe d'apprentissage machine revient donc à mettre en opposition deux outils dont l'utilité diffère.

Troisièmement, ils ont choisi de comparer les deux approches en se basant uniquement sur l'AUROC comme mesure de performance d'un modèle, ce qui est un peu réducteur. Bien qu'il s'agisse d'une méthode fréquemment utilisée, elle est influencée par des régions moins utiles de la courbe ROC et ne rapporte que la discrimination. La méthodologie de développement de modèle prédictif TRIPOD[46,47] recommande de présenter une mesure de discrimination, mais aussi de calibration et de performance clinique (sensibilité, spécificité, valeur prédictive positive ou négative).

Ensuite, cette revue de la littérature se limite aux articles ayant utilisé des types de données utilisables à la fois par une régression logistique et une approche d'apprentissage machine, ce qui est une fois encore réducteur. Cette approche induit un biais de sélection différentiel. L'utilité la plus impressionnante des approches d'apprentissage machine est justement d'utiliser des types de données complexes, qu'une approche traditionnelle ne peut pas utiliser (image, temporelle, haute dimension, etc.).

Finalement, les auteurs ont choisi de limiter la période d'étude à 2016-2017 menant à un relativement petit nombre d'études empêchant de faire des analyses de sous-groupes par approche d'apprentissage machine qui auraient été indispensables. Ces deux approches sont des outils de prédiction différents et il semblerait étonnant qu'un outil soit systématiquement supérieur à un autre dans la totalité des scénarios. Dans certains cas, une approche plus simple (régression logistique ou autre) pourrait surpasser une approche complexe alors qu'une approche d'apprentissage machine plus complexe apparait utile pour gérer des données massives[42]. De plus, à l'image de la météo, le rapport signal sur bruit est parfois tel qu'aucune prédiction n'est possible, peu importe l'approche utilisée. Une étude de sous-groupe par type de scénario clinique, par complexité d'algorithme et par quantité de données, aurait été nécessaire pour pouvoir répondre de manière convaincante à la question.

Ils démontrent toutefois le besoin d'une plus grande rigueur méthodologique[46,48] et suggèrent que « l'apprentissage machine » n'est pas une panacée, particulièrement dans des situations avec des données tabulaires simples. Le fait qu'une revue systématique soit nécessaire pour prouver un point qui peut sembler évident met en lumière les attentes irréalistes qu'a suscitées l'engouement pour ce domaine[49].

En conclusion, la question qu'ils soulèvent est une question d'importance. Cependant, leurs choix méthodologiques empêchent d'arriver à une conclusion et n'apportent, finalement, que peu d'information utilisable en pratique. D'un côté, l'utilisation d'un algorithme d'apprentissage machine reste pertinent dans un scénario de prédiction clinique. D'autre part, l'utilisation d'une approche logistique, sans être une régression logistique formelle, reste pertinente en modélisation par algorithme. En effet, la complexité n'étant pas un gage de performance, il est important d'utiliser un modèle plus simple comme référentiel, comme contrôle et comparatif pour le modèle plus complexe.

3.3 Réseaux de neurones

3.3.1 Bref historique

En 1943, MCulloch et Pitts développent le concept du neurone informatique à l'aide de circuit électrique[38] en se basant librement sur un neurone biologique. Il s'agissait d'un petit circuit qui faisait une somme pondérée de plusieurs valeurs et additionnait une valeur de biais. Le résultat était ensuite transformé par une fonction d'activation pour fournir une valeur finale. Dans le cas du neurone de Mculloch-Pitts, il s'agissait de la fonction de Heaviside qui vaut « 0 » pour une valeur négative et « 1 » pour une valeur positive. Cette idée a été reprise par Rosenblatt en 1958[50]. Il a développé une manière d'apprendre automatiquement les poids de la somme pondérée et le biais. Cette architecture relativement simple reste aujourd'hui pertinente. En effet, les RNA actuels sont constitués d'un grand nombre de ce type de neurone, interreliés et disposés en couches les unes sur les autres, en remplaçant simplement la fonction d'activation Heaviside par une plus complexe[51] (ReLU, sigmoïde, tanh, etc.). On peut aussi faire un parallèle avec la régression logistique, laquelle est une forme de neurone artificielle dont la fonction d'activation est une fonction logistique.

Avec l'amélioration notable des techniques de programmation et du matériel informatique, les RNA ont été au cœur d'une révolution numérique qui a pu être constatée notamment par la victoire de l'ordinateur sur l'humain à de nombreux jeux de société (Échecs, Jeopardy, Go, Super Mario, etc.) ou par l'émergence de la conduite autonome. L'implication d'entreprises majeures (Google, Microsoft, Facebook, Apple, Amazon, etc.) a eu un effet d'accélérateur et a engendré une expansion majeure de ce domaine.

Au cœur de cette révolution reste la capacité extraordinaire des RNA pour la reconnaissance de *patterns*. C'est particulièrement vrai dans des situations qui sont faciles pour un être humain, mais qui sont excessivement complexes à faire en programmation classique. Un exemple est la reconnaissance d'images. Le cas prototype est celui de la compétition ImageNet[52] dont le but est de classifier un million d'images appartenant à milles catégories différentes. La totalité des données d'entrainement comprend plus de quatorze millions d'images. Le but est de fournir pour chaque image cinq choix de ce que l'image contient. Il s'agit d'une erreur si la vraie

réponse ne se trouve pas dans ces cinq choix. Classifier ainsi un million d'images serait, pour un humain, une tâche excessivement laborieuse, mais facile (5% d'erreur); le moindre enfant d'âge scolaire aurait probablement un faible taux d'erreur. Toutefois, avant l'arrivée des RNA, les meilleurs algorithmes avaient un taux d'erreur de 25%-30%[53]. En 2012, AlexNet, le premier réseau de neurones utilisé dans cette compétition, a eu un taux d'erreur de 16,4%. Cette avancée magistrale a mis les RNA sous la lumière des projecteurs. Les réseaux actuels ont un taux d'erreur de 2%, inférieur à l'humain.

3.3.2 Réseaux denses

Un réseau de neurones est constitué de nombreux neurones informatiques interreliés, à l'image d'un cerveau. Les différents neurones sont organisés en couches. Le nombre de neurones par couches et le nombre de couches sont au choix du créateur du réseau. La sortie de chaque neurone d'une couche devient l'entrée de la couche suivante. Il y a toujours une couche d'entrée et une de sortie. La forme de la couche de sortie permet de spécifier le type de donnée produite par le réseau (catégorielle, continue, binaire, etc.). Entre ces deux couches se trouve un nombre arbitraire de couches dites cachées. La Fig. 3 illustre un exemple de réseau avec trois données en entrée, trois couches cachées de cinq neurones chacune et une sortie avec une seule valeur.

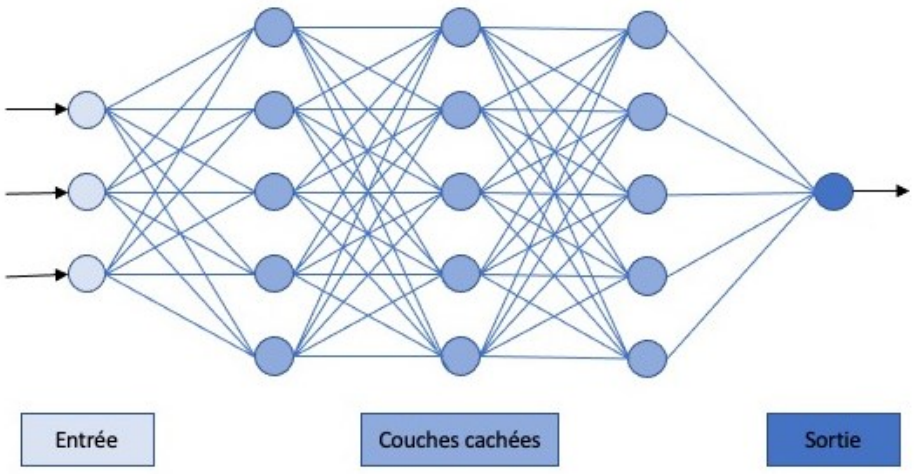


Fig. 3 Exemple de réseau avec trois couches cachées et une de sortie

Sur la figure, chaque ligne correspond à un poids qui doit être appris par le réseau. Plus le nombre de neurones choisi est grand, plus la complexité du réseau est grande et plus le nombre

de poids est grand. Pour entrainer un réseau complexe, il faut un grand nombre de données. Un nombre insuffisant de données mènera le réseau à apprendre le groupe d'entrainement (sur apprentissage ou *overfitting)* au lieu d'apprendre les structures ou *patterns* sous-jacents.

3.3.3 Réseaux convolutifs

Les réseaux convolutifs sont un une forme particulière de réseau dense, particulièrement importants en imagerie. La différence par rapport à un réseau dense usuel est qu'au lieu de faire une somme pondérée comme opération mathématique de base, ils font une autre opération mathématique appelée une convolution.

La convolution est un concept mathématique complexe qui permet notamment d'accentuer certains *patterns* dans une image ou dans une série temporelle à l'aide d'une matrice appelée « masque ». Ce masque est habituellement choisi arbitrairement. Selon le masque choisi, différents aspects de l'image seront accentués (lignes verticales, lignes horizontales, courbes, etc.). Dans un RNA convolutif, le contenu du masque n'est pas choisi, mais plutôt appris par le réseau lors de l'entrainement. La Fig. 4 illustre la structure du réseau convolutif AlexNet, un réseau qui reste relativement simple par rapport aux réseaux plus récents. Il comporte cinq couches convolutives successives (en turquoise) et trois couches denses (mauve et orange) par la suite.

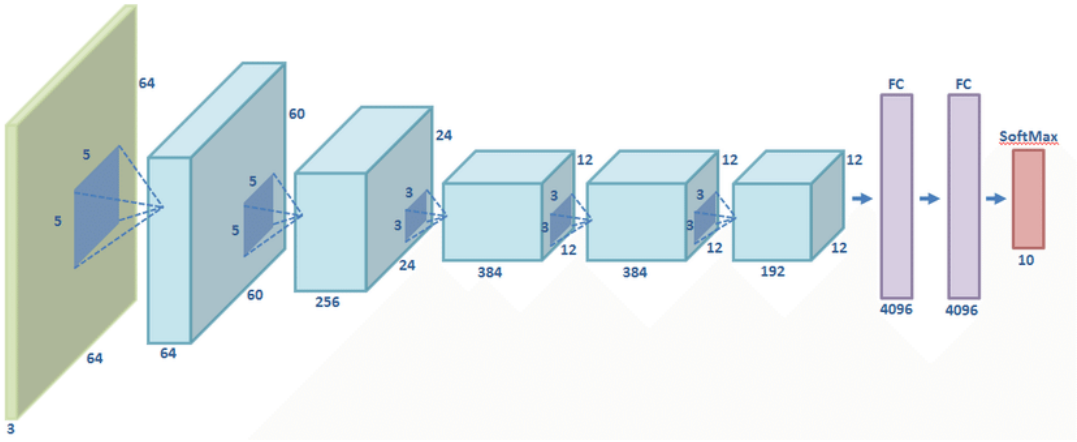


Fig. 4 Architecture du réseau AlexNet. Reproduit libre de droit de Llamas et coll.[54]

Il existe plusieurs réseaux convolutifs qui ont fait leurs preuves[55,56], typiquement en démontrant un bas taux d'erreur sur la compétition de classification d'images ImageNet[52], décrite

ci-dessus. Ces réseaux sont très larges et existent sous un format préentrainé pour faciliter leur utilisation. Vu qu'il existe autant d'architecture de réseau de neurones que l'imagination le permet, il est fréquent que les chercheurs se basent sur des architectures publiées et démontrées plutôt que de partir de zéro. De plus, bien que la majorité de ces réseaux aient été développés pour des applications en deux dimensions, la convolution s'applique aussi en unidimentionnel[55] (données temporelles, traitements de signal) et en tridimensionnel[57] (vidéos, imagerie médicale).

3.3.4 Autoencodeur

L'autoencodeur[58] est un sous-type de réseau qui mérite une attention particulière. C'est un réseau de neurones, qui peut être dense ou convolutif, dont l'intérêt principal est qu'il fonctionne par apprentissage non supervisé. Son fonctionnement est basé sur une architecture dans laquelle l'entrée et la sortie sont du même type, avec un espace latent au milieu qui joue le rôle d'un goulot d'étranglement. L'entrainement se fait en plaçant à l'entrée et à la sortie les mêmes données. La présence de l'espace latent va amener le réseau à devoir synthétiser ou résumer les données pour ensuite pouvoir reconstituer l'entrée avec la plus petite erreur moyenne possible. La Fig. 5 montre un exemple de cette architecture avec trois entrées et sorties et dont l'espace latent est une seule valeur.

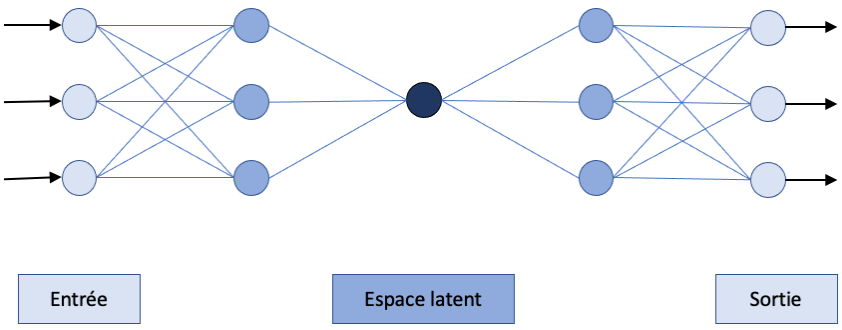


Fig. 5 Exemple simpliste d'autoencodeur

Bien que les règles de bases soient fixes (entrée=sortie et espace latent réduit), le reste de l'architecture est libre. Un autoencodeur peut donc être très complexe ou très simple, constitué de couches convolutives ou non.

Cette architecture en apparence simple permet un nombre d'applications impressionnant. Puisque le format des données de l'espace latent est modifiable, il est possible de transformer des données en un format différent. Une fois l'autoencodeur entrainé, il est possible de ne conserver que la première moitié et d'utiliser le réseau pour transformer un type ou un format de données dans un autre type ou format[58–60]. Par exemple, l'entrée peut être une image, des données temporelles ou des données de très grande dimension, alors que la sortie (espace latent) peut être de plus petite taille et unidimensionnel. Lorsqu'utilisé de cette manière, les valeures contenues dans l'espace latent sont parfois appelé caractéristiques ou *features*. Une autre utilisation est en tant que préentrainement, en faisant un premier entrainement sur des données de manière non supervisée avant de finir l'entrainement de manière supervisée sur une base de données de plus petite taille. Finalement, les autoencodeurs sont aussi utilisés comme détecteurs d'anomalies[41,61]. En entrainant le réseau sur une base de données ne contenant pas l'anomalie en question, il n'apprend pas son existence. Si par la suite on lui présente en entrée une anomalie, il risque reconstituer la sortie de manière erronée. Cette erreur de reconstitution élevée est un moyen de détecter l'anomalie.

3.4 Enjeux et difficultés spécifiques aux données médicales

D'un point de vue clinique, le but du modèle développé dans ce mémoire est de détecter de potentiels donneurs d'organes. D'un point de vue d'apprentissage machine, il s'agit d'une problématique qui est récurrente en médecine : la détection d'événements rares à l'aide de données cliniques incluant des séries temporelles. La détection d'événements rares, ou d'anomalies, n'est pas propre à la médecine; on peut par exemple penser la détection de fraudes. La recherche de solutions face à ce problème est un champ de recherche actif dans le domaine de l'apprentissage machine[61]. Au-delà de la problématique de détection d'événements rares, les données médicales présentent plusieurs enjeux et particularités qui sont décrits ci-dessous.

3.4.1 Types de données

Les données médicales sont une numérisation de multiples points de vue sur un individu. Il existe donc une immense variété de types de données[62] provenant de nombreuses sources[63]. Ces données peuvent être continues (résultats de laboratoire, données démographiques, etc.),

catégorielles (scores cliniques, type d'imagerie, médication, etc.), binaires (présence/absence diagnostique, certains résultats de laboratoire, etc.), mais aussi en format imagerie 2D (radiologie simple), imagerie 3D (CT-Scan, IRM, etc.), textuelle (notes médicales, rapports de radiologie, etc.). Les données peuvent être temporelles (signes vitaux, résultats de laboratoires, etc.) ou fixes sur un épisode de soin (diagnostic, âge, sexe, etc.). Les données temporelles ont des fréquences d'échantillonnage qui varient grandement (d'une fois par jour à plusieurs fois par minute, ou même par seconde). Finalement, en plus d'une variabilité de types de données, les données médicales apportent de nombreux défis au niveau de l'accès, de l'harmonisation, de la corruption et de la complexité des données. Il est souvent rapporté que, dans un projet utilisant une base de données, la préparation des données prend 80% du temps[64]; cette proportion est probablement augmentée en santé vu la complexité spécifique des données.

3.4.2 Features et transformation de variables

Le terme *features*, qui se traduit par caractéristiques représentatives, fait référence aux valeurs qui sont utilisées dans un modèle d'apprentissage machine. Il peut s'agit des données cliniques brutes, ou des valeurs transformées à partir de ces données. Le choix ou la création de nouvelles caractéristiques représentatives est une technique centrale au traitement des données et à l'apprentissage machine[65]. Le choix de *features* moins nombreux, mais plus informatifs permet de réduire la taille d'un modèle et parfois d'en améliorer les performances. Ce choix peut se faire selon un avis d'expert du domaine ou via un processus itératif plus ou moins structuré. La création de nouvelles caractéristiques représentatives implique la transformation de données brutes. La transformation peut être simple (normalisation ou catégorisation) ou peut impliquer la création de nouvelles données synthétiques (imputation et données manquantes, voir ci-dessous). Finalement ces caractéristiques peuvent être générées par l'humain ou déduites automatiquement par l'algorithme d'apprentissage machine. Une grande utilité de la création de nouvelles caractéristiques représentatives est le changement d'un type de données en un autre.

En effet, la majorité des algorithmes d'apprentissage machine ne peuvent utiliser que des données d'un type défini. Il peut s'agir de données numériques tabulaires (CNN, LSTM) ou de données numériques continues ou binaires. Les données textuelles ou catégorielles nécessitent

une transformation avant de pouvoir être utilisées. Une approche classique de transformation d'une donnée catégorielle est le « *one hot encoding »* où une variable contient, par exemple, trois catégories est transformée en trois différentes variables binaires indépendantes (une par catégorie). Une autre approche est le « *target encoding »* dans laquelle une variable catégorielle est transformée en donnée continue dont la valeur dépend de la proportion de cette catégorie dans le groupe avec l'issue recherchée. Ce type d'encodage nécessite toutefois une forme d'apprentissage, mais évite le problème d'augmentation du nombre de variables.

La transformation, explicite ou implicite, est aussi la seule manière de regrouper différents types de données. La transformation de données complexes (image, temporelle, textuelle, etc.) en un format unique et plus simple (vecteur unidimensionnel numérique) permet le regroupement des différentes données. Cela permet de regrouper l'information provenant de sources très diverses dans un même algorithme d'apprentissage machine, pour arriver à une prédiction qui se base sur une multitude de sources d'information.

Ces enjeux de transformation de données et de choix de caractéristiques représentatives sont propres à l'ensemble des projets utilisant des données. Ils sont toutefois prépondérants dans les projets utilisant des données médicales du fait de la grande variété de types de données et la fréquence importante des données catégorielles et textuelles.

3.4.3 Données manquantes et erronées

Le traitement des données manquantes est un sujet complexe[66]. Les données manquantes peuvent être classifiées en manquantes complètement au hasard (manquement totalement aléatoire), manquantes au hasard (pas totalement aléatoire, mais indépendant de la valeur étudiée et le manquement peut être prédit par les données restantes) ou manquantes non au hasard (probabilité d'absence reliée à la valeur étudiée et le manquement ne peut pas être prédit par les données restantes)[62].

La grande majorité des algorithmes d'apprentissage machine ne peuvent utiliser une donnée manquante. La solution la plus simple est donc de retirer les variables ou les cas comportant des données manquantes pour ne garder que les cas complets. Toutefois, cela réduit grandement les données utilisables et la capacité de généralisation du modèle. L'approche

préférée est celle de l'imputation, où la variable manquante est remplacée par une variable artificiellement générée. La méthode de remplacement est un sujet d'étude en soi et peut être basée sur une technique statistique (imputation par la moyenne, imputation multiple, etc.), sur des approches d'apprentissage machine ou sur des approches mixtes[67]. Il est important de noter que la majorité de ces approches assument que les données sont manquantes complètement au hasard, ce qui n'est souvent pas le cas.

La problématique des données manquantes est fréquente en santé[62]. De plus, indépendamment de la cause, la probabilité qu'une donnée soit manquante est souvent corrélée à la probabilité qu'une autre donnée soit manquante. Une situation commune est celle où le clinicien choisit de ne pas faire un test de laboratoire ou d'imagerie particulier, parce que non cliniquement requis. Les données peuvent aussi ne pas avoir été mesurées par manque de temps clinique. Les données démographiques peuvent aussi manquer parce que le patient est incapable de les donner (démence, inconscience, etc.). Les données temporelles peuvent avoir été échantillonnées à des fréquences différentes entre les patients, créant une forme de données manquantes. Une donnée jugée erronée ou aberrante peut devoir être retirée et devenir ainsi manquante. Finalement, il peut aussi y avoir une différence entre les pratiques cliniques de deux centres hospitaliers, de deux services ou de deux cliniciens, tous menant aussi à une forme de données manquantes. Les exemples sont nombreux, et la compréhension du mécanisme est importante pour choisir la méthode d'imputation la plus appropriée.

3.4.4 Quantité de données

Un dernier enjeu spécifique aux données médicales est la relative petite quantité de données disponibles. Cela est paradoxal, puisqu'une seule hospitalisation peut générer environ 150 000 points de données[68]. Cependant, les données sont souvent inaccessibles, réparties sur de multiples bases de données séparées, non numérisées ou protégées par le secret médical. En comparaison des millions d'images d'ImageNet[52], les bases de données en santé se comptent en milliers de patients[69,70]. De plus, il est requis de séparer les données en groupes d'apprentissage, validation et test, réduisant davantage la quantité de données[65]. Le mauvais couplage entre capacité du modèle et quantité de données augmente le risque de surapprentissage (*overfitting)*.

Il existe, là aussi, de nombreuses approches face à ce problème spécifique. La plus simple est de réduire la complexité des algorithmes en choisissant un algorithme intrinsèquement plus simple (par exemple un modèle linéaire) ou en réduisant la quantité de couches et de paramètres d'un modèle neuronal. Une autre approche est d'utiliser des méthodes de régularisation, qui réduisent le risque de surapprentissage (*overfitting)* et maximisent la généralisation. Les plus utilisées sont l'arrêt précoce (arrêt d'entrainement avant d'atteindre trop de convergence), le *dropout* (blocage aléatoire d'une partie des connexions du réseau neuronal) ou l'utilisation de norme L1 ou L2 (pénalisation sur la complexité du modèle)[38].

La solution idéale est l'augmentation de la quantité de données. Toutefois, cette option est souvent difficile en santé, les projets étant limités par la prévalence de l'issue à l'étude et la taille de la base de données utilisée. Les options envisageables sont de continuer la collecte de données prospectivement ou d'obtenir et d'harmoniser les données provenant d'autres sources, deux options coûteuses en temps et en énergie.

Sans collecter de nouvelles données, il est possible de créer des données partiellement, ou complètement, artificielles. C'est une approche bien implantée en imagerie[71] avec des résultats impressionnants, même en imagerie médicale[72]. Son utilisation en données tabulaires est plus rare, mais reste démontrée[73]. Une seconde option consiste à entrainer le modèle d'apprentissage machine sur une base de données externe de plus grande taille et à transférer ces connaissances dans la base de données interne de plus petite taille. Le premier entrainement peut être fait de manière supervisée ou non, augmentant de beaucoup la quantité de données utilisables. Cette approche permet d'améliorer les résultats en les comparant à une approche sans transfert de connaissance[74–76].

Finalement, un cas particulier qui s'apparente au manque de données est celui de la balance de classes. Le domaine de la santé comporte de très nombreux problèmes où le but est de discerner une situation rare dans une population plus large. Par exemple, seule une faible proportion de patients ayant une radiographie des poumons aura un diagnostic de cancer du poumon, mais il est crucial de ne pas manquer ce diagnostic. Dans ce genre de problème, la prévalence de la classe minoritaire d'intérêt est souvent de 1 :100 à 1 :1 000, parfois plus rare encore. Un algorithme pourrait classifier tous les cas comme négatifs et aurait un score de

prédiction de 99 à 99.9%, tout en étant complètement inutile cliniquement. Pour aborder ce problème, des modifications aux données ou à l'algorithme peuvent être faites[77]. Au niveau des données, on peut réduire la quantité de données de la classe majoritaire (*under-sampling)* ou à l'inverse augmenter la classe minoritaire (*over-sampling)*. Il est aussi possible de générer un surplus de données synthétiques dans la classe minoritaire. Un exemple commun de cette approche est l'algorithme SMOTE (pour *Synthetic Minority Over-sampling Technique)*[78] qui fait une interpolation des données de la classe minoritaire pour générer plus de données. Au niveau des algorithmes, il est possible de pénaliser les erreurs de manière différentielle pour forcer un apprentissage plus important de la classe minoritaire. Aucune approche ne semble supérieure à une autre pour toutes les situations, et deux approches (donnée et algorithmes) peuvent être utilisées conjointement[77].

3.5 Revue de modèles prédictifs d'apprentissage machine en santé

Il existe un nombre exponentiel d'articles appliquant un algorithme d'apprentissage machine pour une prédiction clinique[42,63,79]. Ainsi, bien qu'il ne soit pas réaliste de réviser l'ensemble des articles publiés dans ce domaine, il semble pertinent d'étudier la méthodologie d'une sélection d'articles de grande qualité ayant des similarités avec la problématique de ce mémoire. Aucun ne porte sur le don d'organes, puisqu'il n'existe aucun modèle publié pour cette problématique. Toutefois, ces articles s'apparentent à la problématique de prédiction d'événements rares cliniques à l'aide de données réelles multimodales et incluant des séries temporelles. Il s'agit d'articles publiés dans des revues à haut facteur d'impact. Ce survol permettra de comparer comment ces équipes d'experts ont abordé les problématiques spécifiques au domaine de la santé décrites dans la section précédente.

Le premier article a été publié dans le *Lancet Respiratory Medicine Journal* en 2018 par Alexander Meyer et coll.[80]. Ils ont utilisé une population rétrospective de 47 559 patients ayant eu une chirurgie cardiaque ouverte entre 2000 et 2016. Le but était une prédiction binaire de trois issues cliniques (insuffisance rénale dialysée, retour en salle d'opération pour saignement et décès) à l'aide de 52 données cliniques, collectées durant les premières 24h postopératoires. Ils ont utilisé à la fois des données statiques (4 caractéristiques démographiques, 9 caractéristiques

de la chirurgie) que temporelles (17 laboratoires, 9 gaz sanguins, 11 signes vitaux et 2 bilans liquidiens), échantillonnées aux demi-heures. Les données manquantes ont été réduites en n'utilisant que des *features* disponibles chez au moins 50% de la population. Par la suite, un report par l'avant (*last value carried forward)* a été utilisé. Les données qui demeuraient manquantes à cette étape ont été imputées par une valeur spécifique à chaque *feature,* choisie par un expert clinique. La balance de classe (incidences de 4.9% pour le saignement, 6.2% pour la mortalité et 1% pour la dialyse) a été adressée par un sous-échantillonnage aléatoire de la classe majoritaire pour ramener l'incidence à 50%. Ils ont utilisé un réseau de neurones récurrent qu'ils ont comparé à des modèles de prédictions linéaires validés cliniquement et spécifiques aux issues choisies. Ils ont fait une validation externe en utilisant la base de données MIMIC[69]. Le RNA était supérieur pour le saignement (AUROC 0.87 vs 0.58), pour la mortalité (AUROC 0.95 vs 0.71) et pour l'insuffisance rénale (AUROC 0.96 vs 0.72). La supériorité restait vraie à chaque heure durant les premières 24h. Il y avait une perte de précision attendue dans la cohorte de validation externe (AUROC de 0.75 vs 0.66 pour le saignement, 0.81 vs 0.63 pour la mortalité et 0.91 vs 0.66 pour l'insuffisance rénale).

Le second article a été publié dans la revue *Anesthesiology* en 2018 par Christine Lee et coll.[81]. Ils ont utilisé une population rétrospective de 59 985 patients ayant eu une chirurgie sous anesthésie générale entre 2013 et 2016 dans deux centres hospitaliers californiens. Le but était une prédiction binaire de la mortalité intrahospitalière à l'aide de 87 variables intraopératoires choisies par deux experts du domaine. Ils ont utilisé uniquement des valeurs intraopératoires (principalement des doses de médicaments et des signes vitaux). Ils n'ont pas utilisé de valeurs temporelles, mais ont inclus dans le modèle plusieurs variables descriptives pour chaque valeur temporelle (c.-à-d. maximum, minimum, moyenne et dose cumulée). Les données manquantes ont été imputées par la moyenne. La balance de classe (incidences 0.8%) a été adressée par une génération de valeur synthétique en ré-échantillonnant des patients de la classe minoritaire et en ajoutant un bruit gaussien. Ils ont utilisé un réseau de neurones dense qu'ils ont comparé à des modèles de prédiction linéaire validés cliniquement et à des modèles logistiques développés localement. Ils ont fait plusieurs analyses de sensibilités avec des variantes de variables. L'architecture du réseau, de même que les nombreux hyperparamètres, ont été optimisés par

une approche non décrite. Ils n'ont pas fait de validation externe. Le RNA obtenait un AUROC de 0.90, comparable à une régression logistique laquelle obtenait une AUROC de 0.89). Bien que le résultat global soit très encourageant, il était équivalent au score clinique anesthésique ASA[82] et bien inférieur à d'autres modèles de prédictions validés cliniquement (AUROC 0.97 pour le *risk stratification index).*

Le dernier est un article publié dans la revue *Nature Medicine* en 2020 par Stéphanie Hyland et coll.[83]. Ils ont utilisé une population rétrospective de 55 602 patients ayant eu une hospitalisation aux soins intensifs d'un hôpital suisse entre 2008 et 2016. Le but était de développer et valider un score de prédiction de collapsus hémodynamique afin de prédire un tel événement dans les 8h suivantes. Leur choix de variables cliniques suivait un processus complexe. Ils utilisaient un grand nombre (non quantifié) de variables à la fois en laboratoire, clinique et démographique. À partir de ces variables, ils ont créé 5 278 *features*. De ceux-ci, seules 500 (tirées de 112 variables) ont été gardées pour le modèle complexe et 176 (16 variables) pour le modèle léger. La sélection de *features* a utilisé une approche semi-automatique en les classant selon l'information apportée au modèle. L'aspect temporel des variables a été utilisé par une approche de *shapelets*, isolant automatiquement des profils de progression d'intérêt (augmentation, diminution, stagnation, forme de U, etc.) et utilisant la ressemblance à ces formes comme une variable numérique. Les données manquantes ont aussi été imputées à l'aide d'un report vers l'avant (*last value carried forward),* mais en ajoutant un effet de dérive vers une valeur médiane. Par la suite, une valeur normale est imputée aux données manquantes. La balance de classe a été adressée avec un sous-échantillonnage, mais ce n'était pas clairement décrit. Plusieurs approches ont été testées (ensemble d'arbre de décisions avec gradient *boosté,* régression logistique, réseau de neurones récurrent, perceptron) et l'approche par ensemble d'arbres de décision était supérieure avec un AUROC de 0.94. Une validation externe faite sur la base de données MIMIC conservait une excellente performance avec un AUROC de 0.90. Ils ont fait de nombreuses sous-analyses pour démontrer la résilience du modèle et sa capacité de prévenir de manière précoce l'événement prédit jusqu'à plusieurs heures avant l'événement.

Il n'existe aucun modèle ou publication dont le but est de prédire les potentiels donneurs d'organes. La grande majorité des études publiées en don d'organes se concentrait sur la survie

en attente d'un organe[84] ou après la transplantation[84–86]; sur la prédiction du rejet[87] ou le choix des médicaments[88]. Rabinstein and coll.[89] ont développé un score aidant à prédire le succès pour les patients DDC. Bien que complémentaire, le but n'est pas le même que ce mémoire. Le seul article qui s'approche du but de cette maitrise est publié par Fernandes and coll.[9] en 2015. Ils ont développé un algorithme de dépistage automatique des catastrophes neurologiques. Leur approche démontrait une sensibilité de 77% et une spécificité de 66%. Toutefois, leur approche était basée sur des mots clefs trouvés dans les rapports de radiologie, une approche très novatrice, mais qui reste limitée par le besoin d'obtenir une interprétation préalable des résultats d'imagerie.

3.6 Application clinique et systèmes d'aide à la décision clinique

Peu de modèles pronostics se rendent à une validation clinique, qu'ils soient basés sur une méthode d'apprentissage machine [39] ou non [90,91]. Cette réalité dépasse les modèles prédictifs et touche l'ensemble des modèles d'apprentissage machine clinique[92]. Toutefois, l'utilisation d'outils cliniques informatisés, basés sur des modèles d'apprentissage machine ou non, est prometteuse. Ces outils sont communément appelés systèmes d'aide à la décision clinique (SADC). Leur implémentation et leur utilisation courante sont limitées par la qualité des modèles sous-jacents, mais aussi par des enjeux logistiques, informatiques, financiers ou culturels [93–95]. Toutefois, ils ont un grand potentiel pour la sécurité des patients et l'amélioration de la qualité des soins.

Leur utilisation, dans le contexte particulier de l'aide au diagnostic, ne démontrait pas de signal fort de gain clinique dans une revue systémique récente[96]; une majorité des études utilisées souffraient de biais importants. La moitié des études avaient tout de même démontré une amélioration des résultats étudiés.

Aux soins intensifs, de nombreux SADC ont été testés, particulièrement dans le domaine du diagnostic précoce et de la pronostication[97]. Leur utilisation dans l'aide au diagnostic n'a pas démontré d'amélioration clinique bien que les études étaient, là aussi, rares et de faibles qualités[94]. L'aide au diagnostic du sepsis sévère, par exemple, ne changeait ni la mortalité ni la rapidité de diagnostic[98]. Toutefois, leur utilité semble avoir un intérêt particulier dans l'aide au

diagnostic de maladies rares[99]. C'est une situation difficile et particulièrement susceptible à plusieurs biais cognitifs liés à la nature humaine[100]. Une étude, utilisant une alerte automatique basée sur des critères cliniques, a permis de tripler le nombre de tests de diagnostics pour une maladie rare (déficit en alpha-1 antitrypsine)[101].

L'utilisation de SADC en don d'organes est très émergente. Leur utilisation est surtout limitée à l'attribution des organes[102]. Une étude est en cours pour aider à mieux prédire le temps d'ischémie chaude dans un processus de DDC (NCT04661787).

Une étude unicentrique faite par une équipe d'intensivistes-pédiatres américains s'apparente fortement au sujet de ce mémoire[8]. Ils ont implémenté un système qui réfère automatiquement tous les patients ayant des critères de potentiels DDN à leur ODO. Ils ont démontré une diminution du délai avant référence de 30 heures à 1.7 heure tout en améliorant le taux de conversion de 50% à 90%. À noter qu'il n'y avait pas de potentiels donneurs manqués, ni avant ni après l'introduction du système. Ces résultats sont fortement encourageants pour le succès d'un tel système à plus grande échelle. Les résultats sont limités par le délai de 2 ans entre la période pré et post SADC, pouvant biaiser les résultats. De plus, leur système utilisait un dossier informatisé très complet, lequel comprenait des éléments très précis de l'examen neurologique (taille des pupilles, état de conscience, horodatage des médicaments, etc.). Ce genre d'information n'est pas disponible dans la majorité des hôpitaux. Par ailleurs, leur système n'utilisait pas un modèle prédictif, mais suivait une série de règles prédéfinies testant directement les critères de potentiel donneur d'organes tels que décrits à la section 2.42.4. Cette étude soutient le potentiel d'un système d'aide au diagnostic de potentiels donneurs d'organes. Toutefois, puisque la majorité des dossiers électroniques ne contiennent pas les informations nécessaires à un encodage des critères, il est nécessaire de développer un modèle prédictif qui utilise les informations usuellement disponibles pour permettre d'identifier avec précision les potentiels donneurs d'organes.

4 Méthodologie

4.1 Population à l'étude, critères d'inclusion et d'exclusion et sous populations

La cohorte utilisée dans ce projet est tirée des données cliniques et administratives de CITADEL[103], une base de données cliniques informatisées, rassemblant plus de 20 millions d'épisodes de soins au CHUM chez plus de 3.5 millions de dossiers patients.

La population à l'étude était l'ensemble des patients ayant été admis dans une des différentes unités de soins intensifs adultes (incluant l'unité de soins intensifs coronariens) entre le 1er janvier 2012 et le 31 décembre 2019. Les patients ayant un épisode hospitalier de moins de 16h ont été exclus. Seul le dernier épisode de soins par patient a été conservé. La population a été séparée en deux. La première, composée de patients des soins intensifs, a servi a développé et tester le modèle (voir chapitres 5 et 6). La seconde, composée des patients de l'unité des soins intensifs coronariens, a servi de cohorte de validation externe (voir chapitre 7) et a été complétement exclue du développement du modèle. L'étude et l'extraction des données ont été approuvées par le comité d'éthique de la recherche du centre de recherche du CHUM (numéro de projet 19.158) sans consentement direct des participants, mais avec approbation du Directeur des Services Professionnels du CHUM, vu le bas risque du projet et l'impossibilité d'obtenir le consentement de l'ensemble des patients, en accord avec les articles 3.7A et 5.5A de l'énoncé de politique des trois conseils[104].

L'ensemble des données utilisées provenait d'une agrégation de bases de données cliniques, assemblées et structurées par les scientifiques des données de CITADEL[103]. Les données démographiques, de service clinique d'appartenance, de même que les dates et heures, sont tirées de la base de données administratives de l'hôpital, dont l'exactitude et la complétude sont nécessaires au bon fonctionnement quotidien du CHUM. Bien que le CHUM était originalement basé sur trois sites, l'ensemble de l'administratif, des laboratoires et du service d'imagerie étaient centralisés et son fonctionnement était celui d'un unique hôpital bien avant la fusion physique.

Des variables d'intérêt ont été extraites à la fois pour le développement du modèle et pour la description de la population. Ces variables étaient, au niveau clinique, les laboratoires, la présence ou absence d'une imagerie cérébrale et les diagnostics reliés à l'épisode (code ICD10). Les codes ICD10 sont ajoutés à postériori par le personnel des archives médicales. Seuls les diagnostics principaux ont été utilisés dans la description de la population, pour maximiser la fiabilité de cette donnée. Les résultats de laboratoires ont été choisis comme source d'information principale pour leur impartialité, le fait qu'ils soient standardisés, qu'ils minimisent l'impact humain dans leur traitement tout en reflétant directement l'état clinique du patient et le fait qu'ils soient numérisés dans la totalité des hôpitaux.

Au niveau administratif, ces variables étaient l'âge, le sexe, la durée de l'épisode hospitalier, le statut (mort/vivant) au congé et le service clinique d'appartenance du patient. Les laboratoires, la présence d'imagerie cérébrale et le service d'appartenance du patient étaient les seules variables utilisées dans la création du modèle. Le reste des variables (âge, sexe, durée, statut mort/vivant et ICD10) n'ont été utilisées que pour la description de la population.

Parmi cette population de patients de soins intensifs du CHUM, les potentiels donneurs d'organes ont été identifiés en combinant plusieurs sources d'informations. Les données du bloc opératoire ont permis d'identifier tout patient ayant subi une procédure de prélèvement d'organes. Les données de TQ ont permis d'identifier les patients référés, qu'ils aient été refusés ou non. De plus, un audit continu des patients décédés au CHUM, effectués par des infirmières de TQ, a permis d'identifier les potentiels donneurs d'organes non référés et manqués.

L'audit de décès de TQ, dont les données sont utilisées dans la cohorte, utilisait des critères qui étaient légèrement différents des recommandations du groupe d'experts décrites à la section 2.4 ci-dessus vu qu'ils dataient d'avant ces recommandations. L'audit de décès de TQ définissait deux sous-populations de donneurs d'organes potentiels. La première incluait les patients avec une atteinte neurologique sévère nécessitant une ventilation mécanique et qui sont décédés dans les 24h suivant le passage en soins de confort. La seconde incluait les patients nécessitant une ventilation mécanique, sans atteinte neurologique sévère, mais qui sont décédés dans les 3h suivants le passage en soins de confort. Les patients avec de franches contrindications au don d'organes (cancer récent, actif ou métastatique, infection disséminée ou dysfonction

multiorganique) étaient exclus. L'ensemble de ces donneurs potentiels pouvait être divisé en quatre sous populations cliniquement différentes:

1. Les donneurs transférés (*transfered donors)*. Il s'agissait des patients transférés au CHUM en état de mort neurologique, qui ont été évalués par TQ au CHUM et dont les organes ont été prélevés au CHUM.
2. Les donneurs référés et transplantés (*refered and transplanted*). Il s'agissait des patients hospitalisés au CHUM, qui ont été référés à TQ, confirmés comme donneur et dont les organes ont été prélevés.
3. Les donneurs référés et non transplantés (*refered but ineligible*). Il s'agissait des patients hospitalisés au CHUM, qui ont été référés à TQ, mais qui n'étaient pas éligibles au don d'organe, quelle que soit la raison (néoplasie, refus familial, etc.).
4. Les donneurs potentiels non référés (*not refered*). Il s'agissait des patients hospitalisés et décédés au CHUM, remplissant les critères de l'audit de décès, mais qui n'ont pas été référés à TQ.

4.2 Préparation de la base de données

4.2.1 Données temporelles de laboratoires

4.2.1.1 Préparation initiale

La complexité et l'intérêt des données de ce projet étaient leur aspect multimodal et temporel. La base de données était séparée en deux fichiers différents et non fusionnables. Le premier contenait les données statiques, démographiques et administratives (dates, âge, identifiants, etc.). Le second contenait les données temporelles des laboratoires. Les deux fichiers devaient être nettoyés et formatés en parallèle.

Il y avait une grande variété de types de laboratoires différents allant du très rare au quasi omniprésent. Les types des laboratoires suivaient un format local et ont dû être transformés et uniformisés selon la nomenclature LOINC[105]. Par ailleurs, plusieurs laboratoires étaient fragmentés (gaz artériels, appellation différente selon le site hospitalier d'origine, etc.) et devaient être fusionnés. Ce travail d'harmonisation a été fait manuellement, de manière itérative,

avec un œil clinique expert, afin de minimiser la perte de données. Le contenu de la colonne de résultat étant multimodal, il était en chaine de caractère et non pas en valeur numérique. La majorité des résultats étaient des données continues. Une petite partie était binaire. Une autre partie était catégorielle ordinales (1+, 2+, 3+ ou 1-2, 3-5, 6-10) ou non ordinales (présence/absence/inconnu). Finalement, certaines valeurs étaient entrées manuellement et ont dû être uniformisées (pourcentage et fraction). Le but final était d'obtenir un jeu de données propre et formaté en 4 colonnes (identifiant patient, horodatage de la donnée de laboratoire, identifiant du laboratoire et valeur de laboratoire numérique) avec une perte de données minimale.

4.2.1.2 Approche des données manquantes, aberrantes et formatage

Le but de cette partie était de transformer le jeu de données propre complet en un jeu de données formaté utilisable par un algorithme d'apprentissage machine.

La première étape était de réduire la diversité de type de laboratoire. Le but était de ne garder que des laboratoires courants pour un patient standard de soins intensifs. La rationnelle était d'éviter de fournir à l'algorithme des laboratoires spécifiques aux potentiels donneurs d'organes. De plus, utiliser des laboratoires courants permettait d'augmenter la transférabilité du modèle vers des hôpitaux de plus petite taille. Les analyses de laboratoires rares étaient celles qui n'avaient été faites que chez un petit pourcentage des patients. Un pourcentage limite de 10% a été arbitrairement décidé avec une validation de cette valeur par une étude de sensibilité. Les laboratoires qui étaient présents chez moins de 10% des patients ont donc été exclus du modèle parce que trop rares. Dans la même logique, certains patients n'ont eu que très peu d'analyses de laboratoires. Ainsi, les patients ayant eu moins de 10% des analyses de laboratoires utilisées ont été retirés de la base de données, pour éviter de devoir imputer la quasi-totalité de leurs données.

Par la suite, les laboratoires n'étant pas faits à des fréquences ni des heures régulières, il était nécessaire de les formater de manière standardisée. Le séjour hospitalier du patient a été fragmenté en blocs de 8h. Seule la dernière valeur de laboratoire était gardée dans chaque bloc.

Outre la standardisation, cela avait pour but d'aveugler l'algorithme à des laboratoires faits à plus haute fréquence, comme c'est le cas pour les donneurs d'organes.

Puisque notre objectif était d'utiliser l'évolution dans le temps des laboratoires, nous nous attendions à ce que l'amélioration des laboratoires soit un élément important du modèle. Les donneurs potentiels vivent typiquement un événement critique (arrêt cardiaque, AVC, trauma, etc.) puis une période de stabilité, suivi d'une amélioration. Durant cette période, les organes récupèrent, mais le cerveau a subi des dommages trop importants pour avoir la même amélioration. Nous avons émis l'hypothèse que le patient montrerait donc une amélioration rapide d'au moins une partie des résultats de laboratoires. Pour capturer cette progression, une fenêtre de trois jours (72h) a été choisie comme compromis entre minimiser la quantité de données imputées et maximiser la durée d'évolution. De plus, pour maximiser le signal, les trois derniers jours du patient avant son congé des soins intensifs ont été utilisés.

Une fois les données formatées, certains blocs de 8h restaient vides et devaient être imputés. Comme discuté ci-dessus, il existe de nombreuses méthodes d'imputations. Puisqu'il y avait un lien clair entre l'issue étudiée et les données de laboratoires manquantes, il ne s'agissait pas d'un cas de données manquantes au hasard. Nous avons utilisé une technique qui voulait imiter la réflexion clinique et palier au mécanisme causant le manque. La première étape était de recopier chaque valeur vers l'avant, jusqu'au moment où une nouvelle valeur est mesurée (*last value carried forward*). Ainsi, si un patient avait une donnée dans un bloc de 8h, cette valeur était recopiée dans l'ensemble des blocs subséquents jusqu'à ce qu'elle soit remplacée par une nouvelle valeur plus à jour. Cette stratégie est parfois critiquée [106,107] lors de l'imputation de données manquantes. Cependant, son utilisation était nécessaire pour que le modèle soit fonctionnel dans un contexte clinique réel prospectif, le modèle ne pouvant être conscient d'informations qui surviendraient dans le futur. De plus, l'horizon de temps étant relativement court (72h), il semblait raisonnable de penser que la valeur qui approxime le mieux la donnée manquante est la dernière valeur disponible pour le patient.

Une fois cette première étape complétée, il restait des données complètement manquantes, soit parce qu'elles précédaient la première donnée disponible pour l'analyse de laboratoire en question, soit parce que cette analyse de laboratoire n'avait jamais été faite chez le patient. Nous

avons imputé ces variables par une valeur biologiquement normale. Une imputation basée sur les données existantes aurait eu tendance à biaiser les résultats vers des valeurs anormales, surtout pour des laboratoires peu fréquents. En effet, si une analyse de laboratoire est prescrite par un clinicien, la probabilité pré-test que ce laboratoire soit anormal est plus grande. À l'inverse, si un clinicien n'a aucun indice clinique de vérifier un laboratoire, sa valeur a une plus haute probabilité pré-test d'être normale. Nous avons donc imputé une variable choisie de manière aléatoire, selon une distribution gaussienne et située entre le minimum et le maximum normal fournis par le laboratoire.

4.2.2 Données statiques et démographies

La plupart des données cliniques sont temporelles. Il existe cependant plusieurs données qui peuvent être considérées comme statiques sur une même hospitalisation (âge, sexe, antécédents médicaux, prise de médication, résultats d'imageries, raison d'admission, etc.). Certaines de ces données n'étaient pas accessibles aisément dans notre jeu de données. Les antécédents médicaux et pharmacologiques sont riches d'informations, mais n'étaient pas numérisés de manière standardisée ou n'étaient disponibles seulement une fois l'hospitalisation complétée. L'accès aux données du dossier santé Québec (DSQ) n'est pas autorisé pour la recherche. Dans les données numérisées, la majeure partie (âge, sexe, poids et taille) n'est pas une contrindication à la référence comme donneur potentiel. Nous avons donc choisi de ne pas fournir la plupart de ces données au modèle, même si disponibles, pour éviter d'induire un biais de sélection latent. Les deux seules données statiques qui étaient incluses à ce stade étaient la présence ou l'absence d'une imagerie cérébrale et le service d'admission du patient.

La première (la présence ou l'absence d'une imagerie cérébrale) était une donnée binaire. Elle a servi de proxy simplifié à l'ajout futur des données d'imagerie brute. Il s'agissait d'une valeur facile à obtenir, universellement numérisée et disponible dans la grande majorité des centres hospitaliers. Cette valeur était obtenue par un filtrage par expression régulière (regex) sur une extraction textuelle de la liste des examens d'imageries faits par le patient. Par la suite, la valeur binaire « a eu / n'a pas eu une imagerie cérébrale avant le départ des soins intensifs » pouvait être ajoutée aux données servant à entrainer le modèle.

La seconde valeur statique, service d'admission du patient, était une donnée catégorielle. Lors d'un épisode hospitalier, et particulièrement lors d'une admission aux soins intensifs, le patient est sous la responsabilité d'un service clinique (chirurgie générale, neurochirurgie, cardiologie, etc.). Cette information servait de proxy très général à la raison d'admission. Il existe de nombreuse façon d'encoder des variables catégorielles[108]. Le plus fréquemment utilisé (*one-hot-encoding)* crée une nouvelle colonne binaire pour chaque catégorie. Son utilisation dans notre cas nous semblait inutilement complexe du fait de la probable grande quantité de catégories de services cliniques. C'était une donnée dite à grande cardinalité. Par conséquent, nous avons choisi de transformer les données avec un encodage basé sur la cible (*target encoding*). Cette approche semble la plus performante, indépendamment de l'algorithme choisi[108]. Le but est de donner une valeur numérique à chaque catégorie. Cette valeur augmente selon la fréquence de représentation de la catégorie dans le groupe d'intérêt (ici le groupe de potentiels donneurs d'organes). Dans notre situation, nous nous attendions, par exemple, à ce que la valeur encodée du service d'admission « neurochirurgie, » soit plus élevée que la valeur encodée du service de « chirurgie cardiaque ». Le premier était probablement surreprésenté dans les potentiels donneurs d'organes avec dommages neurologiques sévères alors que le dernier était plus rare dans la sous-population de potentiels donneurs d'organes, malgré qu'il soit fréquent dans l'ensemble de la population.

4.3 Premier modèle : Autoencodeur non supervisé

La problématique au centre de ce mémoire est un cas de détection d'événements cliniques rares. [61] .Bien que plusieurs approches soient possibles pour ce type de problème, l'utilisation d'autoencodeur[38,51,58,109] semblait avantageux par sa capacité d'encoder et de transformer les données multimodales (notamment les données temporelles) en un espace dimensionnel plus réduit, pour ensuite les reconstituer à l'état original. L'hypothèse était que le réseau de neurones reconstituerait avec un plus grand succès les données issues de la distribution dominante et, à l'inverse, ferait plus d'erreurs de reconstitution sur les anomalies (les potentiels donneurs d'organes dans notre cas). Cette erreur de reconstitution, quantifiée en erreur moyenne mise au carré, était la valeur permettant de détecter les anomalies lorsqu'elle dépassait un certain seuil d'erreur. Cette approche a fait ses preuves en traitement de signal[109] et en santé[110], mais n'a

jamais été utilisée pour des données tabulaires temporelles. Cette approche présentait plusieurs avantages potentiels pour notre projet et pour un grand nombre de situations du domaine médical. L'entrainement de l'autoencodeur pouvait se faire de manière non supervisée sur un jeu de donnée bruité, tant que la classe d'anomalie est très largement minoritaire. De plus, en sélectionnant une grande quantité de patients non-donneurs pour son entrainement, on se retrouvait à réduire la problématique de balance de cas dans les groupes de validation et de test par *under-sampling*. Finalement, l'autoencodeur ouvrait la possibilité d'être entrainé avec un jeu de données externes[69,70], et le réseau entrainé pourrait servir de préentrainement pour des projets futurs.

Tel que décrit dans les articles d'exemple (section 3.5), il existe de nombreuses approches possibles pour les séries temporelles. Le choix s'est rapidement porté sur les RNA pour leur capacité d'utiliser des données multimodales, leur utilisation intrinsèque de l'aspect temporel, pour leur plasticité et pour la grande quantité de publications suggérant leur fiabilité dans une variété de situations. Deux types de RNA permettaient d'utiliser les données temporelles : les réseaux convolutifs[55,56] et les réseaux récurrents[38,111]. Il n'existait pas à notre connaissance de supériorité d'une approche sur l'autre, dans le domaine des données temporelles. Toutefois, l'approche par réseaux récurrents était plus utilisée dans le domaine de l'analyse de signal et le domaine du langage. Un de leurs avantages est notamment la possibilité d'utiliser des séries temporelles de durées variables, ce qui n'était pas un besoin dans ce projet. L'approche par réseaux convolutifs se démarquait par son utilisation extensive dans le domaine de l'imagerie. Il existe de très nombreuses architectures publiées[55] basées sur des réseaux convolutifs. Par ailleurs, l'ensemble des données de laboratoires ont été traitées d'un bloc, similaire à une image. Devant l'absence d'une claire supériorité des réseaux récurrents, la plus grande simplicité et la plasticité des réseaux convolutifs et la possibilité d'utiliser des architectures bien démontrées, l'approche s'est rapidement portée vers un réseau convolutif. Toutefois, il serait intéressant de comparer plusieurs approches pour un même problème clinique afin d'aider de futurs cliniciens dans ce choix.

L'approche choisie a donc été de sélectionner une architecture convolutive reconnue dans le domaine de l'imagerie[112] et de l'adapter à des données temporelles (convolution à une

dimension au lieu de deux) tout en réduisant la complexité du réseau. Bien que l'entrainement soit non supervisé, et que le risque de surapprentissage (*overfitting)* dépendait de la grandeur de l'espace latent, une réduction importante de la complexité du réseau était tout de même nécessaire.

L'architecture du réseau autoencodeur était fortement inspirée de l'architecture du réseau Resnet[112]. Toutefois, le bloc résiduel était simplifié et une convolution unidimensionnelle (dans le sens temporel) a été utilisée plutôt que bidimensionnelle. La dimension temporelle était de 9 étapes, fixée par le choix d'avoir trois valeur par jour (une à chaque 8 heures), avec une évolution de trois jours (voir section 4.2.1). Le nombre de blocs résiduels était fixé à 3 par la structure du réseau (espace latent unidimensionel, utilisation d'un *pooling* de 3 et dimension temporelle de 9). Il restait quelques hyperparamètres, notamment la fonction d'activation précédant à l'espace latent. La grandeur de l'espace latent était un hyperparamètre d'importance[113]. Finalement, vu la relativement petite quantité de données, un abandon aléatoire de connexion (*dropout*) a été utilisé[38] pour réduire le surapprentissage (*overfitting)*. Le pourcentage d'abandon était un autre hyperparamètre. L'optimisation des hyperparamètres a été faite par une approche de recherche en grille (*grid search)* permettant de tester de nombreuses combinaisons.

La performance de discrimination du modèle a été mesurée à l'aide de l'AUROC et son potentiel d'application clinique a été mesuré sur trois seuils (sensibilité fixée à 90%, spécificité fixée à 90% et seuil balancé). L'AUROC a été testé dans les sous-populations de donneurs décrites ci-dessus.

4.4 Modèle final : Classificateur supervisé (Article 1)

Le développement et la validation du modèle final sont décrits plus en détails dans le corps du premier article (chapitre 6), lequel est en voie d'être soumis dans la revue en accès libre *Nature Scientific Reports.* L'article a été conçu en respectant la méthodologie de développement de modèles de prédiction TRIPOD[47].

De manière sommaire, l'autoencodeur a été utilisé comme première partie du modèle. L'espace latent a été utilisé comme des *features* auxquelles ont été ajoutées les deux

informations statiques (présence/absence d'imagerie cérébrale et service d'admission). Un classificateur multicouche avec une activation sigmoïdienne a ensuite été utilisé afin d'obtenir une prédiction binaire. Les données ont été séparées en entrainement (60%), validation (20%) et test (20%). Les données finales ont été calculées sur le groupe test. Un modèle logistique développé en parallèle a été utilisé comme point de comparaison puisqu'il n'existait pas de modèle dans la littérature pouvant servir de comparaison. Le terme « modèle logistique » a été utilisé pour éviter une confusion avec une régression logistique, laquelle nécessitait une méthodologie précise qui n'a pas été utilisée.

Les deux modèles ont été comparés en termes d'AUROC, de score de Brier balancé et de sensibilité et spécificité. Une analyse de calibration a aussi été faite. Les deux modèles ont été comparés dans les sous-groupes de donneurs potentiels décrits ci-dessus. Deux analyses de sensibilité ont été réalisées. La première simulait une approche prospective en calculant l'AUROC à 48h, à 24h et à 8h du congé des soins intensifs. La seconde testait la résilience. Les laboratoires les plus rares étaient progressivement retirés, le modèle était ensuite entrainé sans ces analyses et testé à nouveau. Lorsque pertinent, un intervalle de confiance était calculé par technique de *bootstrap*. Finalement, une analyse d'erreur a été faite en révisant manuellement les dossiers des faux positifs définis comme patients décédés aux soins intensifs, non-donneurs selon l'audit, mais avec des « scores » de prédiction supérieurs à 0.8.

4.5 Validation externe du modèle final (Article 2)

Le second article (chapitre 7) a utilisé les patients hospitalisés dans une unité de soins intensifs séparée, l'unité de soins intensifs coronariens (SIC), comme une cohorte de validation externe. Ces patients avaient été complètement exclus du développement du modèle jusqu'à ce stade. Bien que située dans le même hôpital, son utilisation comme validation externe semblait idéale vu qu'il s'agit d'une unité séparée, traitant d'autre type de patient et prise en charge par des cliniciens différents. Pour éviter une contamination des groupes, les patients ayant été à la fois aux soins intensifs (SI) et aux SIC ont été exclus de la cohorte de validation. Cela en faisait une cohorte de validation séparée, mais potentiellement difficile. Les donneurs potentiels identifiés

aux SIC étant transférés aux SI pour le processus d'évaluation de TQ, il ne restait donc que des donneurs potentiels difficiles à identifier.

Une perte de capacité prédictive est attendue dans une cohorte externe[90,91]. Afin d'évaluer le gain du transfert de connaissance en validation externe, plusieurs approches ont été comparées. Le classificateur étant un RNA multicouche, ces couches pouvaient être préentrainées séparément et leur réentrainement (*fine-tuning)* pouvait être bloqué couche par couche. L'article comparait 6 approches différentes avec différentes combinaisons de préentrainement et de blocage de ré-entrainement. Par la suite, une quantité croissante de données a été fournie au modèle pour son ré-entrainement, pour évaluer le minimum de données requises pour atteindre un plateau.

5 Bases de données et modèle non supervisé

5.1 Résultats

5.1.1 Bases de données

Le jeu de données temporel brut comprenait 24.6 millions de lignes de données. Les colonnes d'importances étaient *dw_pid* (identifiant patient unique), *resultdtm* (horodatage du laboratoire), *specimencollectionmethodcd* (type de prélèvement, c.-à-d. artériel, veineux, urinaire, etc.), *servacro*, *longdesc*, *lbres_ck*. Chaque patient avait une médiane de 584 résultats de laboratoires (IQR 338 -1 119). Il y avait un total de 2 460 types de laboratoires différents. Chaque patient avait une médiane de 101 différents types de laboratoires (IQR 78-126). Finalement, chaque type de laboratoire avait un nombre médian de 18 résultats (IQR 3-193).

La base de données finale comportait un total de 103 valeurs de laboratoires qui sont présentées à l'Annexe A. Sur ces 103 valeurs, il y avait 76 valeurs continues, 26 valeurs catégoriques ordonnées et 1 catégorique non ordonné. Les valeurs catégoriques ordonnées étaient typiquement des quantifications d'éléments sanguins ou urinaires anormaux, lesquels sont chiffrés en 0, 1+, 2+, 3+ ou 0-2, 3-5, 6-10.

La population finale comprenait 24 132 patients, incluant 4 669 patients du groupe de validation externe de l'unité de SIC. 397 donneurs étaient présents dans la population de soins intensifs et 36 dans la population de l'unité de SIC.

Un total de 19 463 patients a été utilisé pour l'entrainement, la validation et le test de l'approche par autoencodeur. La répartition des autres approches est décrite dans les sections correspondantes des deux articles.

La population finale comprenait 8 795 (36.4%) femmes avec un âge moyen 67.9 ans (écart type de 14.5 ans). Le nombre d'admissions était stable dans le temps (Fig. 6) avec une baisse en 2018, possiblement liée au déménagement vers le site du nouveau CHUM.

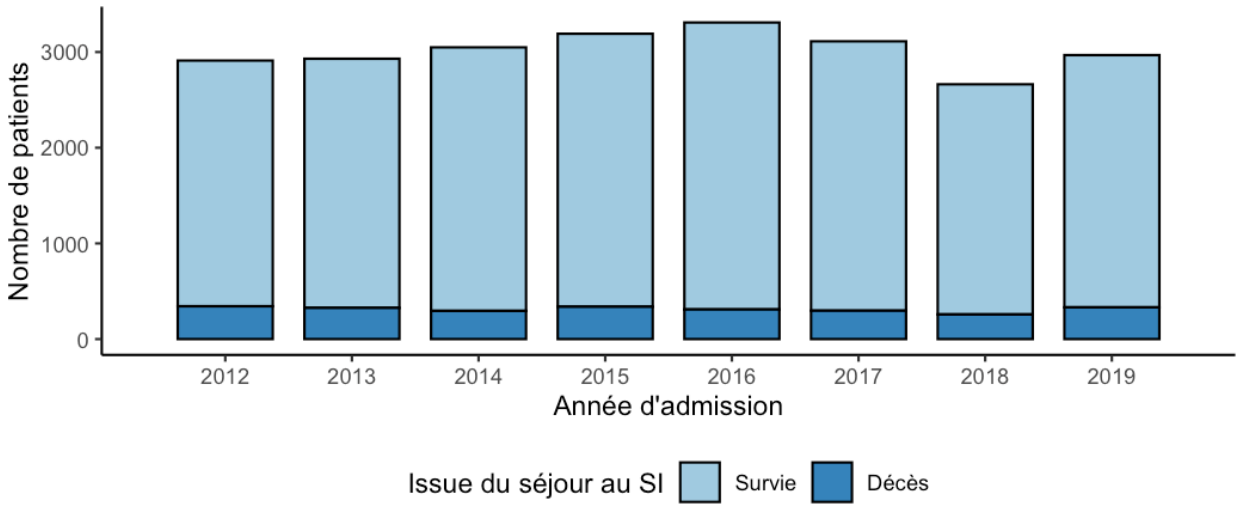


Fig. 6 Nombre de patients admis annuellement

Les patients étaient hospitalisés une durée médiane de 0.97 jour (écart interquartile 0.13-3.2) avant d'être admis aux soins intensifs et y restaient un temps médian de 2.16 jours (écart interquartile 1.09-4.8). La mortalité était de 10.4%. Il y avait un total de 433 donneurs confirmés ou potentiels (1.8%). Le nombre de donneurs du CHUM variait (Fig. 7), mais semblait suivre la variation annuelle des donneurs d'organes au Québec, selon les statistiques de TQ[14] (ligne noire, Fig. 7).

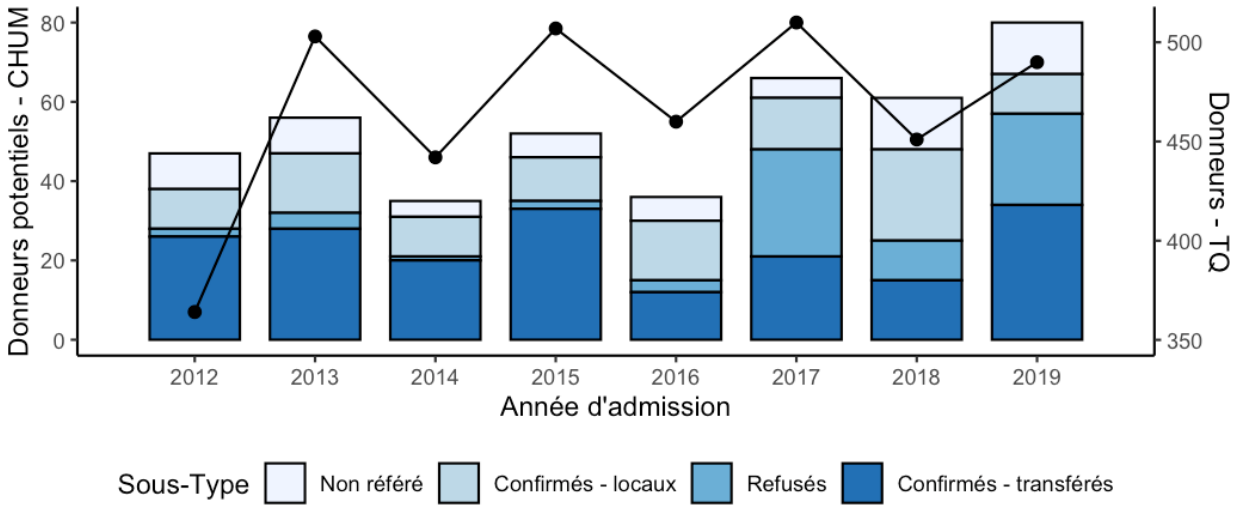


Fig. 7 Nombre de donneurs rapportés et référés annuellement

Le total de 19 463 patients a été séparé de manière aléatoire en trois groupes. 85% (16 214) des patients non-potentiels-donneurs ont été aléatoirement choisis comme groupe d'entrainement de l'autoencodeur. Les patients restants ont été séparés 80% pour la validation et l'ajustement des hyperparamètres (N=2 606 patients) et 20% comme groupe de test (N=644). Le même groupe test a été utilisé pour cette approche et dans le premier article (chapitre 6). Les caractéristiques descriptives des patients peuvent être retrouvées à la Table I. La sélection de 85% des non-potentiels-donneurs permettait un effet d'enrichissement volontaire du groupe de validation et test. Le taux de donneur potentiel passait ainsi de 2.04% (cohorte complète) à 12.2% (groupes de validations et test).

5.1.2 Autoencodeur

L'impact du pourcentage d'abandon aléatoire et la fonction d'activation n'avaient pas un effet majeur sur l'erreur de reconstitution (erreur quadratique moyenne, Fig. 8) sur le groupe d'entrainement ni sur l'aire sous la courbe ROC du groupe de validation (AUC, Fig. 9). La fonction d'activation précédant l'espace latent semblait avoir un impact mineur. Toutefois, la fonction sigmoïde étant continue, elle permettait d'obtenir un espace latent exempt de zéro, une caractéristique utile si le modèle venait à être utilisé comme préentrainement. Une valeur de 0.1 de *dropout* a été choisie. Un espace latent de 200 valeurs semblait la dimension au-delà de laquelle il n'y avait plus de gain d'information. L'architecture finale de l'autoencodeur est présentée à la Fig. 10.

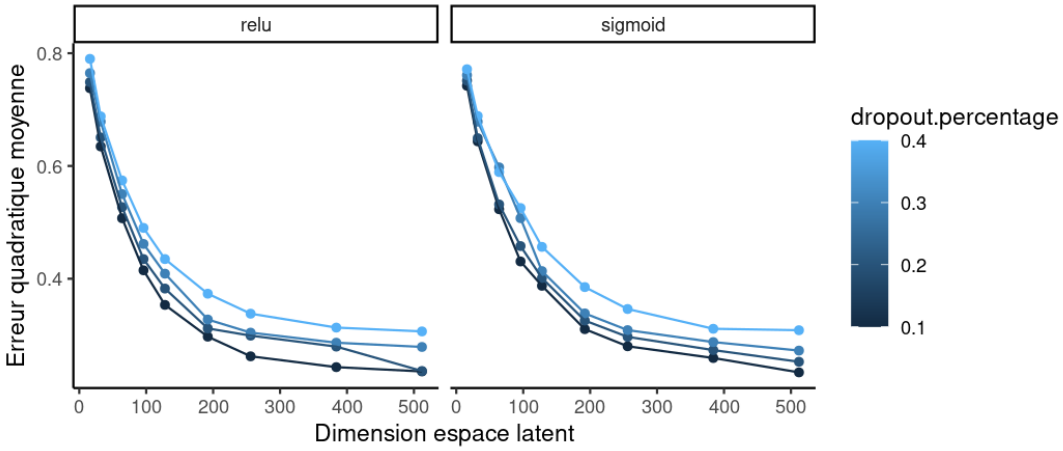


Fig. 8 Impact des hyperparamètres sur l'erreur de reconstitution

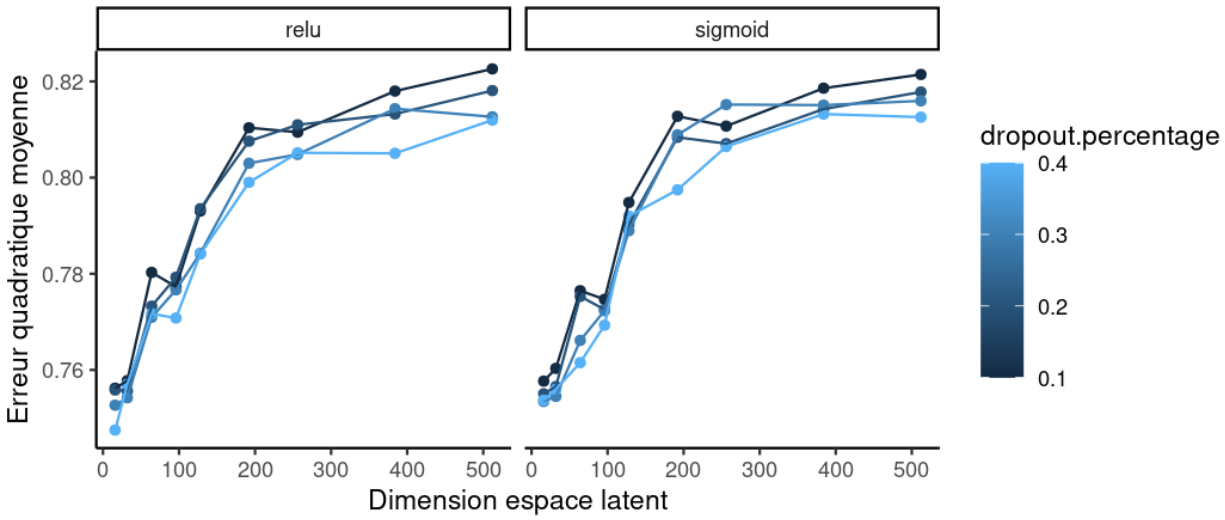


Fig. 9 Impact des hyperparamètres sur l'aire sous la courbe ROC

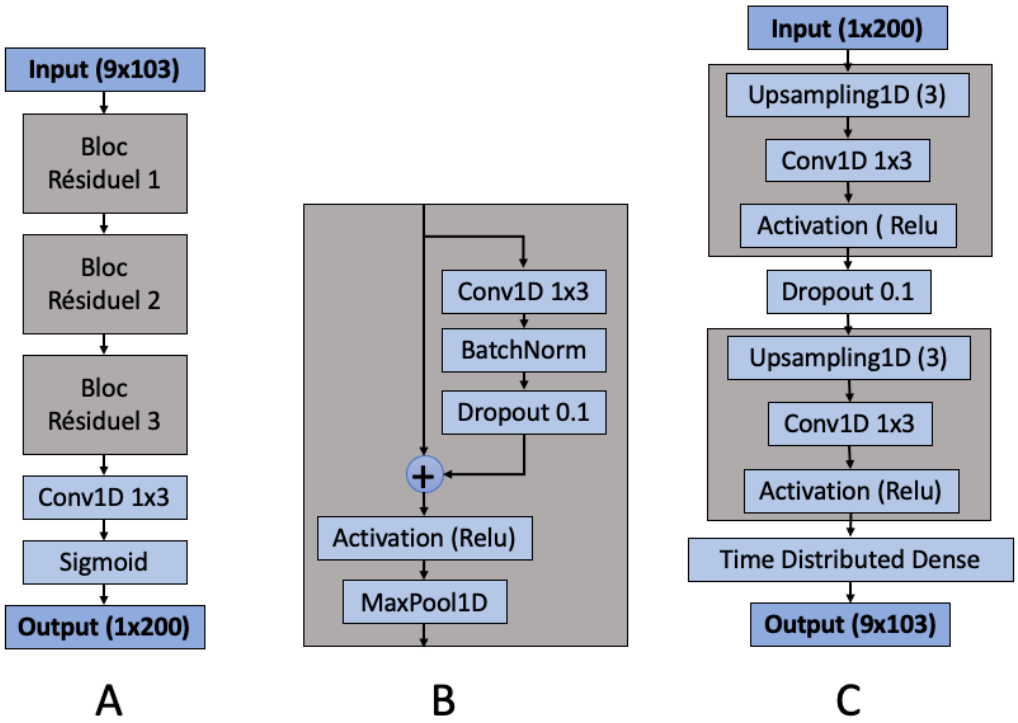


Fig. 10 Schéma du réseau neuronal de l'autoencodeur.
A : Encodeur. B : Détail du bloc Résiduel. C : Décodeur

L'AUROC globale était de 0.855 (Fig. 11), avec sous-division dans les différentes sous-populations de donneurs d'organes. Si l'on choisissait un seuil d'erreur de reconstitution optimal, la sensibilité était de 88% et la spécificité de 73%. Si l'on choisissait un seuil pour une sensibilité fixée de 90%, la spécificité était de 69%. Finalement, si l'on choisissait une spécificité de 90%, la sensibilité était de 51%. Le tableau de contingence de chacune de ces situations est au Tableau 1.

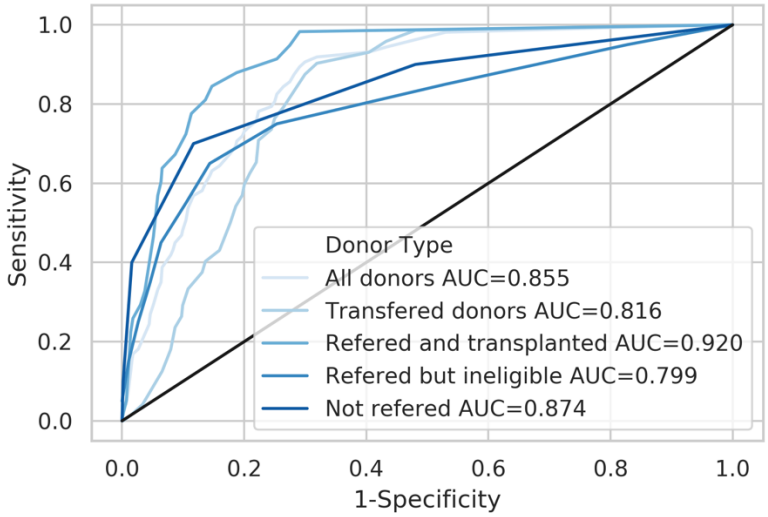


Fig. 11 : Courbe ROC pour le groupe test.

	Seuil optimal		Seuil sensible		Seuil spécifique	
	Donneur+	Donneur-	Donneur+	Donneur-	Donneur+	Donneur-
Prédiction+	70	151	72	174	41	51
Prédiction-	10	413	8	390	39	513

Tableau 1. : Table de contingence de l'autoencodeur

5.2 Discussion

Tout d'abord, l'approche de préparation de la base de données et de sélection des caractéristiques a montré une approche fonctionnelle et efficace. Cette approche d'imputation fortement inspirée de l'approche clinique et utilisable de manière prospective était une force de ce projet. Des travaux subséquents pourraient être faits pour explorer davantage la méthode d'imputation en prélevant des résultats de laboratoire jugés non nécessaires cliniquement.

Ce modèle avait plusieurs forces, notamment un AUROC de 0.855 ce qui était impressionnant pour une approche naïve et non supervisée. Cela supportait la faisabilité de la tâche. De plus, la capacité prédictive était bonne, indépendamment du sous-groupe de potentiel donneur utilisé (Fig. 11). Troisièmement, un seuil optimal donnait une sensibilité de 88%, ce qui semblait encourageant. La spécificité à ce seuil était de 73%. Les résultats obtenus par notre approche non supervisée étaient supérieurs à la seule étude visant à détecter de potentiels donneurs d'organes, publiée par Fernandes et coll.[9] (sensibilité de 77% et une spécificité de 66%). L'approche et les données utilisées sont tellement différentes qu'il semble difficile de pousser plus loin la comparaison. Bien qu'une approche d'autoencodeur soit utilisée dans d'autres domaines, son utilisation en médecine est très émergente et novatrice.

Toutefois, ce modèle a des limitations. Premièrement, l'absence de différence entre les sous-groupes suggère que l'approche non supervisée a commis des erreurs de reconstitution sur l'ensemble des populations minoritaires (potentiels donneurs d'organes ou non), de manière peu spécifique. Cependant, il est possible que le modèle soit moins sensible aux autres populations minoritaires que les potentiels donneurs d'organes, puisqu'il s'agissait de la seule population absente du groupe d'entrainement. Deuxièmement, les résultats sont insuffisants pour une application clinique à ce stade. Avec une spécificité de 73%, une prévalence de donneurs d'organesde 2% donnerait un taux de faux positif important, menant à une valeur prédictive positive de 6% laquelle est incompatible avec une utilisation clinique. La quantité de fausses alarmes qui en découleraient mèneraient à une fatigue et un abandon précoce du système. L'utilisation d'un seuil spécifique augmentait la spécificité mais au prix d'une baisse de la sensibilité à 51%. L'augmentation de valeur prédictive positive restait marginale pour atteindre 10.4%.

Il s'agit, à notre connaissance, de la première utilisation d'un autoencodeur avec des données temporelles de laboratoires. Cette approche a démontré la faisabilité d'utiliser un autoencodeur pour transformer des données temporelles complexes en un format uniformisé linéaire utilisable et potentiellement combinable avec d'autres sources de données. De plus, bien que cette approche n'ait pas démontrée une discrimination suffisante pour ce projet, c'est une approche

qui pourrait notamment être utilisée pour la détection de données ou de patients aberrants à un stade de préparation de base de données.

L'architecture utilisée pour l'autoencodeur a été conservée dans la seconde approche du projet, une approche supervisée cette fois, qui a mené au développement du modèle final. L'autoencodeur y est utilisé comme préentrainement d'un classificateur. La méthodologie et les résultats sont présentés dans le chapitre suivant et dans le premier article qui s'y trouve.

6 Développement et validation du modèle final

6.1 Introduction

L'article qui suit est en cours de révision finale par les co-auteurs pour être soumis à la revue en accès libre *Nature Scientific Reports.* Nicolas Sauthier est l'auteur principal de ce manuscrit. Il a conçu le modèle et le projet, préparé la base de données, conçu et réalisé les analyses qui s'y trouvent. Il a écrit l'ensemble du code informatique utilisé dans ce projet. Rima Bouchakri et Kip Brown ont extrait les données, aidés à la conception du modèle non supervisé et à la révision du manuscrit. Le dernier auteur est le directeur de recherche de ce mémoire, Michaël Chassé. Il a contribué à la conception du projet, à la conceptualisation et la révision des analyses de même qu'à la supervision globale du projet. Michaël Sauthier et François Martin Carrier ont aidé à la conception des analyses et à la révision du manuscrit. Louis-Antoine Mullie a aidé à la révision du manuscrit. Héloïse Cardinal, Marie-Chantal Fortin et Nadia Lahrichi ont aidé à la conception du projet et à la révision du manuscrit.

6.2 Detection of potential organ donors; an automatic deep learning approach on temporal data.

Authors: Nicolas Sauthier, MD; Rima Bouchakri, PhD; Kip Brown, PhD; Michaël Sauthier, MD PhD; François Martin Carrier, MD MSc; Louis-Antoine Mullie, MD;Héloïse Cardinal MD PhD; Marie-Chantal Fortin, MD PhD; Nadia Lahrichi PhD; Michaël Chassé, MD PhD

6.2.1 Abstract

Organ donation is not meeting demand and we may be missing 30-60% of potential donors. It relies on manual identification and referral to Organ Donation Organizations (ODO) and neural networks could be used to help with the identification of potential donors. Using retrospective routine clinical and temporal evolution of laboratory results, we developed and tested a neural network model to automatically identify potential organ donors. We first trained a convolutive autoencoder that learned from the longitudinal changes of over 100 types of laboratory results. We then added a deep neural network classifier. This model was compared to a simpler logistic regression model. We observed an AUROC of 0.966 (CI 0.949-0.981) for the neural network and 0.940 (0.908-0.969) for the logistic regression model. At a prespecified cutoff, sensitivity and specificity were similar with 84% and 93% for both models. The neural network accuracy remained stable across donor subgroups and in a prospective simulation. The linear regression model AUROC decreased when applied in a more challenging potential donors' sub-population as well as in the prospective simulation (0.81 vs 0.71). Our findings suggest the feasibility of using machine learning models to help with the identification of potential organ donors using routine clinical data.

6.2.2 Introduction

Despite a stagnant to slow improvement in total organ donors over the last 20 years[114], organ donation in Canada is still not meeting the demand. In 2021[115], 4 043 patients were on waiting lists, only 2 782 organs were transplanted, and 250 patients died waiting.

Organ transplantation depends on potential organ donor identification and conversion to actual donors. The former is a major challenge that relies heavily on the training of medical teams which is inefficient given the rarity of deceased organ donation. Multiple retrospective cohort studies suggested that between 30% and 60 % of potential organ donors are either not identified or not referred to ODO[3–6]. More efficient identification of potential organ donors could lead to an increase in the total number of referrals to Organ Donation Organization (ODO) and thus substantially increase the number of organ donors. Doing so would result in more organs available for transplantation.

Concurrently, the medical field is seeing a surge in medical data from the rapid development and implementation of electronic health records (EHR) and the interconnection of clinical databases. Advances in the machine learning (ML) field help to make use of those big data[42]. A biological-inspired ML algorithm called neural networks (NN) is capable of feats such as autonomous driving, image recognition, and pattern analysis. Its application in medicine is showing great potential[79] with applications in medical imaging, mortality and readmission prediction[116] or continuous severity scoring[80]. The evolution toward neurological death and organ donation are very complex clinical patterns. It remains unclear how a complex ML approach, such as an NN, would perform in such situations compared to simpler models [117], such as a logistic regression and other traditional epidemiological approaches. A reliable model able to detect subtle clinical evolution patterns could improve the efficiency of organ donor identification.

Regarding organ donors' prediction, a clinical score has been derived to estimate the probability of successful donation after cardiac death[89]. That score has the potential to be complementary to a predictive model identifying potential organ donors, since a patient needs to be identified as a potential organ donor before this score may be used. There is no ML model exist to predict potential organ donors. Most of the current application of machine learning in the field of organ transplantation focuses on predicting recipients' outcomes. Published models improved the prediction of organ recipients' survival while on the waiting list[84], improved the survival prediction after organ transplants[84–86], improved the graft rejection prediction[87] or helped in guiding clinicians in choosing an anti-rejection drug post renal transplant[88].

The primary objective of this study was to develop a predictive model for the identification of potential organ donors among hospitalized patients in an intensive care unit using routinely collected clinical data. Our secondary objectives were to compare the predictive accuracy of a neural-network-based model compared to a logistic regression model used as a baseline, to evaluate our models in organ donor subpopulations, and to evaluate the model in a simulated 48h prospective window. This article followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) checklist[118] and the Guidance for Development and Reporting of Prediction Models[47].

6.2.3 Methods

6.2.3.1 Population and outcomes

The dataset was derived using the electronic health records from intensive care unit (ICU) admissions from Jan 1st, 2012, to Dec 31st 2019, from the CHUM (Centre Hospitalier de l'Université de Montréal, Montreal). We included all adult patients who were admitted to an ICU. We excluded patients with a hospital length of stay shorter than 16h. For patients with multiple ICU admissions, we included data only for the last ICU stay to avoid correlated data.

Our predicted outcome was becoming a potential organ donor. We defined this group of patients as patients following one of the four categories: (1) actual organ donors either through local identification (admitted to the CHUM for a condition that eventually evolved to death and organ retrieval) or (2) through inter-hospital transfer (neurological death diagnosis made in another peripheral hospital and transferred to the CHUM for organ retrieval); (3) patients referred to the ODO for donation but deemed ineligible for donation by the ODO (next of care refusal, medical contraindications detected in the workup, etc.); (4) potential organ donors not referred to ODO. This last category of patients was identified by the ODO through local continuous death audits, which excluded patients with any recent, active, or metastatic cancer, disseminated infection, or multi-organ failure. The death audit defined potential organ donors as either as a patient with a severe neurological condition, mechanically ventilated, who died within 24 hours of the end of care, or as a patient without a severe neurological condition, mechanically

ventilated, who died within 3 hours of the end of care. The study and data extraction were approved by the CHUM Research Ethics Board and individual patient consent was waived given the low risk and retrospective nature of the study.

6.2.3.2 Predictors

Variable selection

We included time-varying laboratory variables and static variables. Since identified organ donor patients may be treated differently than other patients (more frequent or rarer laboratory analysis and investigation, shorter length of ICU stay), we aimed to blind our model to the specifics of medical practice. The intent was to avoid the bias of predicting the medical practice instead of true clinical patterns [47]. This was done in four distinct ways. First, we focused our effort on laboratory analyses, which are impartial, standardized, and accessible electronically even in hospitals without EHR. Second, we excluded rare analyses (defined as present in less than 10% of all ICU patients). Third, we excluded usual demographical values (age, biological sex, height, or weight) since they are not a contraindication for organ donation and may induce a latent bias. Finally, to blind the model to the higher frequency of laboratory analysis often done in potential organ donors, we only kept the last 72 hours of a patient's ICU stay (discharged from the ICU or death), divided into 9 blocks of 8 hours keeping only the last value within each time block.

We included only two static variables: the specialty responsible for the patient (neurosurgery, internal medicine, cardiology, etc.) and the presence or absence of a head radiological imaging exam.

Missing values

Since data missingness pattern was probably associated with the value of the outcome, missingness was considered not to be at random[66]. Missing specialty responsible values were coded as a distinct “unknown” specialty. To mimic a clinician thinking with a missing laboratory value and to increase the model usability in a real-world setting, temporal data was imputed in a two-step computationally efficient way. We first used the last value carried forward (LVCF), i.e. if

a patient had a laboratory value done in the first 8-hour block, it's carried in all the subsequent 8-hour blocks until updated by a new lab result. For values still missing after LVCF, we imputed a normal value randomly sampled from a Gaussian distribution in the normal range of each laboratory analysis. This decision was made to reflect the fact that if a physician has no clinical reason to order a lab test, its pre-test probability of being either similar to the previous value or normal is high.

6.2.3.3 Models' architecture

We fitted two different models: a NN temporal model and a classic simpler logistic model (LM) as a baseline comparison model.

The NN model used the temporal aspect of the laboratory data combined with the static values. Its architecture is schematized in Supplementary S1: and Supplementary S2: (supplementary materials). The temporal values were embedded with a convolutive autoencoder (AE)[58,119]. AE is a subtype of NN architecture used in non-supervised learning. An AE is trained with the same information (image, text, laboratory data, etc.) presented at the entrance and the exit of the network. The data is compressed and transformed into a reduced size called latent space and is then decoded back to its original form. The AE learns to encode the data while minimizing loss. This type of architecture has been shown useful in multiple ways such as an embedding and a dimensionality reduction tool, to reduce noise[59], for transfer learning and pretraining[60] and as an anomaly detection approach. We designed the convolutive AE based on a well-known[55] convolutional neural network (CNN)[112] using Python and Keras with TensorFlow[120–122]. We adapted it to slide only along the temporal dimension of lab results to detect patterns in the temporal evolution of the laboratory. We used a mix of dropout layers and L2 regularization on our models to help avoid overfitting. The final architecture is presented in Fig. S1 and S2 (supplementary materials).

We used our AE to (1) take the maximum out of the temporal component of the data, (2) embed the temporal data in a one-dimensional format, (3) act as an anomaly detector since it was trained only on non-donor patients, and (4) allow a smaller size classifier on top so it could be trained quicker. Our AE is unsupervised and only trained with the temporal data.

For the classifier part, we used a simple four layers deep NN ending with a logistic layer. Temporal embedded data were concatenated with static data. Static data were encoded either as one-hot for binary data or target mean encoding with smoothing for multiclass data.

The simpler logistic model used only the last value for each laboratory analysis combined with each static value.

1.1.1.1 Data structure

Potential organ donors are a rare subpopulation, encompassing around 2% of ICU patients based on preliminary data exploration. We approached the class imbalance problem with a mixture of purposeful subsampling and oversampling[123]. 85% of the non-donor patients were randomly selected and used as the embedding training set, allowing subsampling while using this data for the autoencoder. The rest of the patients (15% of non-donors and all potential donors) were randomly divided into a train/validation/test set (60%/20%/20%). During training, proportionally more weight was put on the minority class. That means that if non-donors outnumber donors 100 to 2 the training weights of donors will be 100/2 = 50 and non-donors 1. The validation dataset was merged with the train dataset to augment the training size of the final model before test data unblinding. Results are all made on the test dataset which was excluded from model development.

1.1.1.2 Statistical analysis

We reported performance, discriminative and calibration properties of our models (NN and LM) compared their different properties. We compared the overall performance of our NN model to our LM models using a scaled Brier score, their discriminative properties using AUROC and their calibration using calibration curves[47]. Confidence intervals were obtained using percentiles of 2000 bootstraps on the test set. Group were compared using an unpaired t-test. We compared the ROC curve of our four subgroups of donors and potential donors. In an approach to maximize potential donor detection, the goal was to choose a threshold giving a high sensitivity. Because the project had a potential prospective application, the optimal cutoff had to be derived from the training data and not from the test data. The cutoff was derived using a 3-fold approach to the

training dataset choosing the average threshold that gave a 90% sensitivity on the training dataset.

We did 2 subgroups analysis. First, we compared the ROC curve of our four subgroups of donors and potential donors. Second, we simulated a prospective approach to compare accuracy 48h, 24h and 8h before ICU discharge or death. Finally, we did two sensitivity analyses. First, to ensure the model was resilient to removal of rare labs, we retrained the models and analyzed the variation in the model performance by removing predictors present in less than 20% of all ICU patients, then 30% and so on. Second, we manually reviewed the files of patients not listed as potential organ donors, but who were predicted as such with a high degree of confidence (>0.75) by either of the models. Statistical analyses were done using R and Python[29,31].

6.2.4 Results

Baseline characteristics of the population are reported in Table I. Our complete dataset used 19 067 patients and included 397 potential donors for a prevalence of the outcome of 2.1 % in the study population but of 12% in the sub-sample used to train, validate and test of our models. After excluding rare laboratory analyses, the NN model and LM were trained on 103 distinct laboratory analyses, reported in the table S3 (supplementary material), as well as other non-laboratory predictors (2 variables).

Table I. Population Characteristics

	AE Train n= 16 213	Train n=1 972	Validation n=634	Test n=644
Sex F n(%)	5 871 (36,2%)	737 (37,4%)	249 (39,3%)	262 (40,7%)
Age mean (std)	67,2 (14,5)	66,4 (14,7)	67,2 (14,2)	65,4 (15,6)
Donors' subtypes n (%)				
Non-donor	16 213 (100%)	1 734 (87,9%)	555 (87,5%)	564 (87,6%)
Transferred	-	121 (6,1%)	32 (5,0%)	36 (5,6%)
Local	-	53 (2,7%)	25 (3,9%)	29 (4,5%)
Referred	-	42 (2,1%)	13 (2,1%)	10 (1,6%)
Not referred	-	22 (1,1%)	9 (1,4%)	5 (0,8%)
Principals' reasons for admission n (%)				
Ischemic heart diseases	3 855 (22,9%)	375 (18,4%)	132 (20,8%)	130 (19,5%)
Other forms of heart disease	1 833 (10,9%)	235 (11,5%)	60 (9,5%)	55 (8,2%)
Cerebrovascular diseases	1 033 (6,1%)	175 (8,6%)	66 (10,4%)	59 (8,8%)
Diseases of the arteries/arterioles/capillaries	677 (4,0%)	75 (3,7%)	22 (3,5%)	18 (2,7%)
Total Diseases of the circulatory system	7 907 (47,0%)	906 (44,5%)	306 (47,0%)	282 (42,3%)
Neoplasms	2 729 (16,2%)	305 (15,0%)	104 (16,0%)	95 (14,2%)
Diseases of the digestive system	1 649 (9,8%)	192 (9,4%)	49 (7,5%)	76 (11,4%)
Consequences of external causes	1 299 (7,7%)	141 (6,9%)	54 (8,3%)	53 (7,9%)
Diseases of the respiratory system	1 063 (6,3%)	132 (6,5%)	37 (5,7%)	43 (6,4%)
Most frequent admission services n (%)				
Cardiac surgery	5 997 (37,0%)	615 (31,2%)	210 (33,1%)	199 (30,9%)
General surgery	1 616 (10,0%)	184 (9,3%)	58 (9,1%)	57 (8,9%)
Hepatology	787 (4,9%)	121 (6,1%)	31 (4,9%)	33 (5,1%)
Internal Medicine	692 (4,3%)	107 (5,4%)	33 (5,2%)	34 (5,3%)
Neurosurgery	660 (4,1%)	99 (5,0%)	37 (5,8%)	34 (5,3%)
Intensive care	519 (3,2%)	176 (8,9%)	54 (8,5%)	52 (8,1%)
Burn unit	587 (3,6%)	64 (3,2%)	19 (3,0%)	27 (4,2%)
Length of stay in hours median [IQR]	50,7 [95,2]	51,6 [93,6]	51,4 [78,3]	48,5 [90,4]
Death in ICU n (%)	1 573 (9,7%)	413 (20,9%)	136 (21,5%)	133 (20,7%)

AE: Auto-Encoder.

AUROC curves for each model as a whole and separated by organ donor subtypes are presented in Figure I. Overall, in the test dataset, the AUROC of the NN model was marginally however statistically superior (p=0.014) with 0.966 (bootstrapped 95% CI 0.949-0.981) compared to the logistical model with an AUROC of 0.940 (bootstrapped 95% CI 0.908-0.969). The scaled Brier score was also statistically superior in the NN model (0.481 vs 0.352, Table II).

The cutoffs used were obtained by a 3-fold approach on the train group to obtain a high sensitivity cutoff, giving a cutoff of 0.4 for the neural network and 0.47 for the logistical model. As expected, there is a difference in the aimed sensitivity (90%) and the obtained one (84% for both NN and LM, see Table III). Both models had similar sensitivities (84%) and specificity (93%) (Table II). Confusion matrices are presented in Table III. Confusion matrices by subgroups are presented in the tables S4 to S8 (supplementary materials). To obtain an actual 90% sensitivity, the actual cutoff on the test set was 0.21 for the NN model and 0.08 for the LM. Confusion matrix at those cutoffs is presented in the table S9 (supplementary materials). At those cutoffs, specificity was 88% for the NN model and 74% for the logistical model.

Calibration curves show that both models tend to underestimate the actual proportion of potential organ donors especially between a predicted probability of 0.3 and 0.8 with better accuracy at low and high predicted probabilities (see Fig S10, supplementary materials).

As a prespecified sensitivity analysis, we manually reviewed the medical files of the 11 false positive, who died in the ICU, while predicted as donor by either model. Results are presented in table S10 (supplementary materials). Of 11 cases, almost half were not potential donors because of neoplasia, even if two (# 10 and #11) could still have been referred because of the low likelihood of metastasis. Two (#3 and #8) were actual potential organ donors missed by the death audit of our ODO. Sensitivity analysis shows that the model was resilient to laboratory removal with only a small decrease in the AUROC when only the most frequent laboratory analyses were kept (Fig S11, supplementary materials). The simulated prospective approach Figure II showed that the AUROC decreases with the increase of the delay between the discharge of ICU and the test point. However, the NN keeps a better accuracy in the two longest delays of 24 and 48h.

Figure I. ROC Curves

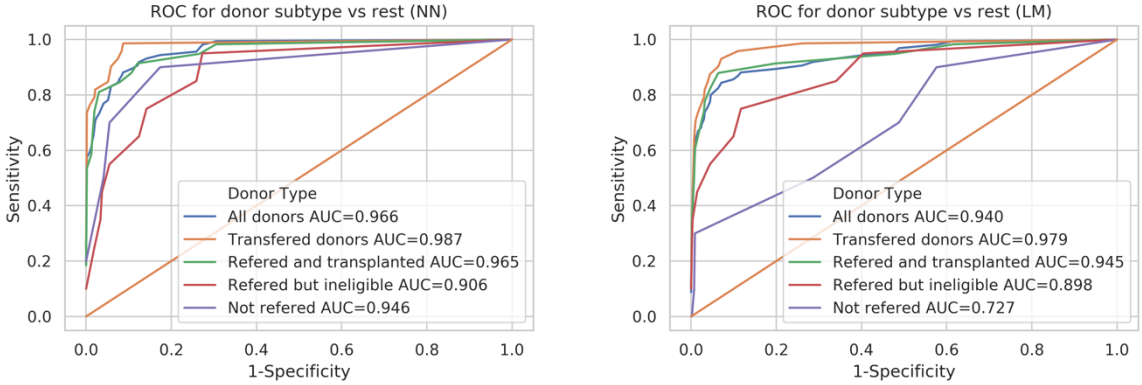


Table II. Models' performance in the test set

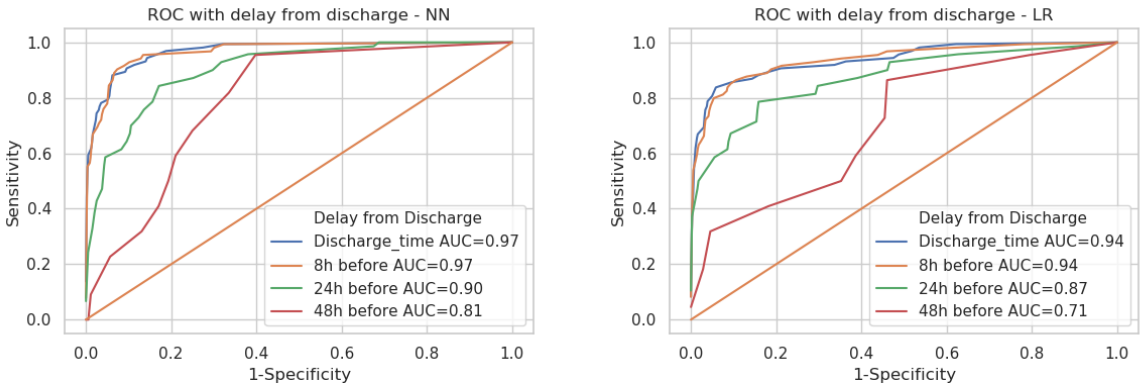
	NN model	Logistic model	p-value
ROC AUC	0.966 (0.949-0.981)	0.940 (0.908-0.969)	0.014
Scaled Brier score	0.481 (0.306-0.614)	0.352 (0.135-0.518)	0.049
Sensitivity	0.838 (0.750-0.914)	0.838 (0.750-0.917)	0.99
Specificity	0.926 (0.903-0.947)	0.934 (0.913-0.954)	0.36

Data presented as bootstrapped median with 95% confidence interval. p-values were calculated based on bootstrapped data.

Table III. Confusion matrix in the test set

Neural Network		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	67	42
Predicted non-potential organ donors	13	522
Logistic model		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	67	37
Predicted non-potential organ donors	13	527

Figure II. Prospective analysis over 48h before ICU discharge



6.2.5 Discussion

We propose an innovative approach to the problem of organ donor identification. We were able to develop and internally validate a NN model with high accuracy to detect potential organ donors based on routinely collected temporal and static clinical data. Our model did not require any human intervention and can use clinical data with minimal pre-processing.

This model is the first evidence to support the use of real-world data to help screen for potential organ donors. The only comparable work was published by Fernandes and al. [9] who reported a model with 77% sensitivity and 66% specificity. However, their model aimed to identify catastrophic neurologic events using specific keyword identification on head CT scan reports[32]. This approach required the scans to be interpreted by a radiologist and thus required human intervention.

The more complex temporal model (NN) marginally outperformed the non-temporal simpler version (LM). The NN kept a good accuracy in the more complex clinical patterns. The NN also outperformed the simpler LM when simulating a prospective identification of donors up to 48 hours before the time of final donor classification. This could be explained by the added value of the clinical temporal evolution and by the fact that the NN had access to more data points than the LM, which only had access to the last laboratory timepoint. Subsequent work is needed to improve the model and specifically reduce the false positive rate. Since a lot of the false positives of the model were not eligible because of known neoplasia, that information could in the future be used to update the models and improve their performance.

In subgroup analyses, we observed that our model performed better on donors that were also identified by the clinicians. These subpopulations represent the largest donor subtype making it likely that the model learned mostly from this subtype. Also, those subtypes may be more clinically distinct with more stable laboratory values, making them easier to detect. Our model was also able to detect a significant proportion of potential organ donors that were missed or not referred by the clinical teams. Although the accuracy of the models was slightly lower in this group, those findings are of significant clinical interest since those patients were missed and did not have the opportunity to be assessed for donation. Since even the detection of one additional patient is of potential clinical benefit, we believe that if externally validated, such models could help support clinicians in the screening of potential organ donors. Interestingly, when we conducted a review of the classification errors of the models, the NN model detected two patients that have been missed by both the manual death audit and by the clinicians, potentially suggesting a higher sensitivity than the manual death audit alone.

Our study has a few limitations. First, our initial design required that a proportion of the data be used to train the autoencoder. Training the model as a single classifier could give different results. However, we think that the difference would be marginal, and our approach allows advantages such as the possibility of easily merging multimedia information in future iterations of the model (radiology images, CT scans, vital signs, etc.). Second, it is a retrospective study based on the data of a single quaternary transplant center, where clinicians are highly trained in the detection of potential organ donors. Truly missed organ donors are rare. In most of the non-referred patients, we found that organ donation was considered by the clinician and the option was not pursued often because of family refusal. However, those patients still have a clinical pattern like a truly missed potential organ donor. Itis unknown how the accuracy of our model would translate in a true, unsimulated, prospective setting, or in a different institution, and will thus require external and prospective validation before being considered for clinical use. Finally, our model is trained using at least 16 hours of temporal data and as such does not apply to neurological catastrophic events that would need a quick decision while in the trauma bay or the emergency room. A decision for those patients will still need to be done by the clinician ideally guided by advanced care planning from the patient. Alternatively, those patients could benefit from an observation period in the ICU, as recommended by some[124,125].

In conclusion, we demonstrated the feasibility to develop two models that can identify potential organ donors using routinely collected clinical data. The models identified patients that were not detected by the medical teams and manual death audits. Future work is required to validate the models externally and prospectively and to further improve their prediction accuracy.

6.2.6 Competing interest

This study was founded by a research innovation grant by the Canadian Donation and Transplantation Research Program (CDTRP). FM Carrier, M Sauthier and M Chassé are recipients of a research career award from the Fonds de recherche du Québec-Santé.

6.2.7 Data availability

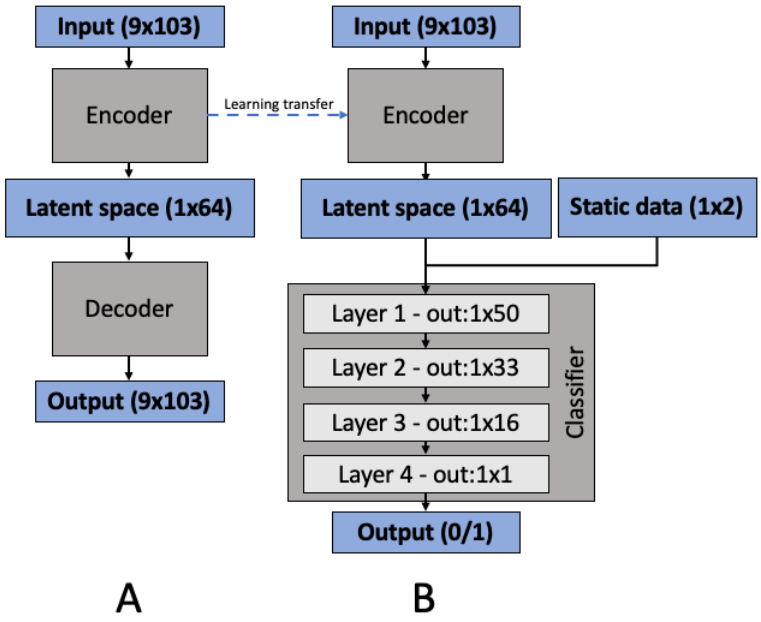
Access to the data is restricted to the research team by the ethical board according to provincial laws. Dataset may be accessible after privacy agreements with the research team.

6.2.8 Bibliography

(Uniformisée avec le mémoire, Voir chapitre 0)

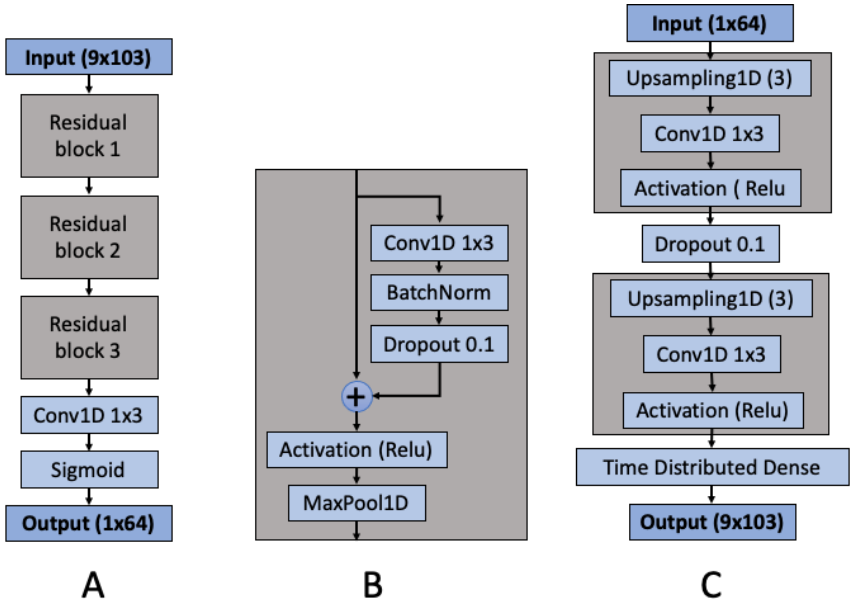
6.2.9 Supplementary materials

Supplementary S1: Neural Network architecture



A) Autoencoder. B) Classifier

Supplementary S2: AutoEncoder architecture



A) Encoder. B) Residual block C) Decoder

Supplementary S3: Laboratory values used in model development

Lab Value	%missing	Median	IQR	Normal range
white_blood_cell_count	0,1%	10,5	6,7	[4-11]
hemoglobin	0,1%	96,0	32,0	[120-196]
hematocrit	0,1%	0,3	0,1	[0,35-0,6]
red_blood_cell_deviation_width	0,1%	14,8	2,7	[11,5-20]
red_blood_cell_count	0,1%	3,2	1,0	[3,8-6,2]
platelet_count	0,1%	188,0	129,0	[140-500]
mean_corpuscular_volume	0,1%	90,9	6,8	[80-101]
mean_corpuscular_hemoglobin_concentration	0,1%	331,0	18,0	[300-365]
mean_corpuscular_hemoglobin	0,1%	30,2	2,5	[24-33,5]
lipemia_presence	0,1%	0,0	0,0	[0-0]
icterus_presence	0,1%	0,0	0,0	[0-0]
hemolysis	0,1%	0,0	0,0	[0-0]
Neuhil_count	0,1%	8,0	6,3	[1,3-7,7]
monocyte_count	0,1%	0,8	0,6	[0-1,6]
lymphocyte_count	0,2%	1,1	1,0	[1-4,1]
sodium	0,2%	140,0	5,0	[135-145]
creatinine	0,2%	81,0	59,0	[42-112]
potassium	0,2%	4,0	0,7	[3,5-5]
eosinophil_count	0,3%	0,1	0,2	[0-0,8]
basophil_count	0,4%	0,0	0,0	[0-0,3]
urea	0,4%	7,2	6,2	[2,8-8,8]
mean_platelet_volume	0,7%	10,3	1,4	[6,5-13,5]
glucose	1,1%	7,5	3,1	[4-6,2]
chloride	1,2%	106,0	7,0	[96-106]
magnesium	1,2%	0,8	0,2	[0,7-1,01]

Lab Value	%missing	Median	IQR	Normal range
phosphate	1,4%	1,1	0,4	[0,72-1,64]
albumin	2,3%	28,0	9,0	[36-52]
total_calcium	2,7%	2,1	0,3	[2,17-2,56]
inr	3,6%	1,1	0,3	[0,8-1,2]
partial_thromboplastin_time	3,8%	29,0	18,0	[22-32]
creatine_kinase	14,2%	195,0	397,0	[24-213]
total_bilirubin	15,0%	13,0	14,3	[7-23]
alanine_aminotransferase	15,0%	30,0	53,0	[8-39]
aspartate_aminotransferase	15,7%	39,0	57,0	[13-39]
hs_troponin_t	16,1%	112,0	333,8	[0-18]
corrected_total_calcium	23,6%	2,3	0,2	[2,2-2,58]
alkaline_phosphatase	24,1%	75,0	65,0	[36-110]
venous_ph	25,8%	7,4	0,1	[7,31-7,43]
fio2	27,2%	0,4	0,2	[0,21-0,21]
venous_pco2	29,1%	44,0	11,0	[38-54]
venous_bicarbonate	29,1%	24,0	5,2	[21-29]
venous_po2	29,1%	39,0	11,0	[35-95]
arterial_po2	29,2%	108,0	57,6	[70-110]
arterial_pco2	29,2%	38,0	10,0	[32-45]
arterial_bicarbonate	29,2%	23,0	4,6	[19-28]
venous_lactic_acid	33,1%	1,3	1,0	[0,56-2,4]
venous_o2_sat	33,6%	0,7	0,2	[0,7-1]
lipase	34,0%	23,0	29,0	[10-102]
arterial_o2_sat	34,2%	1,0	0,0	[0,92-1]
total_protein	37,2%	56,0	14,0	[63-81]
urinary_ph	38,3%	5,0	1,5	[4,8-8]
urinary_density	38,4%	1,0	0,0	[1-1,03]
urinary_protein	38,4%	0,0	0,3	[0-0]

Lab Value	%missing	Median	IQR	Normal range
urinary_glucose	38,5%	0,0	0,0	[0-0]
urinary_blood	38,5%	0,3	25,0	[0-0]
urinary_bilirubin	38,5%	0,0	0,0	[0-0]
urinary_cetones	38,6%	0,0	0,0	[0-0]
urinary_urobilinogen	38,7%	0,0	0,0	[0-0]
urinary_nitrite	38,9%	0,0	0,0	[0-0]
urinary_leucocytes	39,0%	0,0	0,0	[0-0]
ionized_calcium_ph74	42,0%	1,2	0,1	[1,12-1,32]
arterial_ph	44,0%	7,4	0,1	[7,35-7,45]
lactate_dehydrogenase	45,6%	205,0	142,0	[104-205]
fibrinogen	47,8%	3,4	2,2	[2-4,5]
anticoagulant	48,4%	0,0	1,0	[0-0]
gamma_glutamyl_transferase	48,9%	44,0	75,0	[7-47]
amylase	51,5%	55,0	60,5	[20-104]
temperature	53,8%	37,0	0,0	[36-38]
osmolality	55,3%	289,0	16,0	[275-300]
urinary_polychromia	55,4%	1,0	0,0	[0-0]
thrombin_time	56,2%	17,0	7,0	[12-18]
ph	58,6%	7,4	0,1	[7,37-7,43]
nucleated_red_blood_cells	60,2%	0,0	0,0	[0-0,1]
erythrocytes	63,9%	4,0	54,0	[0-2]
ck_mb	65,0%	16,0	24,3	[0-19]
leucocytes_count	65,3%	4,0	7,0	[0-2]
uric_acid	67,4%	310,0	178,0	[167-441]
arterial_lactic_acid	67,9%	1,4	1,4	[0,6-2,4]
anisocytosis_presence	68,2%	1,0	1,0	[0-0]
base_excess	68,8%	-1,4	5,2	[-2,5-2,5]
anion_gap	71,2%	8,0	4,0	[4-14]

Lab Value	%missing	Median	IQR	Normal range
venous_base_excess	71,6%	-0,7	6,0	[-2-3]
cholesterol	72,3%	3,4	1,6	[3,16-7,3]
triglycerides	72,7%	1,4	1,0	[0,43-2,82]
hdl_cholesterol	74,3%	0,9	0,4	[0,8-2,38]
plt_anisocytosis_presence	75,3%	1,0	0,0	[0-0]
urinary_mucus_presence	75,6%	1,0	0,0	[0-0]
urinary_bacteria	76,1%	1,0	0,0	[0-0]
elliptocytes_presence	77,6%	1,0	0,0	[0-0]
urinary_pavimentous_cells_presence	77,9%	1,0	0,0	[0-0]
echinocyts_presence	81,4%	1,0	0,0	[0-0]
thyroid_stimulating_hormone	82,0%	2,1	2,6	[0,35-5,5]
direct_bilirubin	83,9%	19,6	29,6	[0-3,6]
urinary_ac_ascorb	84,1%	0,0	0,0	[0-0]
giant_platelets_presence	84,4%	1,0	0,0	[0-0]
hba1c	85,2%	0,1	0,0	[0,04-0,06]
acanthocytes_presence	85,9%	1,0	0,0	[0-0]
urinary_hyalin_cylinder_presence	86,3%	1,0	3,0	[0-0]
atypia_lympho_presence	87,6%	1,0	0,0	[0-0]
globulins	87,8%	28,0	11,0	[21-34]
toxic_granulation_presence	87,8%	1,0	0,0	[0-0]
target_cells_presence	88,1%	1,0	0,0	[0-0]
doehle_body_presence	88,3%	1,0	0,0	[0-0]

Values are presented in median and interquartile range (IQR). Color are made to reflect the grouping used in supp. Fig 3.

Supplementary S4: Confusion matrix in test set – Transferred donors

Neural Network		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	33	42
Predicted non-potential organ donors	3	522
Logistic model		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	33	37
Predicted non-potential organ donors	3	527

Supplementary S5: Confusion matrix in test set - Referred and transplanted donors

Neural Network		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	24	42
Predicted non-potential organ donors	5	522
Logistic model		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	26	37
Predicted non-potential organ donors	3	527

Supplementary S6: Confusion matrix in test set - Referred and not transplanted donors

Neural Network		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	6	42
Predicted non-potential organ donors	4	522
Logistic model		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	6	37
Predicted non-potential organ donors	4	527

Supplementary S7: Confusion matrix in test set – Not referred donors

Neural Network		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	4	42
Predicted non-potential organ donors	1	522
Logistic model		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	2	37
Predicted non-potential organ donors	3	527

Supplementary S8: Confusion matrix in validation set – All potential donors

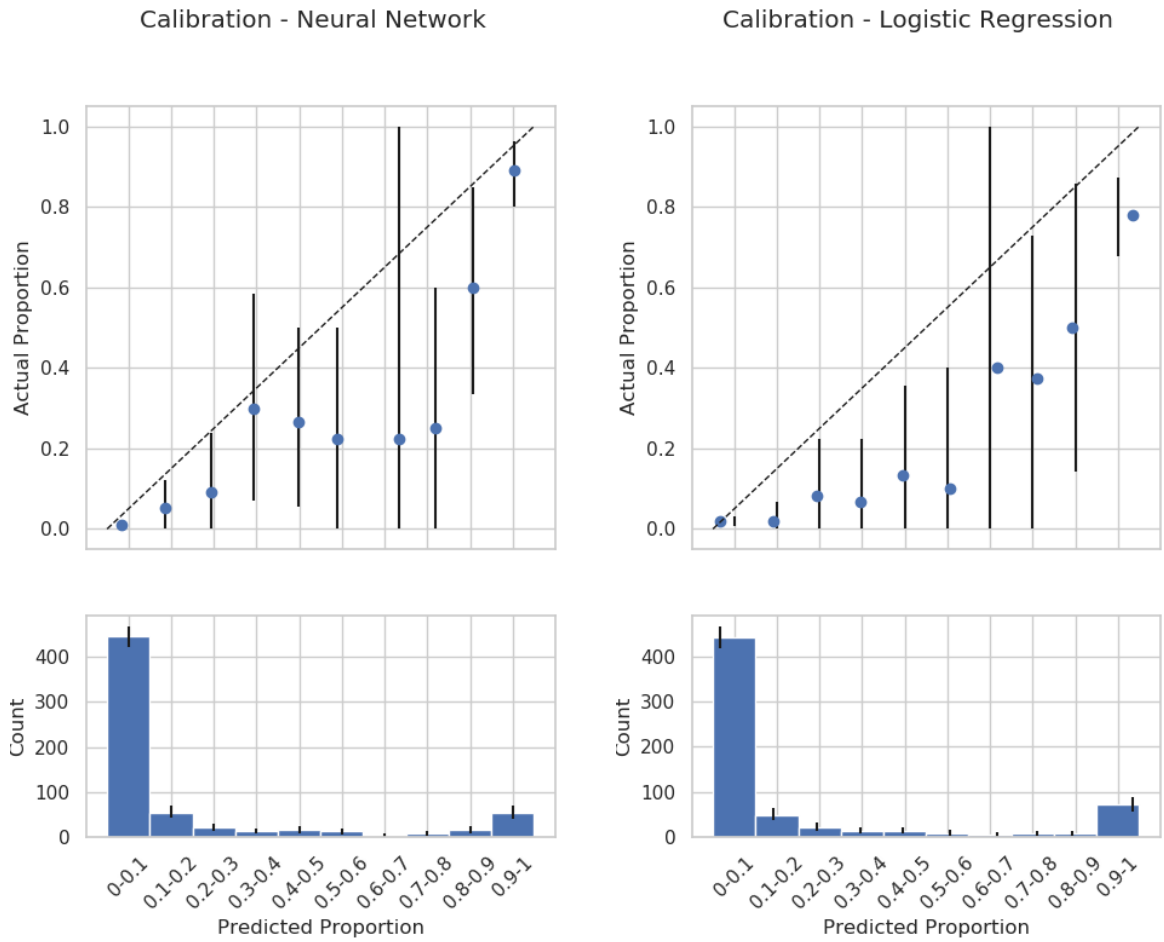
Neural Network		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	68	36

Predicted non-potential organ donors	11	519
Logistic model		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	72	46
Predicted non-potential organ donors	7	509

Supplementary S9: Confusion matrix in test set – All donors – Actual sensitivity 90%

Neural Network		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	72	67
Predicted non-potential organ donors	8	496
Logistic model		
	True potential organ donors	True non-potential organ donors
Predicted potential organ donors	72	144
Predicted non-potential organ donors	8	420

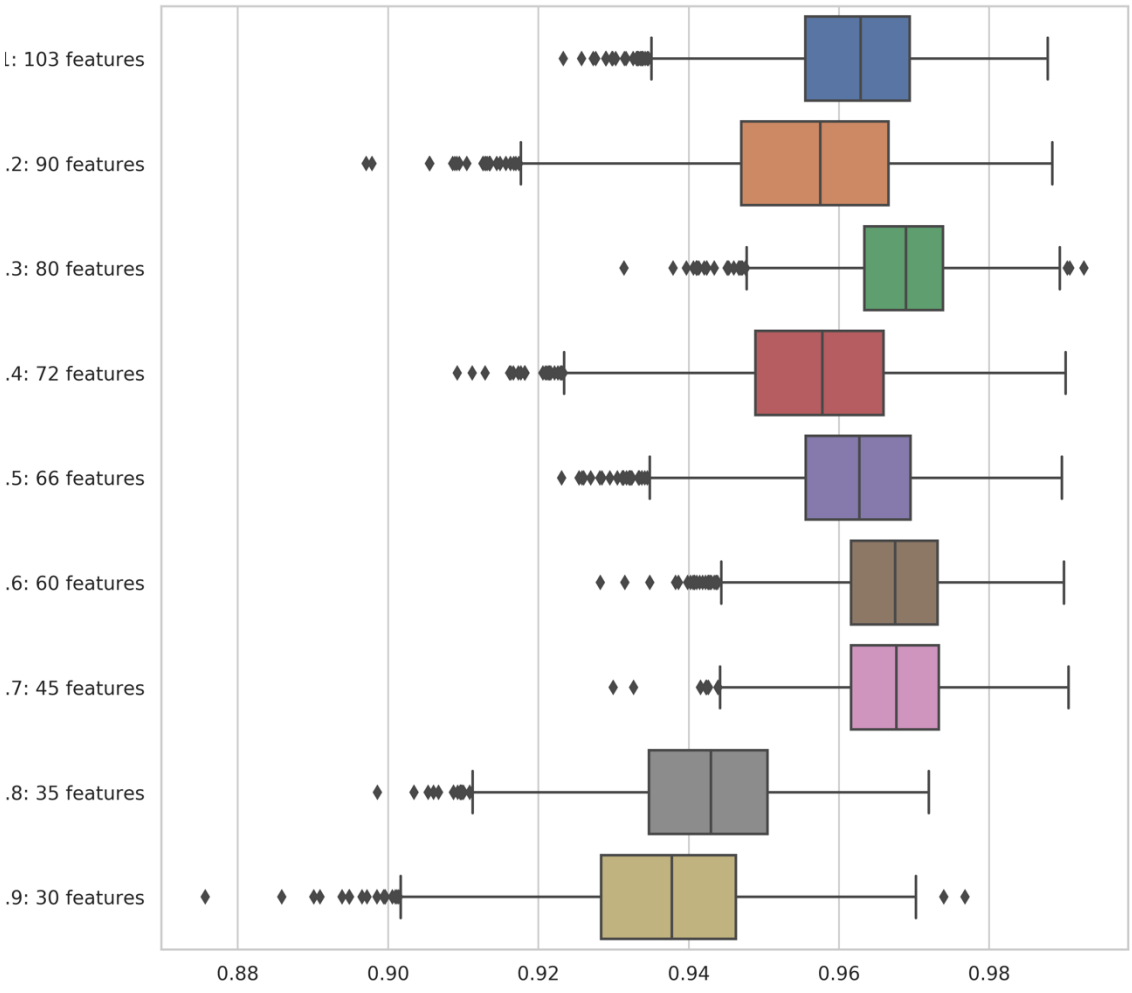
Supplementary S10: Calibration plot



Supplementary S11: False positive error checking

Pt	NN score	LM score	Clinical history
1	0,79	0,45	M85. Pneumonia FiO2 40%. Past medical history of two stroke. Withdrawal of care despite clinical improvement of pneumonia.
2	0,40	1,00	M66. Pneumonia post pneumonectomy for lung neoplasia.
3	0,89	0,59	M70. Hypertensive intracerebral hemorrhage. Death within 24h. Not referred because of "Heavy history of cardiac and pulmonary problems"
4	0,57	0,97	F87. Admission for pulmonary septic shock. Withdrawal of care because of slow evolution
5	0,90	0,27	F68. Aspiration pneumonia. Past medical history of chronic pulmonary obstructive disease dependent on oxygen. Rapid progression to withdrawal of care despite improving labs.
6	0,81	0,60	M72. Pneumonia post craniectomy for intracerebral metastasis.
7	0,55	1,00	M65. Pneumonia post pneumonectomy for lung neoplasia.
8	0,79	1,00	F51. Intracerebral aneurysm rupture. Family refusal before ODO referral.
9	0,26	0,95	H29. Fulminant MS. Withdrawal of care.
10	0,84	0,12	H31. Chondrosarcoma near the basis of the brain here for neurosurgery. Cardiac arrest for 17 min leading to brain ischemia. Withdrawal of care.
11	0,50	0,99	M58. Subarachnoid hemorrhage secondary to a biopsy of a glioblastoma.

Supplementary S12: Sensitivity analysis



Features were removed based on maximum missingness.

7 Validation externe et transfert de connaissances

7.1 Introduction

L'article qui suit est en cours de révision pour être soumis à la revue *Journal of Biomedical Informatics.* Nicolas Sauthier est l'auteur principal de ce manuscrit. Il a conçu et réalisé les analyses, de même qu'écrit l'ensemble du code informatique de ce projet. Rima Bouchakri et Kip Brown ont aidé à l'extraction des données et à la révision du manuscrit, de même que Louis-Antoine Mullie, François-Martin Carrier. Le dernier auteur est le directeur de recherche de ce mémoire, Michaël Chassé. Il a contribué à la conception du projet, à la conceptualisation et la révision des analyses de même qu'à la supervision globale du projet.

7.2 Transfer learning improves the external validation performance of an organ donor detection model in a population with sporadic cases.

Authors: Nicolas Sauthier, MD; Rima Bouchakri, PhD; Kip Brown, PhD; Louis-Antoine Mullie, MD; Brian Potter MD, MD; François-Martin Carrier MD MSc, Michaël Chassé, MD, PhD

7.2.1 Abstract

Organ donation does not meet the increasing demand, and we may be missing 30-60% of potential donors. It relies on manual identification and referral to an organ donor organism, a process that could be improved and partially automated with machine learning. However, potential donors are scattered in various small hospitals, and a prediction model's accuracy decreases in external validation cohorts. To improve a previously developed prediction model, we evaluated transfer learning techniques to reduce this expected drop in accuracy in an external independent cohort. We tested multiple transfer learning approaches to fine-tune our model to the characteristics of this external population. Out-of-the-box results of the model produced an AUROC of 0.82 (0.68-0.95). Retraining the last four layers of a previously fully trained model on a subset of patients improved AUROC to 0.87 (0.73-0.97). Pretraining also reduced the data required to finetune the model to about half. We demonstrated that transfer learning is a realistic approach to fine-tune a predictive model in an external population with a minimal number of cases, mitigating the drop in accuracy.

7.2.2 Introduction

Organ donation has increased slowly over the last 20 years[114] but is not meeting demand. A critical step in the whole process is organ donor identification. Retrospective studies showed that we may miss between 30% and 60% of potential organ donors[3–6]. This problem is present in all hospitals but seems more prevalent in small non-transplant hospitals[29,31]. Strengthened identification and referral to an Organ Donation Organization (ODO) could directly improve the total number of organs transplanted. To try to tackle that challenge, we previously developed a

deep learning model, using only routinely used clinical data (temporal laboratory values, presence of head imaging study, and referral specialty)[126].

Organ donation can happen after a neurological determination of death (NDD)[23] or after death by circulatory death (DCD). Ischemic brain injury secondary to OHCA represents an increasing proportion of potential organ donors and may represent 4% and up to 20% of all cardiac arrest[127]. Patients with OHCA and ROSC can be admitted to various critical care units that care for different populations of patients. For example, they can be admitted to either an intensive care unit (ICU) (managed by an intensivist) or to a coronary intensive care unit (CICU)[128] (managed by a cardiologist). The latter represents an ideal external validation cohort [90,91] for our previously published model[126] as the pathologies, medical and paramedical staff are different and independent. However, accuracy frequently decreases in the external validation of prognostic models[91], reducing the clinical application. Retraining the model in a small cohort of patients with rare events is often non-feasible.

Transfer learning[75,76,129] is a technique that allows a reduction of data required to train a model. It is a two-step process with the pretraining of a model in a usually large, related dataset with a fine-tuning in a generally smaller primary dataset. Previous studies on transfer learning in tabular healthcare data showed promise on ECG[130] and EHR textual data[131] using a pretrained feature extraction approach such as an autoencoder (AE)[58]. An AE is an unsupervised subtype of neural networks where the input and output are identical, with a smaller layer in the middle. This layer, called latent space, forces the network to learn a way to represent the input data in a different and usually smaller form with a minimal loss. Wardi and al.[132] used a transfer of learning on the classifier to predict sepsis shock with emergency room clinical data. They fine-tuned their network in an external validation cohort from another hospital and demonstrated an improvement of AUROC with a fine-tuning cohort of under a thousand patients.

The main objective of this study is to investigate multiple strategies of transfer learning to improve the performance of a previously trained predictive model in a small external and independent validation cohort. Since external validation data is costly to acquire and the event of interest is rare, our secondary objective is to quantify the amount of data required to fine-tune our model.

7.2.3 Method

7.2.3.1 Population

The dataset was derived using the electronic health record (EHR) from ICU and CICU stays from Jan 1st, 2012, until Dec 31st, 2019, in the CHUM (Montreal University Hospital Center, Montreal). The study and data extraction were approved by the CHUM Research Ethics Board and individual patient consent was waived given the low risk and retrospective nature of the study[126].

We included data from the last ICU or CICU stay of all patients with a visit duration longer than 16 h. Patients that went to both ICU and CICU during a single hospital stay were categorized as ICU to keep the validation cohort as independent as possible. We defined our organ donor population as (1) actual organ donors either local (admitted to the CHUM for a condition that eventually evolved to death and organ retrieval) or transferred (neurological death diagnosis made in another peripheral hospital and transferred to the CHUM for organ retrieval) or (2) patients referred to the ODO but deemed ineligible from transplant (next of care refusal, medical contraindications detected in the workup, etc.) and (3) non-referred potential organ donors as defined by a continuous death audit made by the ODO. The death audit defined a potential organ donor as a patient either with a severe neurological condition mechanically ventilated, who died within 24 h of the end of care, or without a severe neurological condition but who died within 3 h. Patients with any recent, active, or metastatic cancer, disseminated infection, or multi-organ failure were excluded.

7.2.3.2 Dataset

We collected routinely used clinical data from the EHR of the patients. The primary type of data was temporal laboratory data. Only frequent laboratory data (sampled in at least 50% of the patients) were kept. The goal was to exclude rare labs that would be too specific to the potential organ donor population. We first used the last value carried forward (LVCF), i.e. if a patient had a laboratory value measured in the first 8 h block, it's repeated in all subsequent 8 h blocks until updated by a new laboratory result. For values still missing after LVCF, we imputed a normal value randomly sampled from a Gaussian distribution in the normal range of each laboratory analysis.

This decision was made to reflect the fact that if a physician has no clinical reason to order a laboratory test, its pre-test probability of being normal is thought to be high.

The presence or absence of cerebral imaging and the specialty responsible for the patient (neurosurgery, internal medicine, cardiology, etc.) were the only static values used. Usual demographic data (age, sex at birth, weight, etc.) weren't used since they aren't a contraindication to ODO referral. Using them may induce bias in the model toward usual clinical practice rather than actual clinical patterns.

Class imbalance in the ICU dataset was addressed with subsampling and oversampling. 85% of randomly selected non-donors in the ICU population were used to train the autoencoder (AE, see below). The remaining 15% of the ICU population trained the classifier. During training, proportionally more weight was put on the minority class. That means that if non-donors outnumber donors 100 to 1, the training weights of donors will be 100 and non-donors 1.

The CICU dataset was randomly separated into a fine-tune and a test set with 75% for fine-tuning and 25% for the test. Confidence intervals were calculated using 100 cross-validations.

To quantify the amount of data required to fine-tune our model, multiple proportions of fine-tuning to test were explored. The minimum of positive cases (i.e. potential organ donors) in the fine-tune cohort was one. That number was increased stepwise from 1 to a maximum of 75% of all positive cases (i.e. 75% of all potential organ donors). At each step, the prevalence of positive cases to negative cases was kept constant. That means that the first fine-tune group was 1 donor and 129 non-donors, the second train group was 2 donors and 258 non-donors, and so on. Confidence intervals were calculated using 100 cross-validations.

7.2.3.3 Deep learning model and transfer learning

The deep learning model is described elsewhere[79,126], but in summary, two steps were used, and TRIPOD methodology was followed. In the first step, we trained a deep learning model called a convolutional neural network, based on a well-published architecture [55,112]. We then trained an unsupervised autoencoder[58,60] to combine its transfer learning and anomaly detection capabilities[58].

Regarding transfer learning from ICU to CICU, we combined two frequently used approaches [75,76,129,133]. The first is to extract features from our temporal laboratory data with our AE with non-supervised training (Figure I A). This encoding process is frozen, which means won't continue learning, and is transferred to the classifier (Figure I B). The second transfer learning is a training of the classifier. Our classifier (Figure I B) is a dense 4 layers neural network with a progressively decreasing size. The layers have 50 neurons (3350 weights), 33 neurons (1683 weights), 16 neurons (544 weights), and finally 1 neuron (17 weights). Learning of each layer can be blocked or allowed. Additionally, the initials weights of each layer can be initialized with pretrained values (transferred from ICU learning) or as random values to start learning from scratch. Two patterns of were tested: either all 4 were trainable (100% of the 5561 weights) or only the last two were trainable (561 of the 5561 or 10% of weights). Since fewer neurons (and therefore fewer weights) are available to learn, the learning capacity decreases. Negative control was implemented by having no trainable layers (0% of weights), thereby blocking all learning. For each fine-tuning pattern, the trainable layers were initialized either as random or as pretrained from the ICU to assess the impact of pretraining.

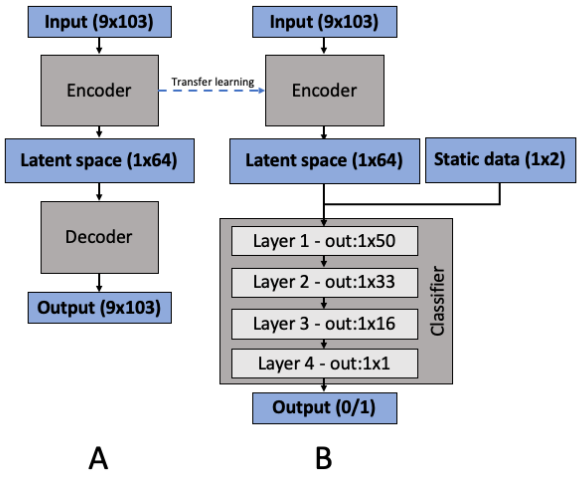


Figure I. Neural Network architecture. A) Autoencoder. B) Classifier

7.2.4 Results

7.2.4.1 Population characteristics

Baseline characteristics of the population are reported in Table I. CICU patients were older, had lower mortality and had a higher proportion of cardiac arrest as the principal diagnosis. Subsampling wasn't used on the CICU population, making potential organ donors a much rarer event than in the ICU population. There were no local or transferred donors and few referred potential donors because detected potential donors were transferred to the ICU, according to hospital protocols.

Table I. Population' characteristics

	AE Train N= 16 213	ICU N=3 250	CICU N=4 669
Sex F N (%)	5871 (36,2%)	1 248 (38,4%)	1 676 (35,9%)
Age mean (std)	67,2 (14,5)	66,3 (14,8)	71,4 (14,0)
Length of stay in hours median [IQR]	50,7 [95,2]	50,6 [90,6]	55,8 [60,6]
Morality N (%)	1 573 (9,7%)	682 (21,0%)	244 (5,2%)
Cardiac arrest as principal diagnostic N (%)	52 (0,3%)	15 (0,4%)	105 (2,2%)
Donors' subtypes n (%)			
Non-donor	16 213 (100%)	2 853 (87,8%)	4 633 (99,2%)
Transferred	-	189 (5,8%)	-
Local	-	107 (3,3%)	-
Referred	-	65 (2,0%)	7 (0,1%)
Not referred	-	36 (1,1%)	29 (0,6%)
Principals' reasons for admission n (%)			
Ischemic heart diseases	3 855 (22,9%)	637 (19,0%)	2 412 (49,9%)

	AE Train N= 16 213	ICU N=3 250	CICU N=4 669
Other forms of heart disease	1 833 (10,9%)	350 (10,4%)	1 584 (32,8%)
Cerebrovascular diseases	1 033 (6,1%)	300 (8,9%)	22 (0,5%)
Diseases of the arteries/arterioles/capillaries	677 (4,0%)	115 (3,4%)	39 (0,8%)
Total Diseases of the circulatory system	7 907 (47,0%)	1 494 (44,5%)	4 122 (85,3%)
Neoplasms	2 729 (16,2%)	500 (15,0%)	73 (1,5%)
Diseases of the digestive system	1 649 (9,8%)	317 (9,4%)	46 (1,0%)
Consequences of external causes	1 299 (7,7%)	248 (7,4%)	133 (2,8%)
Diseases of the respiratory system	1 063 (6,3%)	212 (6,3%)	91 (1,9%)

AE: Auto-Encoder.

7.2.4.2 Model results

We demonstrated that transfer learning from ICU to CICU worked as an out-of-the-box solution. As we can see in Figure II, and Table II, using only the pretrained “knowledge” learned by the model on ICU patients AUROC was 0.820 (CI95% 0.682-0.948), compared to 0.545 (CI95% 0.329-0.765) for the random weights with no further training

When enabling training on the two upper smaller layers, representing only 10% of the total weights of the neural network, we observed no improvement in training accuracy compared to complete transfer learning with AUROC of 0.820 CI95% 0.682-0.948 and 0.820 CI95% 0.682-0.948 (blue and orange curve, central figure, Figure II).

When enabling training on all layers, the pretrained network started training with a higher AUROC than the random initialized one (AUROC 0.809 CI95% 0.643-0.859 vs 0.610 CI95% 0.237-0.769, see blue and orange curve, left figure, Figure II). The pretrained network needed less data to be trained. To exceed a value of 0.85 in AUROC, half of the data was required (9 positive cases vs. 17 positive cases, see Figure II). Both reached a plateau around a training dataset size of 50%

of the external validation dataset as training. Both pretrained and random initialized approaches, with all layers unfrozen and available for training, achieved a similarly high accuracy (Table II). Because of the small number of permutations possible, all results have a wide confidence interval.

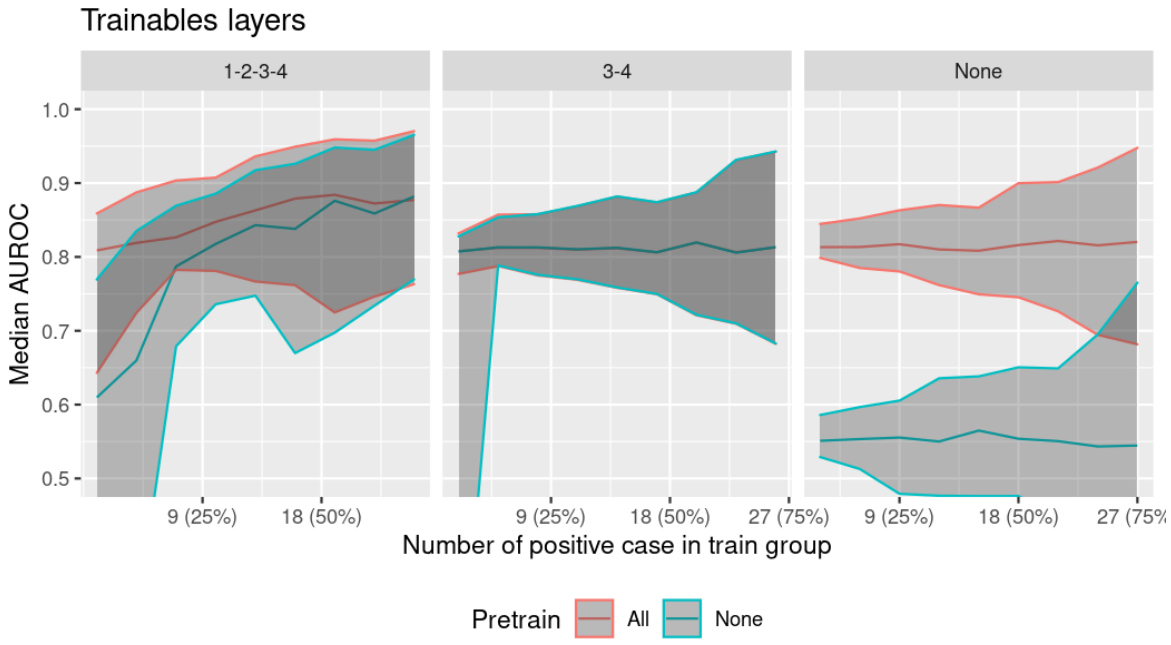


Figure II. AUROC with an increasing proportion of training data, with different patterns of trainable layers, with 95% CI

Table II. Model results at 75%/25% train to test repartition.

Trainable layers	Pretrained layers	AUROC	Sensitivity	Specificity
1-2-3-4	None (all randomized)	0.896 (0.725-0.968)	0.667 (0.333-0.889)	0.927 (0.847-0.962)
1-2-3-4	1-2-3-4	0.874 (0.731-0.974)	0.667 (0.333-0.889)	0.922 (0.833-0.957)
3-4	1-2 (3-4 randomized)	0.820 (0.682-0.948)	0.667 (0.444-0.947	0.907 (0.883-0.922)
3-4	1-2-3-4	0.820 (0.681-0.948)	0.667 (0.444-0.947	0.905 (0.883-0.922)
None	None (all randomized)	0.545 (0.329-0.765)	0.000 (0.000-0.000)	1.000 (1.000-1.000)
None	1-2-3-4	0.820 (0.682-0.948)	0.111 (0.000-0.333)	0.990 (0.985-0.994)

Values are shown as median with a 95% confidence interval calculation on 100 cross-validations.

Layers number refer to Figure I

7.2.5 Discussion

We proposed a promising method to improve a potential organ donor identification model in external small cohorts using very few cases to fine-tune the network. Allowing all four layers to be trainable was the best way to fine-tune our network to the CICU specifics. Allowing only the last two layers to be trainable provided inferior results. We also observed that pretraining the classifier reduced the number of cases required to finetune the network by about half, while maintaining the predictive accuracy of the model.

Previous work on transfer learning in tabular healthcare data worked on ECG [130] and EHR textual data[131]. Both used a pretrained feature extraction approach, such as an AE or a language representation model. This allowed them to keep good precision in a classification task in an external cohort of reduced size. However, they used a population in the thousands with hundreds of positive cases to fine-tune their classification model. Wardi and al.[132] used a transfer of learning approach on the classifier to predict septic shock with emergency room clinical data. They fine-tuned their network in an external validation cohort from another hospital, which had more than three hundred cases, and demonstrated an improvement of AUROC with a fine-tuning. Our approach allowed a sizeable pretrained autoencoder and a smaller pretrained classifier to use less than fifty cases for fine-tuning. This approach demonstrated an increase in prediction capacities in a population with very rare cases, a frequent situation in healthcare. Additionally, our method used real-life complex multimodal data while most studies focused either on temporal or tabular data.

Our approach has some limitations. First, the type of potential organ donors in CICU differs from the missed potential organ donors in ICU and peripheral hospitals. Most organ donors died from a stroke[14], a pathology almost non-existent in CICU. Secondly, our external validation cohort is small and contained only 36 missed potential organ donors, explaining the wide confidence interval. However, it reflects the reality and difficulty of clinical practice. Potential organ donors are more frequent in non-transplant hospitals and smaller hospitals[31], meaning that fine-tuning the external validation cohort is probably small. Finally, our approach gave a relatively low sensitivity. However, every potential donor detected could generate donations that would

otherwise have been lost. A sensitivity of almost 67% in a cohort where most patients were not identified, in this particularly difficult clinical pattern is meaningful, but further work will be required to improve accuracy.

Future work should focus on improving the accuracy of the trained models while maintaining accuracy in external cohorts. The architecture we chose both for initial model development as well as transfer and fine tuning could also be improved. However, those initial results using "real-world" data with minimal preprocessing and our success to transfer this learning into a different population of patients if promising. This strategy could be further tested and improved in other external datasets for similar outcomes, or for other medical conditions using either local datasets or online available clinical datasets[28,29]. Finally, privacy is of great concern when training a model across data from multiple sites with solutions being developed for this issue such as federated learning[134]. The transfer of only trained weights, without gradients, could be a secure approach [135] but more investigations are needed to ensure its safety.

In conclusion, we validated our model in an external and independent cohort of potential organ donors. We increased the accuracy of the out-of-the-box model using a two-step transfer learning approach to fine-tune our model to a different population. These results support the feasibility of fine-tuning pretrained models on small independent cohorts using transfer learning approaches.

7.2.6 References

(Uniformisée avec le mémoire, voir chapitre 0)

8 Discussion

8.1 Retour sur les objectifs du projet

L'objectif principal de ce projet était d'investiguer la capacité d'un modèle temporel de RNA à détecter automatiquement de potentiels donneurs d'organes. Ce modèle devait être développé en utilisant des données rétrospectives cliniques de routine, la majorité avec une composante temporelle. Une fois développé, ce modèle devait être comparé à un modèle simple fréquentiste linéaire. Les trois hypothèses principales étaient que le modèle temporel serait fonctionnel, qu'il serait supérieur au modèle fréquentiste linéaire et qu'il est transférable dans une cohorte externe à l'aide de techniques de peaufinage.

En prévision d'une potentielle application clinique prospective, nous voulions aussi démontrer une capacité de détection précoce, dans une fenêtre de 48h précédant le congé des soins intensifs (ou le décès aux soins intensifs le cas échéant). Notre population de donneurs potentiels étant un agrégat de donneur (confirmés, refusés et non référés), nous voulions aussi déterminer la capacité de prédiction dans chacune des sous-populations de donneurs. Le modèle devait aussi être évalué en termes de calibration, de sensibilité, de spécificité et de taux de faux positifs et de faux négatifs.

Toujours dans un but d'application clinique, nous voulions aussi valider le modèle dans une population externe, c'est à dire l'unité de soins intensifs coronariens, laquelle est cliniquement très distincte des soins intensifs. En prévoyant une baisse de capacité prédictive, nous voulions appliquer et évaluer une technique propre à l'apprentissage machine qu'est le transfert de connaissance. Puisqu'il existe plusieurs approches, nous voulions déterminer si une était supérieure dans notre cas, tout en quantifiant cette supériorité.

8.2 Résultats principaux du projet

Les résultats de ce projet supportent la capacité d'une approche novatrice, basée uniquement sur des données cliniques de routines, utilisant un réseau de neurones pour aider à détecter les potentiels donneurs d'organes. De plus, le modèle de RNA utilisant la progression temporelle des

laboratoires a démontré une supériorité marginale mais significative (AUROC 0.97 vs 0.94, p=0.014) sur un modèle logistique plus simple et non temporel. Cette supériorité se confirmait dans les sous-groupes d'intérêt et dans une approche prospective simulée. Ce modèle a été validé à la fois dans une cohorte interne séparée et dans une cohorte externe cliniquement distincte.

L'approche initiale utilisée était une approche non supervisée, utilisant uniquement l'erreur de reconstitution d'un autoencodeur. Ce projet suggère que, bien que cette approche soit fonctionnelle et hautement intéressante, l'ajout de couches pour en faire une approche supervisée améliore grandement sa capacité prédictive (AUROC de 0.86 vs 0.97). Ce gain global était cliniquement significatif avec une augmentation marquée de la spécificité (73% vs 93%), un élément très important dans un cas d'événement rare. Toutefois, la sensibilité du modèle s'en trouvait un peu réduite (88% vs 84%).

Ensuite, les résultats de notre projet ont suggéré que les sources principales d'erreur du modèle, chez les patients décédés aux soins intensifs, étaient la présence d'une néoplasie ou d'un âge avancé. Dans l'analyse des erreurs du modèle, plus de la moitié des patients étaient hospitalisés avec une néoplasie active ou récidivante, les rendant non éligible au don d'organes. De plus, deux patients de plus de 85 ans ont eu une décision de soins de confort vu que l'évolution clinique globale était lente, bien que positive. Ce type d'erreur pouvait être attendu sachant que ni l'âge ni les antécédents n'étaient fournis au modèle. Ces deux types d'antécédents peuvent expliquer avec un passage en soins de confort, malgré une amélioration clinique lente. Ceci était cohérent avec notre hypothèse que le modèle identifierait des patients avec des analyses de laboratoires normales ou en amélioration.

Finalement, ce projet visait aussi à évaluer l'utilisation du transfert de connaissance pour peaufiner le modèle et améliorer les résultats dans une cohorte externe. Comme attendu, la capacité prédictive du modèle était réduite dans la cohorte de validation externe (AUROC 0.82 vs 0.97). Toutefois, le projet a permis d'identifier des stratégies de ré-entrainement, permettant de peaufiner le réseau et d'améliorer l'AUROC (0.896). Cela supporte la possibilité de peaufiner un modèle prédictif aux réalités spécifiques d'une cohorte externe.

8.3 Forces du modèle et du projet

Christodoulou et coll. ont relevé que 68% des publications[43] de modèles en apprentissage machine avaient un risque de biais, particulièrement au niveau de la validation. L'approche utilisée dans ce mémoire a suivi une méthodologie rigoureuse de développement et validation de modèle de prédiction[47], une qualité rare dans les articles de modèle diagnostique utilisant l'apprentissage machine[48]. En plus de valider le modèle dans une population interne séparée et dans une population externe, la capacité prédictive du modèle a été évaluée dans des analyses de sous-groupes et dans des analyses de sensibilités, ce qui augmente la validité des conclusions.

Deuxièmement, le projet a démontré la fonctionnalité d'une approche en deux parties avec un autoencodeur permettant la transformation de données temporelles en un format linéaire. Ce format transformé a pu ensuite être combiné avec d'autres données statiques pour produire une prédiction unique combinant des informations multimodales complexes. Outre les performances de cette architecture, elle a été choisie pour son grand potentiel de flexibilité qui permettrait aisément d'ajouter d'autres types de données (imagerie cérébrale, données textuelles, signes vitaux, etc.). Sachant que ce mémoire se voulait une première étape d'un projet de plus grande envergure, cette flexibilité était un avantage notable. Par ailleurs, bien que cette approche ait été développée pour le don d'organes, elle est potentiellement utilisable pour le dépistage d'autres maladies rares aux soins intensifs, en ré-entrainant simplement le classificateur pour prédire la pathologie d'intérêt.

Troisièmement, le modèle a été développé dans une base de données clinique massive, comptant près de 20 000 patients. Il n'existait, avant ce projet, aucune base de données clinique de cette envergure étudiant les potentiels donneurs d'organes. Il existait des bases de données cliniques de soins intensifs plus larges. Toutefois, les potentiels donneurs d'organes n'y sont pas identifiés et le fait qu'elles provenaient de systèmes de santé différents (majoritairement américains) les rendait moins spécifiques à la population d'intérêt de donneurs potentiels québécois ou canadiens. L'utilisation d'un grand nombre de patients a donc permis d'avoir un grand nombre de donneurs potentiels, ce qui a permis le développement d'un modèle de RNA complexe tout en minimisant le risque de surentrainement.

Finalement, le modèle a démontré une résilience dans les populations de donneurs qui n'ont pas été référés à TQ. C'était une population minoritaire et difficile à étudier du fait de sa rareté. Toutefois, c'était une population importante pour ce projet, sachant qu'elle s'approchait de la population d'intérêt de donneurs potentiels manqués.

8.4 Limitations générales du projet

Ce projet avait plusieurs limites intrinsèques à sa conception. Tout d'abord, il s'agit d'un projet rétrospectif, ce qui est toutefois nécessaire à l'entrainement d'un modèle d'apprentissage machine.

Les données utilisées étaient de haute qualité et ont été choisies pour la quasi-absence de manipulations dans leur processus, afin de minimiser la présence d'inévitables erreurs humaines. Toutefois, l'étiquette provenant de l'audit de décès provenait d'une revue humaine de dossier. Il est clair que la sensibilité de cet audit de décès n'est pas de 100% puisqu'au moins un cas a été détecté par les modèles, lequel n'avait pas été correctement étiqueté par l'audit de décès. Cependant, cela nous semblait un risque de biais mineur du fait de la faible proportion de ces quelques cas par rapport au grand nombre de non-donneurs correctement étiquetés. De plus, la tendance du nombre de donneurs du CHUM et de TQ (voir Fig. 7) étant similaire, il semblait peu probable qu'une grande quantité de donneurs potentiels aient été mal étiquetés.

Une autre limitation était que les données provenaient d'un centre de transplantation quaternaire. Le groupe de « donneurs potentiels » était un agrégat de sous-types de donneurs (75% de donneurs confirmés, 15% de refusés et seulement 10% de donneurs potentiels non référés). Une révision manuelle des dossiers des donneurs non référés démontrait qu'ils n'ont jamais été réellement manqués par le clinicien. Bien qu'ils auraient dû être référés (amélioration de la qualité de l'acte et réponse aux questions de la familles[26]), ils ne l'ont pas été, le plus souvent par refus familial précoce. Par conséquent, la population utilisée comme « donneur potentiel » était différente des 30 à 60% de donneurs manqués que le projet veut aider à identifier. Toutefois, il aurait été impossible de créer un modèle de cette complexité avec le petit nombre de donneurs non référés. De plus, bien qu'il existait une possible différence entre un donneur manqué et un donneur confirmé, le modèle a pu apprendre des éléments pertinents de chaque sous-

population. Par ailleurs, le modèle est resté performant dans la sous-population de donneurs non référés, laquelle ressemblait probablement le plus aux donneurs manqués. Finalement, si, lors de la validation prospective à venir, le modèle se révélait moins performant sur les « vrais » donneurs manqués, l'approche de transfert de connaissance et de peaufinage pourrait permettre d'ajuster le modèle afin de le rendre plus spécifique.

Troisièmement, l'architecture finale du modèle comprenait une étape d'encodage, suivie d'une étape de classification binaire. Les modèles d'apprentissage machine ont la réputation d'être une « boite noire » qui fournit une réponse sans qu'il soit possible de comprendre les éléments justifiant ce choix. Il existe plusieurs méthodes permettant de mieux comprendre quels *features* (c.-à-d. quels laboratoires ou données statiques) ont été importants pour la prédiction[136]. Deux méthodes parmi les plus faciles d'utilisations sont SHAP[137] et LIME[138.] Toutefois, l'architecture en deux étapes de notre approche menait à un découplage, rendant impossible l'utilisation de la quasi-totalité des méthodes. Cette contrainte semblait peu importante étant donné que ces méthodes d'analyses donnent des conclusions difficiles à interpréter lorsqu'elles sont utilisées avec des données temporelles ou une combinaison de types de données. Ainsi, même si elles avaient été utilisables, l'utilité de leurs conclusions aurait probablement été mineure. Par ailleurs, notre approche minimisait l'impact de cette limitation avec une analyse de sensibilité, de même qu'avec une vérification d'erreur. De plus, l'architecture en deux étapes a permis une grande flexibilité de développement et d'analyse du modèle sans devoir être réentrainée dans son ensemble à chaque fois.

Quatrièmement, bien que le modèle ait démontré une performance encourageante de prédiction d'événement rare, il présentait un taux de faux positif relativement élevé. Une simulation théorique peut être faite avec les performances obtenues du modèle (incidence annuelle de 2%, population de 3 000 patients par année, sensibilité de 84% et spécificité de 93%, voir chapitre 6). Ainsi, durant une année moyenne, le modèle détecterait 50 donneurs potentiels et en manquerait 10. Sur la même période, le modèle générerait 176 faux positifs pour un total de 226 patients sélectionnées annuellement par le modèle. Les résultats du modèle montraient donc que le taux de faux positif serait élevé (78%). Toutefois, le modèle avait un potentiel important d'amélioration et de réduction de faux positifs en ajoutant des données telles que la

présence de néoplasie. De plus, même à ce stade de développement préliminaire, la présélection faite par le modèle est appréciable et réduit le nombre de patients de 3 000 à 226, un nombre beaucoup plus réaliste pour une révision humaine.

Finalement, le but était d'utiliser l'évolution temporelle des laboratoires. Pour minimiser l'imputation et maximiser la durée d'évolution, un minimum de 16h d'évolution et une période d'observation de 72h ont été choisis. Il est vrai qu'une partie des donneurs manqués le sont dans les premières heures, à l'urgence. Toutefois, les recommandations actuelles sont d'admettre pour une observation de 72h tous les patients se présentant à l'urgence avec une atteinte neurologique sévère[124,125]. Cela permet de diminuer la pression de décision sur les urgences et laisse un certain recul pour évaluer les options, parmi lesquelles se trouve le don d'organe. Notre modèle était donc cohérent avec l'approche clinique recommandée.

8.5 Développements futurs

8.5.1 Amélioration du modèle

L'architecture choisie l'a été pour des raisons logistiques plus que pour une supériorité démontrée du type d'algorithme d'apprentissage machine. Du fait de l'absence de démarcation évidente du modèle temporel sur une version non temporelle (modèle logistique), il serait pertinent de comparer d'autres approches d'apprentissage machine comme les arbres de décision ou XGBoost. La combinaison multimodale pourrait être ensuite faite par une méthode d'ensemble[38], le cas échéant.

Deuxièmement, l'approche des *features* choisie était d'être le plus objectif possible. L'analyse de sensibilité a permis de constater que retirer la majorité des laboratoires n'avait que peu d'impact sur la performance globale. Une réduction des laboratoires choisis par un choix subjectif basé sur la clinique permettrait de simplifier le modèle et de le rendre plus généralisable.

Ensuite, le modèle pourrait être amélioré par l'ajout de données. Augmenter la quantité de données, surtout au niveau de l'autoencodeur, permettrait de développer un modèle plus complexe. L'utilisation des bases de données de recherche de soins intensifs[69,70,83] pourrait permettre d'entrainer un autoencodeur plus complexe, lequel pourrait améliorer la performance.

De plus, ce modèle pourrait éventuellement être partagé publiquement de manière pré-entrainée, à l'image de ce qui est fait avec les architectures convolutives pré-entrainée sur ImageNet[52,112].

Finalement, bien que d'augmenter le nombre de patients puisse être utile, il est probable que d'augmenter la diversité de données serait plus efficace pour améliorer la spécificité du modèle. En effet, comme il est possible de constater, en regardant les faux positifs, l'ajout de contrindications au don (néoplasie, choc profond, etc.) permettrait d'en réduire le nombre et d'améliorer l'applicabilité du modèle. L'architecture utilisée dans ce projet a démontré la possibilité de combiner des éléments temporels et statiques tout en gardant la possibilité d'ajouter aisément d'autres modalités. Les antécédents néoplasiques pourraient être extraits par une analyse textuelle des résultats des rapports de pathologies ou d'imageries. Par ailleurs, deux éléments d'importance, l'atteinte cérébrale majeure et la ventilation mécanique, sont absents du modèle. Les ajouter au modèle pourrait grandement en améliorer les performances. Une analyse par réseau convolutif des examens d'imageries cérébrales permettrait d'ajouter l'atteinte neurologiques. La présence de ventilation mécanique pourrait être aisément ajoutée avec un réseau convolutif basé sur les rayons X pulmonaires, un examen quasi uniformément fait chez les patients intubés aux soins intensifs.

8.5.2 Évaluation prospective unicentrique

Parallèlement à l'amélioration du modèle, une implémentation prospective du modèle, dans un cadre de projet de recherche, serait particulièrement pertinente. D'une part, cela permettrait une validation prospective du modèle et, d'autre part, cela permettrait de référer dès maintenant des donneurs potentiels qui ne l'auraient pas été sinon. L'évaluation prospective permettrait aussi une rétroaction constante pour améliorer le modèle de manière plus dynamique. Cette approche au développement logiciel, appelée Agile, est bien décrite, mais encore peu utilisée dans le domaine médical[139,140].

8.5.3 Validation externe multicentrique

La validation faite à l'unité coronarienne était une première étape de validation externe. La suivante serait de valider le modèle de manière rétrospective dans différentes unités de soins intensifs. Au-delà d'être une étape fondamentale à la validation du modèle avant son utilisation prospective multicentrique, cela permettrait de mettre en place et de tester la structure informatique nécessaire à l'extraction des données. S'il était impossible d'extraire les données nécessaires à une évaluation rétrospective dans un centre, il serait bien évidemment impossible de le faire en prospectif dans ce centre. Bien que l'approche la plus simple serait de centraliser les laboratoires dans un centre de coordination, c'est aussi l'approche la plus complexe au niveau logistique et légal. Une approche alternative serait celle utilisée pour la validation externe à l'unité coronarienne, c'est-à-dire peaufiner un modèle préentrainé. Les données propriétaires resteraient locales et des communications sécurisées pourraient être mises en place pour les résultats du modèle.

Finalement, un autre axe de développement multicentrique serait de modifier le modèle pour une utilisation uniquement rétrospective en validation de la qualité de l'acte. L'audit de décès utilisé dans ce projet est fait par TQ dans un but qui dépassait la recherche. Ce type d'audit est fait dans plusieurs hôpitaux de la province, mais le temps requis pour l'analyse manuelle des dossiers est conséquent. Un modèle, spécifiquement entrainé sur les patients avec un décès aux soins intensifs, permettrait possiblement d'accélérer grandement ce processus.

8.5.4 Limitations logistiques, déontologiques et éthiques

Bien qu'il reste plusieurs étapes avant que le modèle ne puisse être utilisé de manière routinière (amélioration du modèle, interconnexion des bases de données cliniques, etc.), il semble important de prévoir dès maintenant la manière dont pourrait être implanté cet outil tout en réfléchissant aux potentielles barrières logistiques et éthiques.

D'un point de vue logistique, l'approche qui semblerait la plus efficace serait d'avoir un algorithme qui classifie chaque patient dans une unité de soins intensifs à intervalles réguliers, une fois par jour par exemple. Une approche possible serait que chaque patient classifié comme

potentiel donneur d'organe, dans l'ensemble des hôpitaux que surveille le modèle, soit ensuite référé à une personne-ressource en transplant d'organe. Cette personne recevrait une liste de patients, sur laquelle un tri manuel devrait être fait pour exclure les patients avec des contrindications évidentes. Par la suite, les médecins traitant les différents patients pourraient être contactés directement pour confirmer la situation et suggérer, le cas échéant, une référence formelle à TQ. La limitation logistique principale est l'accessibilité aux données d'un maximum de centre hospitalier, en particulier les centres non-transplanteurs, où le taux de conversion est plus faible.

D'un point de vue réglementaire, dans le cadre d'un projet de recherche, l'accès aux données médicales requiert l'approbation par un comité d'éthique à la recherche (article 2.3, énoncé de politique des trois conseils pour l'éthique de la recherche avec des êtres humains[104]). Il est prévu que le comité d'éthique peut autoriser, pour une utilisation secondaire (càd pour de la recherche), l'accès aux données cliniques identificatoires (article 5.5A) et non identificatoires (article 3.7A), sans consentement des patients. Cela n'est possible que si les données identificatoires sont essentielles à la recherche, que le risque prévu de conséquences négatives sur les patients est faible, que les données sont protégées adéquatement et qu'il est pratiquement impossible de solliciter le consentement. À noter aussi qu'un couplage de données requiert aussi une autorisation du comité d'éthique, vu que cela pourrait créer des renseignements identificatoires. Hors du contexte de la recherche et de la qualité de l'acte, le secret professionnel protège les données médicales du patient, ce qui pourrait limiter l'accès. Toutefois, si seul le système avait accès aux données identificatoires et confidentielles, et ne faisait que traiter les données pour faire une recommandation au médecin traitant, il est possible que le code de déontologie médicale de même que le secret médical soit respecté. Avant une implémentation prospective clinique, il serait indispensable d'obtenir l'avis de juristes spécialisés en droit médical pour s'assurer de la légalité du système.

Finalement, avant un déploiement à grande échelle, l'aspect éthique devrait être abordé. Un système surveillant l'ensemble des patients des soins intensifs québécois (voir même canadiens) à la recherche de potentiels donneurs d'organes pourrait être considéré comme

opportuniste. Toutefois, ce projet semble respecter les principes fondamentaux de la bioéthique que sont la justice, la non-malfaisance, la bienfaisance et l'autonomie.

Tout d'abord, au niveau de la justice, le but du projet est justement d'offrir à l'ensemble des patients et leurs familles de discuter de la possibilité du don d'organes. Chaque potentiel donneur manqué est un patient qui n'a pas eu accès à cette discussion.

Le principe de non-malfaisance semble respecté sachant que le but du projet n'est pas de pousser les familles à un don d'organes, mais bien de s'assurer que la possibilité ait été abordée.

Au niveau de la bienfaisance, le don d'organesest un processus souvent apprécié des familles, qui pourrait même réduire le deuil pathologique[141]. Le projet semble tout particulièrement cohérent avec ce principe.

Finalement, le principe d'autonomie semble lui aussi respecté. Le but n'est pas de forcer un don d'organe, mais d'offrir aux cliniciens, aux familles et aux patients l'expertise des personnes-ressources de TQ pour avoir le consentement le mieux éclairé possible devant ce choix important. Si le projet est déployé cliniquement, il me semble important que les familles approchées soient questionnées à distance de l'événement pour évaluer l'impact de cette approche novatrice sur le processus du don.

9 Conclusion

La transplantation d'organes est une procédure cruciale pour la survie de nombreux patients. Ce processus ne répond pas à la demande et dépend d'une identification clinique des potentiels donneurs d'organe. Cette étape imparfaite manque entre 30% et 60% des potentiels donneurs d'organes.

Ce projet a démontré la faisabilité et le potentiel d'un modèle d'apprentissage machine automatique basé sur des données temporelles cliniques de routines pour aider à dépister les potentiels donneurs d'organes. Ce projet a étudié plusieurs approches et un modèle de réseau de neurones convolutif supervisé semblait supérieur. Ce modèle a été validé dans un groupe interne séparé, dans des sous-groupes cliniques d'intérêt, dans une approche prospective simulée et dans une population externe. Ce projet a aussi démontré la possibilité d'utiliser du peaufinage et du transfert de connaissances pour améliorer les performances du modèle dans la population de validation externe.

Ce modèle, le premier en son genre, est relativement performant et prometteur pour le domaine du don d'organe. Avec l'interconnexion des différents dossiers médicaux électroniques, ce modèle a le potentiel de modifier et d'améliorer radicalement la détection des potentiels donneurs d'organes. Bien que notre modèle ait été développé dans le contexte du don d'organes, l'approche utilisée pourrait aussi servir à dépister d'autres types de pathologies rares, une problématique fréquente dans le domaine de la science des données en santé. Des étapes de validation prospectives et d'amélioration du modèle, notamment l'ajout de données spécifiques, sont toutefois nécessaires avant une utilisation clinique de routine.

Références bibliographiques

1. Transplant Québec. *Statistiques officielles 2021*. https://www.transplantquebec.ca/sites/default/files/bilan_2021_final_public.pdf (2022).
2. Gouvernement du Québec. Naissances, décès et mariages par mois et par trimestre, Québec, 2010-2022. https://statistique.quebec.ca/fr/produit/tableau/naissances-deces-et-mariages-par-mois-et-par-trimestre-quebec (2021).
3. Krmpotic, K., Payne, C., Isenor, C. & Dhanani, S. Delayed Referral Results in Missed Opportunities for Organ Donation After Circulatory Death. *Crit Care Med* **45**, 989–992 (2017).
4. Kutsogiannis, D. J., Asthana, S., Townsend, D. R., Singh, G. & Karvellas, C. J. The incidence of potential missed organ donors in intensive care units and emergency rooms: a retrospective cohort. *Intensive Care Med* **39**, 1452–9 (2013).
5. Sairanen, T. *et al.* Lost potential of kidney and liver donors amongst deceased intracerebral hemorrhage patients. *Eur J Neurol* **21**, 153–159 (2014).
6. Opdam, H. & Silvester, W. Identifying the potential organ donor: an audit of hospital deaths. *Intensive Care Med* **30**, 250–254 (2004).
7. Matesanz, R., Domínguez-Gil, B., Coll, E., de La Rosa, G. & Marazuela, R. Spanish experience as a leading country: What kind of measures were taken? *Transplant International* **24**, 333–343 (2011).
8. Zier, J. L., Spaulding, A. B., Finch, M., Verschaetse, T. & Tarrago, R. Improved Time to Notification of Impending Brain Death and Increased Organ Donation Using an Electronic Clinical Decision Support System. *American Journal of Transplantation* **17**, 2186–2191 (2017).
9. Fernandes, A. P., Gomes, A., Veiga, J., Ermida, D. & Vardasca, T. Imaging screening of catastrophic neurological events using a software tool: Preliminary results. in *Transplantation Proceedings* (2015). doi:10.1016/j.transproceed.2015.03.021.

10. Transplant Québec. Diagnostic du décès neurologique. *2012* https://www.transplantquebec.ca/sites/default/files/leg-for-001f_v1_0.pdf.

11. Rossall, J., Hibbitt, T. & The Canadian Bar Association. Ontario court leaves definition of death in doctors hands. *2019* https://www.cba.org/Sections/Health-Law/Articles/2019/Ontario-court-leaves-definition-of-death-in-doctor.

12. Shemie, S. D. *et al.* Le don apres un deces d'origine cardiocirculatoire au Canada. *Can Med Assoc J* **175**, SF1–SF1 (2006).

13. Shemie, S. D. *et al.* De l'atteinte cerebrale grave au diagnostic de deces neurologique : recommendations issues du Forum canadien. *Can Med Assoc J* **174**, SF1–SF13 (2006).

14. Transplant Québec. *Statistiques officielles 2020*. https://www.transplantquebec.ca/sites/default/files/bilan_2020_public_v2.pdf (2021).

15. Transplant Québec. Statistiques officielles 2019. (2020).

16. Matesanz, R. & Dominguez-Gil, B. Strategies to optimize deceased organ donation. *Transplant Rev* **21**, 177–188 (2007).

17. Matesanz, R., Domínguez-Gil, B., Coll, E., Mahíllo, B. & Marazuela, R. How Spain Reached 40 Deceased Organ Donors per Million Population. *American Journal of Transplantation* **17**, 1447–1454 (2017).

18. World Health Organization. Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death (2020).

19. Institut canadien d'information sur la santé. *Statistiques annuelles sur les transplantations d'organes au Canada : Dialyse, transplantation et don d'organes, 2010 à 2019*. (2020).

20. Ferrari, N., Gagné, V., Désy, F., Gonthier, C. & Institut national d'excellence en santé et en services Sociaux. *Utilisation de l'oxygénation extracorporelle par membrane (ECMO) chez l'adulte au Québec*.

https://www.inesss.qc.ca/fileadmin/doc/INESSS/Rapports/Cardio/INESSS_Avis_ECMO.pdf (2019).

21. Institut national d'excellence en santé et en service Sociaux. *Évaluation des données probantes sur les dispositifs d'assistance ventriculaire gauche HeartMate II® et HeartWare® pour le traitement de l'insuffisance cardiaque chronique terminale*. https://www.inesss.qc.ca/fileadmin/doc/INESSS/Rapports/Cardio/INESSS_resume_DAV_FR.pdf (2012).

22. Shemie, S. D. *et al.* Severe brain injury to neurological determination of death: Canadian forum recommendations. *Cmaj* **174**, S1-13 (2006).

23. Greer, D. M. *et al.* Determination of Brain Death/Death by Neurologic Criteria: The World Brain Death Project. *JAMA - Journal of the American Medical Association* **324**, 1078–1097 (2020).

24. Statistique Canada. Tableau 13-10-0715-01 Décès, selon le lieu de décès (en milieu hospitalier ou ailleurs qu'en milieu hospitalier). (2020) doi:https://doi.org/10.25318/1310071501-fra.

25. Québec., I. de la statistique du. Causes de décès (liste détaillée) selon le sexe, Québec, 2000-2021. (2022).

26. Zavalkoff, S. *et al.* Potential organ donor identification and system accountability: expert guidance from a Canadian consensus conference. *Canadian Journal of Anesthesia* **66**, 432–447 (2019).

27. Squires, J. E. *et al.* Criteria to identify a potential deceased organ donor: A systematic review. *Crit Care Med* **46**, 1318–1327 (2018).

28. Opdam, H. I. & Silvester, W. Identifying the potential organ donor: An audit of hospital deaths. *Intensive Care Med* (2004) doi:10.1007/s00134-004-2185-9.

29. Canadian Institute for Health Information. Deceased Organ Donor Potential in Canada. 1–35 (2014).

30. Kompanje, E. J. O., Bakker, J., Slieker, F. J. A., Ijzermans, J. N. M. & Maas, A. I. R. Organ donations and unused potential donations in traumatic brain injury, subarachnoid haemorrhage and intracerebral haemorrhage. *Intensive Care Med* **32**, 217–222 (2006).

31. Redelmeier, D. A., Markel, F. & Scales, D. C. Organ donation after death in Ontario: A population-based cohort study. *CMAJ* **185**, (2013).

32. O'Brien, Y. *et al. Predicting Expected Organ Donor Numbers in Australian Hospitals Outside of the Donate-Life Network Using the ANZICS Adult Patient Database*. *Transplantation* vol. 102 (2018).

33. Procaccio, F., Rizzato, L., Ricci, A., Venettoni, S. & Costa, A. N. Do 'silent' brain deaths affect potential organ donation? *Transplant Proc* **42**, 2190–2191 (2010).

34. Mccallum, J., Ellis, B., Dhanani, S. & Stiell, I. G. Solid organ donation from the emergency department - A systematic review. *Canadian Journal of Emergency Medicine* **21**, 626–637 (2019).

35. McCallum, J., Yip, R., Dhanani, S. & Stiell, I. Solid organ donation from the emergency department - Missed donor opportunities. *Canadian Journal of Emergency Medicine* **22**, 701–707 (2020).

36. Blackstock, M., McKeown, D. W. & Ray, D. C. Controlled organ donation after cardiac death: Potential donors in the emergency department. *Transplantation* **89**, 1149–1153 (2010).

37. Gatward, J. J., O'Leary, M. J., Sgorbini, M. & Phipps, P. R. Are potential organ donors missed on general wards? A 6-month audit of hospital deaths. *Medical Journal of Australia* **202**, 205–209 (2015).

38. Courville Aaron Goodfellow lan, B. Y. *Deep Learning - Ian Goodfellow, Yoshua Bengio, Aaron Courville - Google Books*. *MIT Press* (2016).

39. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat Med* **25**, 30–36 (2019).

40. Malinchoc, M. *et al.* A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* **31**, 864–871 (2000).

41. Chen, Z., Yeo, C. K., Lee, B. S. & Lau, C. T. Autoencoder-based network anomaly detection. in *2018 Wireless Telecommunications Symposium (WTS)* vol. 38 1–5 (IEEE, 2018).

42. Beam, A. L. & Kohane, I. S. Big Data and Machine Learning in Health Care. *JAMA* **319**, 1317 (2018).

43. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* vol. 110 12–22 Preprint at https://doi.org/10.1016/j.jclinepi.2019.02.004 (2019).

44. Field, A. *Logistic regression Logistic regression Logistic regression*. *Discovering Statistics Using SPSS* (Springer New York, 2012). doi:10.1007/978-1-4419-1742-3.

45. Breiman, L. *Statistical Modeling: The Two Cultures*. *Statistical Science* vol. 16 (2001).

46. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *The Lancet* **393**, 1577–1579 (2019).

47. Leisman, D. E. *et al.* Development and Reporting of Prediction Models. *Crit Care Med* **48**, 623–633 (2020).

48. Yusuf, M. *et al.* Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open* **10**, e034568 (2020).

49. Chen, J. H. & Asch, S. M. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *New England Journal of Medicine* **376**, 2507–2509 (2017).

50. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* **65**, 386–408 (1958).

51. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

52. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* **115**, 211–252 (2015).

53. Alyafeai, Z. & Ghouti, L. A fully-automated deep learning pipeline for cervical cancer classification. *Expert Syst Appl* **141**, (2020).

54. Llamas, J., Lerones, P. M., Medina, R., Zalama, E. & Gómez-García-Bermejo, J. Classification of architectural heritage images using deep learning techniques. *Applied Sciences (Switzerland)* **7**, (2017).

55. Khan, A., Sohail, A., Zahoora, U. & Qureshi, A. S. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev* **53**, 5455–5516 (2020).

56. Aloysius, N. & Geetha, M. A review on deep convolutional neural networks. in *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017* vols 2018-Janua 588–592 (Institute of Electrical and Electronics Engineers Inc., 2018).

57. Singh, S. P. *et al.* 3d deep learning on medical images: A review. *Sensors (Switzerland)* **20**, 1–24 (2020).

58. Bank, D., Koenigstein, N. & Giryes, R. Autoencoders. (2020).

59. Charte, D., Charte, F., del Jesus, M. J. & Herrera, F. A Showcase of the Use of Autoencoders in Feature Learning Applications. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 11487 LNCS 412–421 (2019).

60. Ferreira, M. F., Camacho, R. & Teixeira, L. F. Autoencoders as Weight Initialization of Deep Classification Networks for Cancer versus Cancer Studies. (2020).

61. Pang, G., Shen, C., Cao, L. & Hengel, A. van den. Deep Learning for Anomaly Detection: A Review. *ACM Comput Surv* **54**, (2021).

62. Johnson, A. E. W. *et al.* Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE* **104**, 444–466 (2016).

63. Sanchez-Pinto, L. N., Luo, Y. & Churpek, M. M. Big Data and Data Science in Critical Care. *Chest* **154**, 1239–1248 (2018).

64. Gil, P. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. *Forbes* https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=6e2303c56f63 (2016).

65. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min* **10**, 1–17 (2017).

66. Enders, C. K. *Applied missing data analysis [electronic resource]*. *Applied missing data analysis.* (Guilford Press, 2010).

67. Emmanuel, T. *et al. A survey on missing data in machine learning*. *Journal of Big Data* vol. 8 (Springer International Publishing, 2021).

68. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat Med* **25**, 24–29 (2019).

69. Johnson, A. *et al.* MIMIC-IV. Preprint at https://doi.org/10.13026/S6N6-XD98 (2021).

70. Pollard, T. J. *et al.* The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data* (2018) doi:10.1038/sdata.2018.178.

71. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J Big Data* **6**, (2019).

72. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* **5**, 493–497 (2021).

73. Weldon, J., Ward, T. & Brophy, E. Generation of Synthetic Electronic Health Records Using a Federated GAN. (2021).

74. Deng, Y. *et al.* Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. *NPJ Digit Med* **4**, (2021).

75. Ebbehoj, A., Thunbo, M., Andersen, O. E., Glindtvad, M. V. & Hulman, A. Transfer learning for non-image data in clinical research: a scoping review. *medRxiv* 2021.10.01.21264290 (2021) doi:10.1101/2021.10.01.21264290.

76. Raghu, M. & Zhang, C. Understanding Transfer Learning for Medical Imaging. *Google AI blog* https://ai.googleblog.com/2019/12/understanding-transfer-learning-for.html.

77. Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A. & Seliya, N. A survey on addressing high-class imbalance in big data. *J Big Data* **5**, (2018).

78. Chawla, N. v., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2011).

79. Hinton, G. Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA* **320**, 1101 (2018).

80. Meyer, A. *et al.* Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med* **6**, 905–914 (2018).

81. Lee, C. K., Hofer, I., Eilon, G., Baldi, P. & Cannesson, M. Development and Validation of a Deep Neural Network Model for Prediction of Postoperative In-hospital Mortality. *Anesthesiology* 1–14 (2018).

82. Woerlee, G. M. ASA — Physical Status Classification. *ASA,* 5–6 (2019) doi:10.1007/978-94-009-1323-3_2.

83. Hyland, S. L. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med* **26**, 364–373 (2020).

84. Spann, A. *et al.* Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review. *Hepatology* **71**, 1093–1105 (2020).

85. Mark, E., Goldsman, D., Gurbaxani, B., Keskinocak, P. & Sokol, J. Using machine learning and an ensemble of methods to predict kidney transplant survival. *PLoS One* **14**, e0209068 (2019).

86. Medved, D. *et al.* Improving prediction of heart transplantation outcome using deep learning techniques. *Sci Rep* **8**, 3613 (2018).

87. Senanayake, S. *et al.* Machine learning in predicting graft failure following kidney transplantation: A systematic review of published predictive models. *Int J Med Inform* **130**, 103957 (2019).

88. Tang, J. *et al.* Application of Machine-Learning Models to Predict Tacrolimus Stable Dose in Renal Transplant Recipients. *Sci Rep* **7**, 42192 (2017).

89. Rabinstein, A. A. *et al.* Prediction of potential for organ donation after cardiac death in patients in neurocritical state: A prospective observational study. *Lancet Neurol* **11**, 414–419 (2012).

90. Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. M. Prognosis and prognostic research: Validating a prognostic model. *BMJ (Online)* **338**, 1432–1435 (2009).

91. Moons, K. G. M., Altman, D. G., Vergouwe, Y. & Royston, P. Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ (Online)* vol. 338 1487–1490 Preprint at https://doi.org/10.1136/bmj.b606 (2009).

92. Plana, D. *et al.* Randomized Clinical Trials of Machine Learning Interventions in Health Care. *JAMA Netw Open* **5**, e2233946 (2022).

93. Bates, D. W. & Gawande, A. A. Improving Safety with Information Technology. *New England Journal of Medicine* **348**, 2526–2534 (2003).

94. MacK, E. H., Wheeler, D. S. & Embi, P. J. Clinical decision support systems in the pediatric intensive care unit. *Pediatric Critical Care Medicine* **10**, 23–28 (2009).

95. Kwan, J. L. *et al.* Computerised clinical decision support systems and absolute improvements in care: Meta-analysis of controlled clinical trials. *The BMJ* **370**, (2020).

96. Vasey, B. *et al.* Association of Clinician Diagnostic Performance with Machine Learning-Based Decision Support Systems: A Systematic Review. *JAMA Netw Open* **4**, 1–15 (2021).

97. Hong, N. *et al.* State of the Art of Machine Learning-Enabled Clinical Decision Support in Intensive Care Units: Literature Review. *JMIR Med Inform* **10**, 1–15 (2022).

98. Downing, N. L. *et al.* Electronic health record-based clinical decision support alert for severe sepsis: A randomised evaluation. *BMJ Qual Saf* **28**, 762–768 (2019).

99. Schaaf, J., Sedlmayr, M., Schaefer, J. & Storf, H. Diagnosis of Rare Diseases: a scoping review of clinical decision support systems. *Orphanet J Rare Dis* **15**, 263 (2020).

100. Bordini, B. J. Undiagnosed and Rare Diseases in Critical Care: The Role of Diagnostic Access. *Crit Care Clin* **38**, 159–171 (2022).

101. Jain, A., McCarthy, K., Xu, M. & Stoller, J. K. Impact of a clinical decision support system in an electronic health record to enhance detection of α1-antitrypsin deficiency. *Chest* **140**, 198–204 (2011).

102. Knight, S. R. *et al.* Development of a Clinical Decision Support System for Living Kidney Donor Assessment Based on National Guidelines. *Transplantation* **102**, e447–e453 (2018).

103. CITADEL. CITADEL. https://citadel-chum.com/.

104. Conseil de recherches en sciences humaines, Conseil de recherches en sciences naturelles et en génie du Canada & Instituts de recherche en santé du Canada. *Énoncé De Politique Des Trois Conseils, Éthique De La Recherche Avec Des Êtres Humains.* (2018).

105. Regenstrief Institute. LOINC standard. https://loinc.org/ (2022).

106. Kenward, M. G. & Molenberghs, G. Last Observation Carried Forward: A Crystal Ball? *J Biopharm Stat* **19**, 872–888 (2009).

107. Lachin, J. M. Fallacies of last observation carried forward analyses. *Clinical Trials* **13**, 161–168 (2016).

108. Pargent, F., Pfisterer, F., Thomas, J. & Bischl, B. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Comput Stat* (2022) doi:10.1007/s00180-022-01207-6.

109. Chen, Z., Yeo, C. K., Lee, B. S. & Lau, C. T. Autoencoder-based network anomaly detection. *Wireless Telecommunications Symposium* **2018-April**, 1–5 (2018).

110. Fernando, T., Gammulle, H., Denman, S., Sridharan, S. & Fookes, C. Deep Learning for Medical Anomaly Detection A Survey. *ACM Comput Surv* **54**, 1–28 (2022).

111. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci Rep* **8**, (2018).

112. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). doi:10.1109/CVPR.2016.90.

113. Mai Ngoc, K. & Hwang, M. Finding the Best k for the Dimension of the Latent Space in Autoencoders. in 453–464 (2020). doi:10.1007/978-3-030-63007-2_35.

114. Canadian Institute for Health Information. *Annual Statistics on Organ Replacement in Canada: Dialysis, Transplantation and Donation, 2010 to 2019*. https://www.cihi.ca/sites/default/files/document/corr-dialysis-transplantation-donation-2010-2019-snapshot-fr.pdf (2019).

115. Canadian Institute for Health Information. *Summary statistics on organ transplants, wait-lists and donor - 2021 Statistics*. (2022).

116. Barbieri, S. *et al.* Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk. *Sci Rep* **10**, 1111 (2020).

117. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* **110**, 12–22 (2019).

118. Moons, K. G. M. *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* **162**, W1–W73 (2015).

119. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science (1979)* **313**, 504–507 (2006).

120. Chollet, F. Keras. Preprint at (2015).

121. Python Software Foundation. Python.

122. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2016).

123. Branco, P., Torgo, L. & Ribeiro, R. P. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Comput Surv* **49**, 1–50 (2016).

124. Healey, A. *et al.* CAEP Position Statement - Management of devastating brain injuries in the emergency department: Enhancing neuroprognostication and maintaining the opportunity

for organ and tissue donation. *Canadian Journal of Emergency Medicine* **22**, 658–660 (2020).

125. Souter, M. J. *et al.* Recommendations for the Critical Care Management of Devastating Brain Injury: Prognostication, Psychosocial, and Ethical Management: A Position Statement for Healthcare Professionals from the Neurocritical Care Society. *Neurocrit Care* **23**, 4–13 (2015).

126. Sauthier, N., Bouchakri, R., Carrier, F.-M. & Chassé, M. Detection of potential organ donors; an automatic deep learning approach on temporal data. *(To be published)*.

127. Ho, A. F. W. *et al.* Assessing unrealised potential for organ donation after out-of-hospital cardiac arrest. *Scand J Trauma Resusc Emerg Med* **29**, 1–8 (2021).

128. le May, M. *et al.* From Coronary Care Units to Cardiac Intensive Care Units: Recommendations for Organizational, Staffing, and Educational Transformation. *Canadian Journal of Cardiology* **32**, 1204–1213 (2016).

129. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. *Transfusion: Understanding Transfer Learning for Medical Imaging*. http://arxiv.org/abs/1902.07208 (2019).

130. Jang, J. H., Kim, T. Y. & Yoon, D. Effectiveness of transfer learning for deep learning-based electrocardiogram analysis. *Healthc Inform Res* **27**, 19–28 (2021).

131. Shickel, B. *et al.* Deep Multi-Modal Transfer Learning for Augmented Patient Acuity Assessment in the Intelligent ICU. *Front Digit Health* **3**, (2021).

132. Wardi, G. *et al.* Predicting Progression to Septic Shock in the Emergency Department Using an Externally Generalizable Machine-Learning Algorithm. *Ann Emerg Med* **77**, 395–406 (2021).

133. Boit, J. The Effectiveness of Transfer Learning Systems on Medical Images. (2020).

134. Dayan, I. *et al.* Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med* **27**, 1735–1743 (2021).

135. Zhu, L., Liu, Z. & Han, S. Deep Leakage from Gradients. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).

136. Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Book* (2019).

137. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?' Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session* 97–101 (2016) doi:10.18653/v1/n16-3020.

138. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* **2017-Decem**, 4766–4775 (2017).

139. Lei, H. *et al.* Agile clinical research: A data science approach to scrumban in clinical medicine. *Intell Based Med* **3–4**, 100009 (2020).

140. Jackson, S., Yaqub, M. & Li, C. X. The Agile Deployment of Machine Learning Models in Healthcare. *Front Big Data* **1**, 1–7 (2019).

141. Kentish-Barnes, N. *et al.* Grief symptoms in relatives who experienced organ donation requests in the ICU. *Am J Respir Crit Care Med* **198**, 751–758 (2018).

Annexes

Annexe A : Valeur de laboratoires incluses dans le modèle

Lab name	Type de Donnée	Manquant %	Normal min	Normal max	Moyenne	Std	Médiane	IQR
white_blood_cell_count	Continu	0,1%	4	11	11,8	7,7	10,5	6,7
hemoglobin	Continu	0,1%	120	196	100,2	22,5	96,0	32,0
hematocrit	Continu	0,1%	0,35	0,6	0,3	0,1	0,3	0,1
red_blood_cell_deviation_width	Continu	0,1%	11,5	20	15,3	2,4	14,8	2,7
red_blood_cell_count	Continu	0,1%	3,8	6,2	3,3	0,7	3,2	1,0
platelet_count	Continu	0,1%	140	500	211,1	126,1	188,0	129,0
mean_corpuscular_volume	Continu	0,1%	80	101	91,0	6,0	90,9	6,8
mean_corpuscular_hemoglobin _concentration	Continu	0,1%	300	365	330,4	14,0	331,0	18,0
mean_corpuscular_hemoglobin	Continu	0,1%	24	33,5	30,1	2,2	30,2	2,5
lipemia_presence	Cat. Ord.	0,1%	0	0	0,1	0,3	0,0	0,0
icterus_presence	Cat. Ord.	0,1%	0	0	0,1	0,5	0,0	0,0
hemolysis	Cat. Ord.	0,1%	0	0	0,0	0,2	0,0	0,0
neutrophil_count	Continu	0,1%	1,3	7,7	9,1	5,6	8,0	6,3
monocyte_count	Continu	0,1%	0	1,6	0,8	0,7	0,8	0,6
lymphocyte_count	Continu	0,2%	1	4,1	1,4	3,5	1,1	1,0
sodium	Continu	0,2%	135	145	139,7	4,5	140,0	5,0
creatinine	Continu	0,2%	42	112	110,9	100,4	81,0	59,0
potassium	Continu	0,2%	3,5	5	4,0	0,6	4,0	0,7
eosinophil_count	Continu	0,3%	0	0,8	0,2	0,3	0,1	0,2
basophil_count	Continu	0,4%	0	0,3	0,0	0,1	0,0	0,0
urea	Continu	0,4%	2,8	8,8	9,3	6,8	7,2	6,2

Lab name	Type de Donnée	Manquant %	Normal min	Normal max	Moyenne	Std	Médiane	IQR
mean_platelet_volume	Continu	0,7%	6,5	13,5	10,3	1,2	10,3	1,4
glucose	Continu	1,1%	4	6,2	8,1	2,8	7,5	3,1
chloride	Continu	1,2%	96	106	105,6	5,7	106,0	7,0
magnesium	Continu	1,2%	0,7	1,01	0,9	0,2	0,8	0,2
phosphate	Continu	1,4%	0,72	1,64	1,2	0,4	1,1	0,4
albumin	Continu	2,3%	36	52	28,5	6,5	28,0	9,0
total_calcium	Continu	2,7%	2,17	2,56	2,1	0,2	2,1	0,3
inr	Continu	3,6%	0,8	1,2	1,3	0,6	1,1	0,3
partial_thromboplastin_time	Continu	3,8%	22	32	37,2	19,1	29,0	18,0
creatine_kinase	Continu	14%	24	213	795,6	6729,5	195,0	397,0
total_bilirubin	Continu	15%	7	23	31,9	69,3	13,0	14,3
alanine_aminotransferase	Continu	15%	8	39	138,4	528,2	30,0	53,0
aspartate_aminotransferase	Continu	16%	13	39	181,7	829,6	39,0	57,0
hs_troponin_t	Continu	16%	0	18	492,1	1492,2	112,0	333,8
corrected_total_calcium	Continu	24%	2,2	2,58	2,3	0,2	2,3	0,2
alkaline_phosphatase	Continu	24%	36	110	108,4	118,3	75,0	65,0
venous_ph	Continu	26%	7,31	7,43	7,4	0,1	7,4	0,1
fio2	Continu	27%	0,21	0,21	0,4	0,2	0,4	0,2
venous_pco2	Continu	29%	38	54	45,0	11,3	44,0	11,0
venous_bicarbonate	Continu	29%	21	29	24,3	5,1	24,0	5,2
venous_po2	Continu	29%	35	95	42,1	15,9	39,0	11,0
arterial_po2	Continu	29%	70	110	123,9	61,0	108,0	57,6
arterial_pco2	Continu	29%	32	45	38,9	9,9	38,0	10,0
arterial_bicarbonate	Continu	29%	19	28	23,2	4,7	23,0	4,6
venous_lactic_acid	Continu	33%	0,56	2,4	1,8	2,1	1,3	1,0
venous_o2_sat	Continu	34%	0,7	1	0,7	0,1	0,7	0,2

Lab name	Type de Donnée	Manquant %	Normal min	Normal max	Moyenne	Std	Médiane	IQR
lipase	Continu	34%	10	102	53,3	157,5	23,0	29,0
arterial_o2_sat	Continu	34%	0,92	1	1,0	0,0	1,0	0,0
total_protein	Continu	37%	63	81	56,8	10,6	56,0	14,0
urinary_ph	Continu	38%	4,8	8	31,1	3676,6	5,0	1,5
urinary_density	Continu	38%	1	1,03	1,0	0,0	1,0	0,0
urinary_protein	Cat. Ord.	38%	0	0	0,3	0,9	0,0	0,3
urinary_glucose	Cat. Ord.	38%	0	0	2,9	9,8	0,0	0,0
urinary_blood	Cat. Ord.	38%	0	0	44,3	85,3	0,3	25,0
urinary_bilirubin	Cat. Ord.	39%	0	0	2,5	11,3	0,0	0,0
urinary_cetones	Cat. Ord.	39%	0	0	0,4	1,8	0,0	0,0
urinary_urobilinogen	Cat. Ord.	39%	0	0	7,9	25,3	0,0	0,0
urinary_nitrite	Cat. Ord.	39%	0	0	0,0	0,2	0,0	0,0
urinary_leucocytes	Cat. Ord.	39%	0	0	42,7	121,6	0,0	0,0
ionized_calcium_ph74	Continu	42%	1,12	1,32	1,1	0,1	1,2	0,1
arterial_ph	Continu	44%	7,35	7,45	7,4	0,1	7,4	0,1
lactate_dehydrogenase	Continu	46%	104	205	344,5	899,8	205,0	142,0
fibrinogen	Continu	48%	2	4,5	3,6	1,7	3,4	2,2
anticoagulant	Cat.	48%	0	0	0,4	0,5	0,0	1,0
gamma_glutamyl_transferase	Continu	49%	7	47	90,4	141,9	44,0	75,0
amylase	Continu	52%	20	104	96,2	343,0	55,0	60,5
temperature	Continu	54%	36	38	37,0	0,6	37,0	0,0
osmolality	Continu	55%	275	300	290,7	14,2	289,0	16,0
urinary_polychromia	Cat. Ord.	55%	0	0	1,0	0,2	1,0	0,0
thrombin_time	Continu	56%	12	18	29,8	32,2	17,0	7,0
ph	Continu	59%	7,37	7,43	7,4	0,1	7,4	0,1
nucleated_red_blood_cells	Continu	60%	0	0,1	0,1	0,7	0,0	0,0

Lab name	Type de Donnée	Manquant %	Normal min	Normal max	Moyenne	Std	Médiane	IQR
erythrocytes	Continu	64%	0	2	27,4	34,2	4,0	54,0
ck_mb	Continu	65%	0	19	29,7	88,8	16,0	24,3
leucocytes_count	Continu	65%	0	2	16,0	26,4	4,0	7,0
uric_acid	Continu	67%	167	441	323,7	147,5	310,0	178,0
arterial_lactic_acid	Continu	68%	0,6	2,4	2,2	2,6	1,4	1,4
anisocytosis_presence	Cat. Ord.	68%	0	0	1,6	0,8	1,0	1,0
base_excess	Continu	69%	-2,5	2,5	-1,3	4,9	-1,4	5,2
anion_gap	Continu	71%	4	14	8,5	3,9	8,0	4,0
venous_base_excess	Continu	72%	-2	3	-0,7	5,5	-0,7	6,0
cholesterol	Continu	72%	3,16	7,3	3,6	1,4	3,4	1,5
triglycerides	Continu	73%	0,43	2,82	1,7	2,0	1,4	1,0
hdl_cholesterol	Continu	74%	0,8	2,38	1,0	0,4	0,9	0,4
plt_anisocytosis_presence	Cat. Ord.	75%	0	0	1,0	0,1	1,0	0,0
urinary_mucus_presence	Cat. Ord.	76%	0	0	1,1	0,3	1,0	0,0
urinary_bacteria	Cat. Ord.	76%	0	0	1,2	0,4	1,0	0,0
elliptocytes_presence	Cat. Ord.	78%	0	0	1,0	0,2	1,0	0,0
urinary_pavimentous_cells _presence	Cat. Ord.	78%	0	0	1,0	0,2	1,0	0,0
echinocyts_presence	Cat. Ord.	81%	0	0	1,2	0,5	1,0	0,0
thyroid_stimulating_hormone	Continu	82%	0,35	5,5	3,7	8,4	2,1	2,6
direct_bilirubin	Continu	84%	0	3,6	40,8	67,0	19,6	29,6
urinary_ac_ascorb	Cat. Ord.	84%	0	0	3,3	10,3	0,0	0,0
giant_platelets_presence	Cat. Ord.	84%	0	0	1,0	0,1	1,0	0,0
hba1c	Continu	85%	0,04	0,06	0,1	0,0	0,1	0,0
acanthocytes_presence	Cat. Ord.	86%	0	0	1,1	0,3	1,0	0,0
urinary_hyalin_cylinder _presence	Cat. Ord.	86%	0	0	3,8	5,7	1,0	3,0
atypia_lympho_presence	Cat. Ord.	88%	0	0	1,0	0,0	1,0	0,0

Lab name	Type de Donnée	Manquant %	Normal min	Normal max	Moyenne	Std	Médiane	IQR
globulins	Continu	88%	21	34	29,4	8,8	28,0	11,0
toxic_granulation_presence	Cat. Ord.	88%	0	0	1,0	0,1	1,0	0,0
target_cells_presence	Cat. Ord.	88%	0	0	1,2	0,5	1,0	0,0
doehle_body_presence	Cat. Ord.	88%	0	0	1,0	0,0	1,0	0,0