Université Paris Cité

Université Paris Cité
École doctorale 393 Pierre Louis de santé publique
Centre de recherche des Cordeliers

Élaboration d'un modèle prédictif d'hypertension artérielle secondaire

Par Jean-Baptiste DE LA POIX DE FREMINVILLE

Thèse de doctorat d'INFORMATIQUE MÉDICALE

Dirigée par Bastien RANCE
Et co-dirigée par Michel AZIZI
Présentée et soutenue publiquement
le 10/01/2025

Devant un jury composé de :
Emmanuel CHAZARD, PU-PH, Université de Lille, Rapporteur
Théodora BÉJAN-ANGOULVANT, PU-PH, Université de Tours, Rapporteur
Guillaume ASSIÉ, PU-PH, Université Paris Cité, Examinateur
Claire LE HELLO, PU-PH, Université de Saint-Etienne, Examinateur
Laurence AMAR, PU-PH, Université Paris Cité, Membre invité
Anita BURGUN, PU-PH, Université Paris Cité, Membre invité
Michel Azizi, PU-PH, Université Paris Cité, Co-directeur de thèse
Bastien Rance, PU-PH, Université Paris Cité, Directeur de thèse

Titre : Élaboration d'un modèle prédictif d'hypertension artérielle secondaire

Résumé :

L'hypertension artérielle (HTA) secondaire représente environ 10% des cas d'HTA. L'hyperaldostéronisme primaire en particulier est probablement sous-diagnostiqué, alors qu'il est d'importance du fait du retentissement de cette pathologie sur plusieurs organes cibles tels que le cœur, les reins ou les artères en général.

Nous avons dans ce travail réalisé dans un premier temps une étude épidémiologique de l'HTA secondaire chez les sujets jeunes, sur la base d'une cohorte bicentrique, et montré une prévalence majeure chez ces patients, encourageant à réaliser un bilan complet dans cette population. Nous avons ensuite exploré les facteurs associés à l'aide de méthodes statistiques robustes basées sur le bootstrap. Cependant, il n'a pas été possible d'isoler une population à faible risque d'HTA secondaire qui n'aurait donc pas d'indication à un dépistage.

Nous avons ensuite utilisé les données du centre expert de l'HEGP pour développer un modèle prédictif d'HTA secondaire, que nous avons validé sur une cohorte externe du CHU de Bordeaux. Il s'agit d'un algorithme de stratification du risque d'HTA secondaire, et plus particulièrement de l'hyperaldostéronisme primaire. Nous avons par ailleurs développé des algorithmes d'apprentissage automatique, et comparé les performances des deux approches. Notre algorithme a montré une meilleure valeur prédictive négative sur la cohorte de validation que les algorithmes d'apprentissage automatique, avec comme avantage une interprétabilité qui pourrait faciliter son utilisation et son acceptation en médecine de ville. Nous avons finalement développé une application utilisable facilement pour l'utilisation de cet algorithme de stratification du risque. L'étape suivante serait la validation de cet outil sur une cohorte prospective en médecine de ville.

Mots clefs :

Hypertension artérielle secondaire, Hyperaldostéronisme primaire, Hypertension rénovasculaire, Apprentissage automatique, Médecine de précision

Title : Elaboration of a predictive model for secondary hypertension

Abstract :

Secondary arterial hypertension (HTN) accounts for approximately 10% of HTN cases. Primary aldosteronism, in particular, is most likely underdiagnosed, despite its significance due to its impact on multiple target organs such as the heart, kidneys, and arteries in general.

In this work, we first conducted an epidemiological study on secondary HTN in young individuals, based on a bicentric cohort, and demonstrated a high prevalence in these patients, highlighting the need for comprehensive evaluation in this population. We then explored associated factors using robust statistical methods based on bootstrapping. However, it was not possible to identify a low-risk population for secondary HTN that would not require screening.

Next, we used data from the expert center at HEGP to develop a predictive model for secondary HTN, which we validated on an external cohort from the Bordeaux University Hospital. This model is a risk stratification algorithm for secondary HTN, particularly for primary hyperaldosteronism. Additionally, we developed machine learning algorithms and compared the performance of both approaches. Our algorithm demonstrated better negative predictive value on the validation cohort than the machine learning algorithms, with the added advantage of interpretability, which could facilitate its use and acceptance in general practice.

Finally, we developed an easily usable application for applying this risk stratification algorithm. The next step would be to validate this tool in a prospective cohort in general practice.

Keywords :

Secondary Hypertension, Primary aldosteronism, Renovascular hypertension, Machine learning, Precision Medicine

Remerciements

Au docteur Bastien Rance.

Je te remercie de ton enthousiasme, dès la proposition de ce sujet, et pour la bienveillance et la patience dont tu as toujours fait preuve dans ton encadrement. Grâce à ta sérénité, j'ai toujours avancé avec le sentiment que nous arriverions quelque part, ce qui n'est pas toujours évident à certaines périodes où les résultats se font attendre. Je n'oublie pas non plus nos échanges sur la montagne, la voile, et les podcasts que tu m'as recommandés (et je compte bien arriver devant toi sur Virtual Regatta).

Au Professeur Michel Azizi.

Je vous remercie sincèrement, tout d'abord de m'avoir accueilli dans votre service, avec la confiance que je pouvais apporter quelque chose à votre équipe. Je suis arrivé en HTA avec la ferme intention de tout faire pour rejoindre ma famille à Tours, mais dans d'autres circonstances, j'aurais apprécié rester dans votre équipe. Merci d'avoir accepté d'encadrer ce travail de thèse. Que ce soit dans ce cadre ou dans le service, vous avez essayé de me transmettre une partie de votre savoir, avec une exigence qui m'a toujours poussé à chercher à mieux faire.

A la professeure Laurence Amar.

Il y a beaucoup de choses pour lesquelles je te remercie. La première est l'intelligence que tu as eue de permettre ma venue dans le service d'HTA, et toute l'aide que tu m'as fournie pour faire aboutir mon projet hospitalo-universitaire à Tours. Je n'en serai pas là où j'en suis aujourd'hui sans toi. Merci ensuite de tes conseils avisés au cours de de cette thèse, et de tes multiples relectures et aides à la réécriture de ce premier article. Enfin, merci de ton encadrement dans ma formation clinique d'HTA et de médecine vasculaire. Je suis très heureux de te connaitre et de pouvoir continuer à te voir régulièrement.

A la professeure Théodora Béjan-Angoulvant.

Merci d'avoir accepté d'être rapporteur de ma thèse. Ton expertise dans le domaine de l'HTA fait que ton avis et ton évaluation de mon travail seront une grande plus-value pour la suite de ce projet. Je suis ravi d'avoir rejoint l'équipe d'HTA de Tours et de pouvoir collaborer avec toi dans les années à venir, dans le cadre de la prise en charge des patients, ou de la recherche.

Au professeur Emmanuel Chazard.

Je vous remercie de me faire l'honneur d'être rapporteur de ma thèse. Votre évaluation de mon travail compte tenu de votre expertise en informatique médicale représente beaucoup pour moi, et permettra j'en suis sûr de l'améliorer et de fiabiliser les implications cliniques.

A la professeur Claire Le Hello.

Merci d'avoir accepté de faire partie de mon jury de thèse. J'ai désormais le plaisir de collaborer avec vous au sein du collège des enseignants de médecine vasculaire, et dans peu de temps au sein du groupe de travail « IAvasc ». Votre regard sur mon travail, nourri par votre connaissance de l'HTA, votre expertise clinique, et votre intérêt pour l'IA et les modèles prédictifs, revêt une grande importance pour moi.

Au professeur Guillaume Assié.

Je vous remercie d'avoir accepté de participer à mon jury de thèse. Vous avez la double expertise d'endocrinologue et de spécialiste de l'intelligence artificielle, ce qui vous donne un recul sur les modèles prédictifs d'HTA particulièrement rare. Votre critique et votre analyse de mon travail est donc particulièrement important, et je suis ravi d'avoir la chance de pouvoir en bénéficier

A la professeur Anita Burgun.

Je vous remercie d'avoir accepté de participer à mon jury de thèse, mais surtout de m'avoir accueilli au sein de votre équipe. Par l'intermédiaire de Bastien, et avec votre bénédiction, j'ai pu pendant ces années de thèse travailler à ce projet au sein d'une équipe dynamique, ambitieuse et soudée.

Je profite de ce chapitre pour remercier mes collègues et amis de l'HEGP, Julien, Nassim, Lina, Guillaume (malgré quelques années de distance), Alexis, Nour, Louise, Nicole, Aurélien, Erika, Inès, grâce à qui ces 4 années à distance de ma famille ont été plus faciles à vivre. Merci à Tristan de m'avoir tant appris, et pour les moments musicaux.

Pour finir, je remercie ma famille de son soutien. Mes parents, premier vecteur de confiance et soutien inconditionnel, qui m'ont aussi appris l'exigence. Mes sœurs, dont la plus proche est souvent le soutien le plus fort et le plus doux. Mon cousin, toujours. Et bien sûr ma femme, qui a accepté beaucoup de choses pour me permettre de poursuivre mes projets. Je suis enfin revenu à la maison, et Gaël nous a rejoint et change chaque jour un peu plus notre vie.

Liste des principales abréviations

ACM : correspondances multiples

ACP : analyses en composantes principales

AMT : automesures tensionnelles

ARA2 : antagonistes des récepteurs de l'angiotensine 2

CCAM : Classification Commune des actes Médicaux

CIM-10 : Classification Internationale des Maladies - 10ème version

DFG : débit de filtration glomérulaire

HAP : hyperaldostéronisme primaire

HTA : hypertension artérielle

IEC : inhibiteurs de l'enzyme de conversion de l'angiotensine

LASSO : Least Absolute Shrinkage and Selection Operator

MAPA : mesure ambulatoire de la pression artérielle

NER : Named Entity Recognition

NLP : Natural Language Processing

PA : pression artérielle

PMSI : Programme de Médicalisation des Systèmes d'Information

PPGL : Phéochromocytomes et paragangliomes

VPP : Valeur prédictive positive

VPN : Valeur prédictive négative

UMAP : uniform manifold approximation and projection

XGBoost : eXtreme Gradient Boosting

Sommaire

INTRODUCTION 9

CHAPITRE 1 - CONTEXTE MEDICAL 11

La pression artérielle 12

Hypertension artérielle secondaire 17

CHAPITRE 2 – L'INTELLIGENCE ARTIFICIELLE ET LES MODELES PREDICTIFS EN MEDECINE ET DANS L'HYPERTENSION ARTERIELLE SECONDAIRE 24

Modèles prédictifs en médecine 25

L'apprentissage automatique : grands principes et principaux algorithmes 28

Traitement Automatique des Langues Naturelles (*Natural Language Processing*) 37

Réseaux de neurones et apprentissage profond (*Deep learning*) 38

Les biais liés aux algorithmes d'intelligence artificielle dans le domaine de la médecine 40

Etat de l'art des algorithmes prédictifs dans le domaine de l'HTA secondaire 40

Aide à la décision – aide au diagnostic 45

CHAPITRE 3 : CONSTRUCTION D'UNE BASE DE DONNEES DE PATIENTS HYPERTENDUS 47

Description de l'entrepôt de données de l'HEGP 48

Méthodes d'identification des diagnostics d'hypertension artérielle secondaire 53

Mise en forme des bases de données finales 64

Base de données Bordeaux 71

Aspects technico-règlementaires 71

CHAPITRE 4 : HYPERTENSION ARTERIELLE SECONDAIRE CHEZ LE SUJET JEUNE 73

Introduction 74

Objectifs 75

Présentation du manuscrit 76

Complément de méthodes et résultats 104

Complément de discussion : apport de l'algorithme dans la stratégie de dépistage des patients jeunes 106

CHAPITRE 5 : MODELE PREDICTIF D'HYPERTENSION ARTERIELLE SECONDAIRE.....108

Introduction 109

Objectifs 110

Présentation du manuscrit 111

Complément de méthodes et résultats..... 142

Complément de discussion..... 170

CONCLUSION ET PERSPECTIVES GENERALES 173

BIBLIOGRAPHIE 176

TABLE DES MATIERES 187

TABLE DES ILLUSTRATIONS.....189

Introduction

L'hypertension artérielle (HTA) est un problème majeur de santé publique en raison de sa prévalence dans le monde entier. Elle touche en effet plus d'un milliard d'individus, et environ 1/3 de la population adulte mondiale, et une proportion similaire en France. Cette prévalence et son rôle dans le développement de maladies cardiovasculaires en fait le facteur de risque cardiovasculaire modifiable, c'est-à-dire sur lequel on peut avoir une action, le plus fréquent dans le monde[1–4].

Elle ne s'accompagne le plus souvent d'aucun symptôme, ce qui en fait une maladie très sous-diagnostiquée. Ainsi, on estime qu'en France seulement une personne sur 2 est dépistée, une personne hypertendue sur deux est traitée, et qu'une personne hypertendue sur quatre seulement a une pression artérielle contrôlée[5].

La charge socio-économique liée à cette maladie est majeure, principalement du fait de la prise en charge médicale des complications associées, que sont la maladie coronaire, l'insuffisance cardiaque, la fibrillation atriale, les accidents vasculaires cérébraux et la démence, et la maladie rénale chronique[6,7].

L'HTA est le plus souvent multifactorielle (facteurs génétiques, environnementaux), mais sans cause spécifique unique identifiée[8]. On parle alors d'HTA primaire, ou essentielle qui nécessite une prise en charge par des règles hygiéno-diététiques associées le plus souvent à un traitement médicamenteux poursuivi tout au long de la vie. Dans environ 10% des cas cependant, on peut retrouver une cause sous-jacente, ce qui a une grande importance pour plusieurs raisons. Tout d'abord, une cause secondaire d'HTA peut être une cause de résistance, c'est-à-dire de non-contrôle tensionnel malgré de nombreuses thérapeutiques médicamenteuses[9–11]. Ensuite, certaines causes d'hypertension artérielle ont des conséquences néfastes spécifiques, indépendamment du niveau de pression artérielles[12,13]. Enfin, il existe selon les cas des traitements spécifiques, qui peuvent améliorer le contrôle tensionnel, parfois même guérir l'hypertension artérielle, et prévenir le développement des complications spécifiques.

Les conséquences néfastes de l'hypertension artérielle sont cumulatives dans le temps. Il est donc important de s'atteler à sa prise en charge dès le diagnostic. Pour cette raison, les patients jeunes constituent une priorité, car le développement précoce d'une hypertension

artérielle aura pour conséquence, en l'absence de traitement, une exposition prolongée à ce facteur de risque, et donc des conséquences désastreuses à terme[14,15]. C'est notamment pour cette raison, et parce qu'on suspecte que la prévalence de l'hypertension artérielle y est élevée que son dépistage chez les patients est un sujet dans toutes les dernières recommandations sur l'hypertension artérielle[16–18]. Toutefois sa prévalence reste mal évaluée, et les recommandations ont fluctué ces dernières années, proposant la recherche d'HTA secondaire en dessous de 30 ans (ISH), 40 ans (ESH), ou dernièrement chez tous les patients hypertendus (ESC).

La recherche d'HTA secondaire demande des moyens logistiques et du temps[8,19,20]. Certains paramètres cliniques ou biologiques permettent de suspecter ou non une cause, mais pour une rentabilité moyenne, ce qui explique que cette recherche n'est pas systématiquement réalisée. Toutefois le sous-diagnostic d'un certain nombre de causes, notamment l'hyperaldostéronisme primaire, la forme la plus fréquente d'HTA secondaire, pousse à rechercher des moyens de dépistage plus précis[21–25].

L'avènement de nouveaux outils pour l'aide au diagnostic en médecine, dans diverses spécialités incluant la cardiologie, et notamment dans le domaine de l'intelligence artificielle[26–32], est une opportunité en ce sens. Nous avons donc débuté ce travail avec pour objectif l'amélioration du dépistage de l'hypertension artérielle secondaire pour cibler des populations à risque nécessitant une recherche diagnostique.

Organisation du manuscrit

Cette thèse a pour objectif général le développement d'outils d'aide à la décision pour le dépistage de causes d'hypertension artérielle.

Le contexte médical et la problématique de l'hypertension artérielle secondaire sera d'abord présenté dans la partie 1, puis des notions sur les modèles statistiques et informatique pouvant être applicables aux données médicales dans la partie 2. La partie 3 présentera les résultats de l'élaboration d'une base de données structurées pour l'étude de l'hypertension artérielle secondaire. La seconde contribution, présentée en partie 4, portera sur la description épidémiologique de l'hypertension artérielle chez les patients jeunes, et l'élaboration d'un premier modèle statistique pour l'exploration des facteurs associés. La partie 5 présentera la troisième contribution, l'élaboration d'un modèle prédictif d'HTA secondaire, utilisable comme outil de dépistage en médecine de ville. Nous présenterons pour finir les perspectives de ce travail.

Chapitre 1 - Contexte médical

La pression artérielle

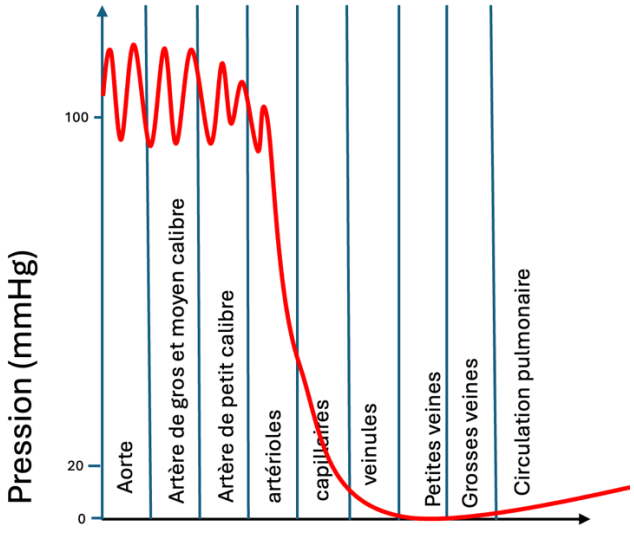
La pression artérielle (PA) est la pression (force par unité de surface) exercée par le sang sur la surface de la paroi artérielle[33].

La circulation systémique est un système à haute pression. La contraction du myocarde, et du ventricule gauche, est responsable de l'onde de pression permettant la distribution du sang. La pression est donc maximale à la sortie du cœur, puis décroît au niveau du système artériolaire, qui a notamment pour rôle l'amortissement de cette onde de pression afin de préserver les organes en aval.

Le cœur agissant comme une pompe, la pression artérielle est pulsatile, correspondant au moment de la contraction du ventricule gauche et à l'éjection du sang dans l'aorte avec une valeur maximale appelé systolique, et une valeur minimale appelée diastolique, correspondant à la relaxation et au remplissage du ventricule gauche.

La pression artérielle moyenne est en situation physiologique autour de 100 mmHg, avec une systolique autour de 120 mmHg et une diastolique autour de 80 mmHg.

Figure 1 - pression artérielle à différents niveaux du système circulatoire chez une personne en position allongée



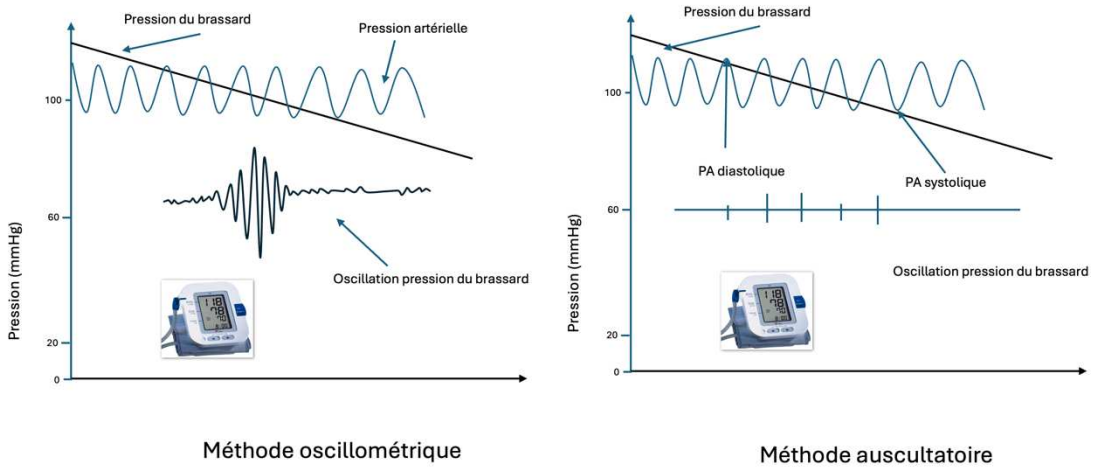
La mesure de la pression artérielle

La pression artérielle systémique peut être estimée à l'aide de différents dispositifs.

La méthode historique manuelle, ou auscultatoire, utilise un sphygmomanomètre, constitué d'un brassard gonflable et d'un manomètre. Elle n'est plus recommandée de nos jours. Elle consiste à contrôler le brassard au niveau du bras pour comprimer l'artère humérale, puis à le dégonfler progressivement jusqu'à entendre à l'aide d'un stéthoscope l'apparition des bruits de Korotkoff en aval au pli du coude[34]. L'apparition de ces bruits signale la pression artérielle systolique, puis leur nouvelle disparition signale la pression artérielle diastolique, et le retour à un régime circulatoire laminaire, le brassard étant gonflé en dessous de la pression artérielle diastolique et donc l'artère n'étant plus comprimée.

Actuellement sont utilisés des tensiomètres automatiques, se gonflant automatiquement, et estimant la pression artérielle moyenne, et par extension la systolique et la diastolique, grâce à des capteurs intégrés. Cette méthode est appelée oscillométriques car la pression artérielle moyenne correspond au régime le plus turbulent et au maximum des oscillations détectées.

Figure 2 - Mesure non-invasive de la pression artérielle, mesure auscultatoire vs. mesure oscillométrique



Il existe également d'autres sites de mesures de la pression artérielle :

- Mesure de la pression artérielle centrale de manière invasive grâce à un cathéter muni d'un capteur de pression. Cette méthode n'est pas recommandée en routine.
- Mesures automatiques de la pression artérielle au niveau de l'artère radiale par un tensiomètre bracelet. Cette méthode est également non recommandée.

Les appareils recommandés en routine sont des tensiomètres automatiques utilisant la méthode oscillométrique.

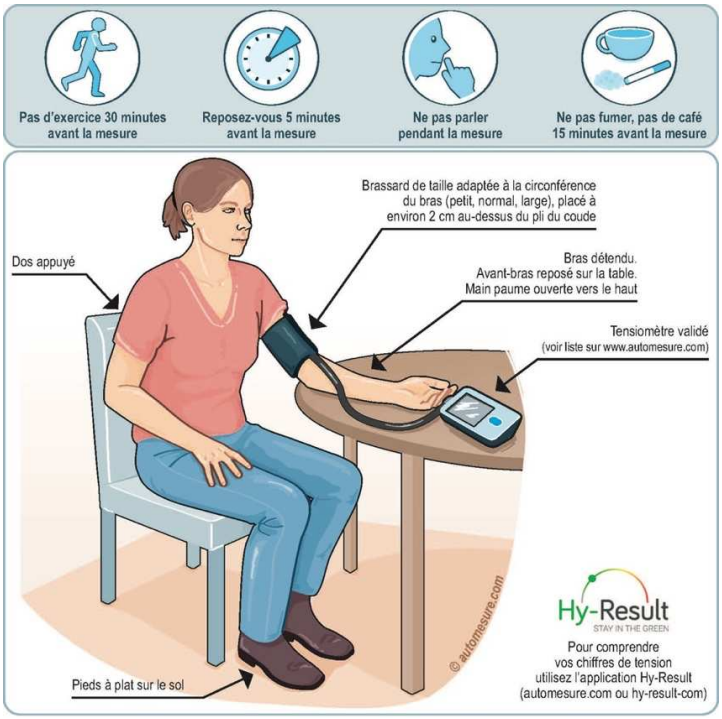
Figure 3 - tensiomètre oscillométrique



Source : www.hy-result.com

La mesure de la pression artérielle se réalise selon un protocole standardisé (Figure 4).

Figure 4 - mesure de la pression artérielle



Source : www.hy-result.com

On distingue mesure clinique et mesure ambulatoire de la pression artérielle.

La **mesure clinique** correspond à la mesure réalisée à l'hôpital ou au cabinet du médecin. Il existe plusieurs protocoles possibles, la plus courante en France étant celle recommandée par la Société Française d'Hypertension Artérielle (SFHTA) :

- Effectuer 3 mesures consécutives
- Préférentiellement avec un appareil à déclenchement automatique
- Brassard adapté à la circonférence du bras
- La moyenne des deux dernières mesures détermine le niveau de PA
- Mesurer la PA aux deux bras initialement pour dépister une asymétrie tensionnelle, puis par la suite au bras ayant la PA la plus élevée
- Dépister une hypotension orthostatique (chute de la pression artérielle à l'orthostatisme)

Les **mesures ambulatoires** de la pression artérielle sont de deux types :

- Les **automesures tensionnelles** (AMT) sont des mesures réalisées à domicile avec un tensiomètre avec brassard osccillométrique. Il s'agit de réaliser 3 mesures consécutives le matin et le soir, pendant plusieurs jour (au minimum 3). Les règles de mesure montrées Figure 4 s'appliquent également.
- La **mesure ambulatoire de la pression artérielle** (MAPA) consiste à mesurer de manière automatique la pression artérielle du patient à intervalle régulier pendant 24H (en général 3 mesures par heure le jour et 2 mesures par heure la nuit).
- Les mesures ambulatoires sont considérées plus fiables et plus reproductibles que les mesures cliniques. Le diagnostic et le suivi de l'HTA doit donc idéalement reposer sur ces dernières.

L'hypertension artérielle

L'hypertension artérielle est définie comme une augmentation chronique de la pression artérielle systolique ou diastolique systémique[35].

Les études épidémiologiques montrent une association continue entre le niveau de pression artérielle et les évènements cardiovasculaires. C'est-à-dire que plus la pression artérielle est élevée, plus le risque de complications cardiovasculaires est important.[36–40].

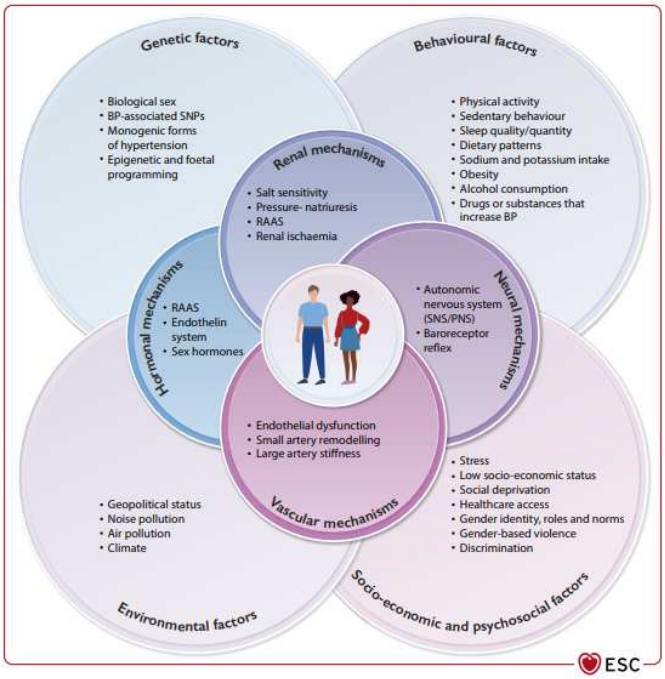
Actuellement, l'hypertension artérielle est définie selon les recommandations de l'ESC et de l'ESH comme :

- Une pression artérielle systolique ≥ 140 mmHg en mesure clinique **OU**
- Une pression artérielle diastolique ≥ 90 mmHg en mesure clinique
- **Confirmées** par une mesure ambulatoire (automesures ou MAPA), ou répétée lors d'une seconde consultation

Hypertension artérielle secondaire

La physiopathologie de l'HTA est complexe et implique de multiples déterminants environnementaux, génétique, hormonaux, hémodynamique, comportementaux.

Figure 5 – facteurs impliqués dans la physiopathologie de l'Hypertension artérielle



Source : 2024 ESC Guidelines for the management of elevated blood pressure and hypertension[18]

Dans certains cas, l'HTA est dite secondaire, c'est-à-dire résultant d'une cause identifiable, parfois curable. Elle représente environ 10% des causes d'hypertension artérielle, même si sa réelle prévalence reste difficile à estimer[8,41,42]. On l'oppose à l'hypertension artérielle primaire, ou essentielle, dont la cause est idiopathique et probablement multifactorielle, liée à des facteurs environnementaux tels que l'alimentation, la consommation d'alcool, l'activité physique, et des facteurs génétiques, certains polymorphismes étant associés à une élévation de pression artérielle [43].

Parmi les causes d'HTA secondaire, les principales sont

- Endocriniennes, liées principalement à des hypersécrétions d'origine surrénaliennes : hyperaldostéronisme primaire, syndrome de Cushing, et phéochromocytome et paragangliomes sécrétants

- Rénovasculaires : sténoses artérielles rénales athéromateuses, dysplasie fibromusculaire (sténoses ou dissections d'une artère rénale), néphropathies, coarctation de l'aorte
- Iatrogènes : consommation de réglisse, contraception oestro-progestative, médicaments vasoconstricteurs…

Il existe également des causes rares d'HTA secondaires monogéniques, par exemple le syndrome de Liddle, le syndrome de Gordon, ou l'excès apparent de minéralocorticoïdes.

Hyperaldostéronisme primaire

L'hyperaldostéronisme primaire (HAP) est la cause d'hypertension artérielle secondaire la plus fréquente, même si sa prévalence est probablement sous-estimée[21,22,24,44].

Il est lié à une sécrétion inappropriée et autonome d'aldostérone par la zone glomérulée d'une ou des deux glandes surrénales. En effet, les trois zones de la glande surrénale présentent trois rôles distincts dans la production des hormones stéroïdiennes.

Chez les humains, la 11βhydroxylase (codée par le gène CYP11B1) est responsable de l'étape finale de la biosynthèse du cortisol chez l'humain, l'aldostérone synthase (codée par le gène CYP11B2) catalyse les étapes finales de la biosynthèse de l'aldostérone. L'expression de CYP11B2 est restreinte à la zone glomérulée, tandis que CYP11B1 est exprimée dans la zone fasciculée et la zone réticulée.

Les HAP peuvent être bilatéraux, c'est-à-dire qu'ils sont liés à une sécrétion inappropriée par les deux glandes surrénales, ou unilatéraux, liés à un adénome ou à une hyperplasie d'une des glandes. Il existe également des formes familiales, plus rares, liées principalement à des mutation de canaux ioniques ou ATPases (KCNJ5, CACNA1D, CACNA1H, et CLCN2), qu'on peut retrouver au niveau germinal ou somatique[45] . L'hyperaldostéronisme familial de type 1, ou hyperaldostéronisme suppressible par les glucocorticoïdes, est un cas particulier d'HAP familial, car dû à un gène hybride entre les gènes CYP11B1 (codant pour la 11β-hydroxylase) et CYP11B2 (codant pour l'aldostérone synthase). Il doit être recherché en cas d'HTA sévère précoce. Son traitement repose un traitement glucocorticoïde, ayant un effet freinateur sur la synthèse glucocorticoïde via l'ACTH. Enfin, il existe également des carcinomes sécrétants.

La conséquence de cette hypersécrétion autonome est une hypertension artérielle due à une réabsorption hydrosodée accrue sous l'effet de l'activation du canal ENaC (Epithelial

Sodium Channel) dans le canal collecteur rénal, ainsi que dans certains cas une hypokaliémie du fait d'une fuite urinaire de potassium via le canal ROMK. La sécrétion de rénine est freinée par l'action de l'aldostérone, ce qui permet de définir le diagnostic en fonction des dosages de rénine et d'aldostérone, après arrêt des traitements interférents [46].

Son dépistage est important car il existe des traitements spécifiques de cette pathologie. Le traitement repose sur les antagonistes minéralocorticoïdes dans la plupart des cas, ou un traitement chirurgical par surrénalectomie en cas de latéralisation de la sécrétion d'aldostérone par une seule des deux glandes surrénales. Par ailleurs, en l'absence de traitement spécifique, il est bien démontré que la sécretion inappropriée d'aldostérone est responsable sur le long terme de dommages cardiaques, rénaux, et vasculaires, indépendamment du niveau de pression artérielle[12,13,47].

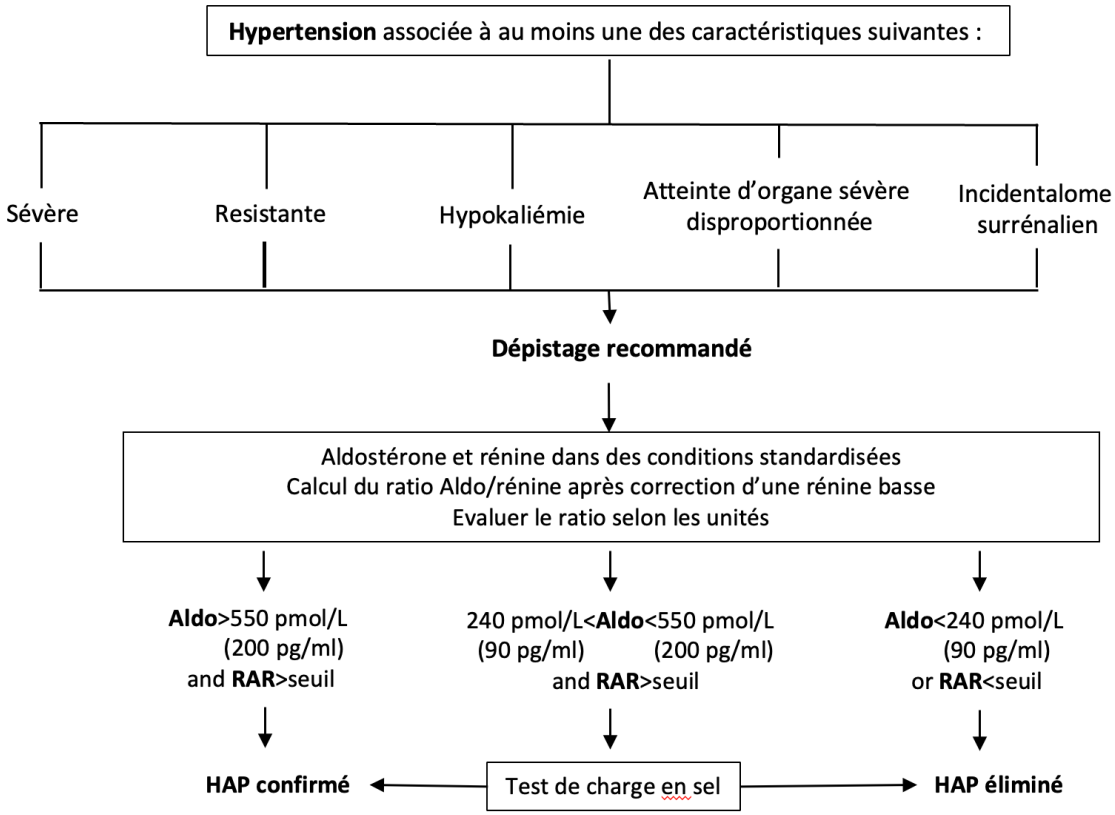
En pratique, selon les recommandations françaises, le diagnostic repose sur le dosage de la rénine et de l'aldostérone, en conditions standardisées[48] :

- Le matin, plus de 2h après le réveil
- Après 5 à 15 minutes en position assise
- En régime normosodé (natriurèse entre 100 et 200 mmol/24h)
- Avec une normokaliémie
- Après arrêt des traitements antihypertenseurs depuis 2 semaines (6 semaines pour les anti-aldostérone), à l'exception des inhibiteurs calciques non dihydropyridiniques, dans antihypertenseurs centraux, et des alpha-bloquants
- Après arrêt d'une contraception oestro-progestative depuis 6 semaines

Le diagnostic est posé si les critères décrits sur la Figure 6 sont vérifiés.

Dans certains cas, un test de freination est requis pour trancher.

Figure 6 - algorithme diagnostic de l'hyperaldostéronisme primaire



Source : adapté d'après SFE/SFHTA/AFCE primary aldosteronism consensus [46])

Plusieurs raisons sont évoquées pour expliquer le fait que cette pathologie soit sous-diagnostiquée[22,25,49] :

- La présentation clinico-biologique est atypique. L'hypokaliémie est absente dans 50% des cas selon certaines études[50] et les niveaux de pression artérielle sont variables, pouvant même être normaux. Plus rarement, on peut retrouver une hypokaliémie sans HTA.
- L'HAP est encore souvent considéré comme une cause rare, ce qui renforce la faible suspicion clinique et conduit à un taux de dépistage limité[49]
- Le diagnostic nécessite la mesure de la rénine et de l'aldostérone dans des conditions standardisées, avec notamment un arrêt des médicaments antihypertenseurs, ce qui n'est pas toujours facile à mettre en place
- Les seuils des dosages hormonaux choisis dans la littérature sont variables, dépendant des méthodes de dosages et des unités

Il apparait donc nécessaire d'élargir les stratégies de dépistage de l'HAP. Ainsi, les dernière recommandations de l'ESC ont suggéré de réaliser le dépistage de cette pathologie chez tous les patients hypertendus[18,51,52], mais cette stratégie se heurte à des difficultés de réalisation et à un surcoût certain.

Hypertension artérielle rénovasculaire

L'HTA rénovasculaire est secondaire à une sténose ou une occlusion d'une artère rénale.

Cela a pour conséquence une diminution du flux sanguin rénal, et ainsi une sécretion de rénine par l'appareil juxta-glomérulaire.

Cette activation du système rénine-angiotensine-aldostérone, occasionne donc biologiquement un hyperaldostéronisme secondaire (rénine et aldostérone élevées).

L'HTA rénovasculaire peut être secondaire à :

- La maladie athéromateuse, avec formation de plaque dans la paroi artérielle, pouvant être secondairement responsables de sténoses ou d'occlusions [11].
- La dysplasie fibromusculaire (sténoses le plus souvent multifocales des artères rénales, ou des tronc supra-aortiques, avec aspect en « collier de perle »)[53]
- D'autres maladies plus rares peuvent être responsables d'infarctus rénaux, de dissections d'artères rénales, ou de sténoses (neurofibromatose de type 1, maladie de Takayasu, syndrome d'Ehlers-Danlos vasculaire)

D'autre malformations vasculaires peuvent être responsables d'hypertension artérielle liées à une diminution du débit sanguin rénal, comme la coarctation de l'aorte (sténose de l'aorte au niveau de la crosse), ou le mid-aortic syndrome (sténose de l'aorte abdominale, que l'on retrouve notamment dans des maladies génétiques telles que la neurofibromatose de type 1, ou le syndrome d'Alagille).

L'HTA rénovasculaire est dans certains cas partiellement ou complètement réversible après revascularisation de l'artère rénale. L'HTA rénovasculaire liée à une sténose artérielle rénale athéromateuse n'a souvent pas d'indication à une revascularisation , en l'absence de bénéfice prouvé sur l'HTA dans les grands essais thérapeutiques [54]. En revanche, le bénéfice de la revascularisation dans la dysplasie fibromusculaire peut être important, et doit être systématiquement évalué, notamment en cas d'HTA sévère ou de découverte récente[53].

En l'absence d'indication à une revascularisation, le traitement repose sur le blocage du système rénine-angiotensine par des inhibiteurs de l'enzyme de conversion de l'angiotensine (IEC) ou des antagonistes des récepteurs de l'angiotensine 2 (ARA2). Ces traitements doivent être maniés avec précaution car ont une influence majeure sur l'hémodynamique rénale en bloquant la vasoconstriction de l'artériole glomérulaire efférente liée à l'angiotensine 2.

Phéochromocytomes et paragangliomes (PPGL)

Les phéochromocytomes sont des tumeurs neuroendocrines surrénaliennes, développées dans la médullosurrénale, qui peuvent produire des catécholamines[55,56]. Ces tumeurs sont rares, responsable d'environ 0,6 cas pour 100,000 personnes-années. Elles peuvent être sporadiques ou associées à des mutations génétiques[55].

Ces tumeurs peuvent être malignes ou bénignes. Elles peuvent également s'accompagner de la sécrétion de catécholamines est responsable d'une hypertension artérielle consécutive à la vasoconstriction artérielle, ainsi que de divers symptômes comme des céphalées, des sueurs, ou des palpitations. Ces symptômes sont fréquents et peuvent se présenter jusque chez 20% des patients[41].

Les complications de cette pathologie sont potentiellement graves et peuvent aller jusqu'au décès, en passant par l'HTA maligne, la cardiomyopathie de stress (syndrome de Takotsubo), des accidents vasculaires cérébraux, ou des troubles du rythme cardiaque[57].

La prise en charge consiste à contrôler la pression artérielle, avant l'exérèse chirurgicale de la tumeur[58].

Les paragangliomes sont des tumeurs neuroendocrines similaires aux phéochromocytomes, mais qui ne se développent pas aux dépens des glandes surrénales mais à partir des cellules des ganglions sympathiques, et donc de localisation variable [55]. Ils peuvent être non sécrétants, ce qui est le cas à 90% lorsqu'ils se développent au niveau du cou. Ils ne sont alors pas liés à l'HTA et nous ne traiterons pas dans cette thèse de ces cas.

Le diagnostic des PPGL repose sur le dosage des dérivés des catécholamines : métanéphrines et normétanéphrines. Il peut être réalisé sur des urines de 24h, ou sur un échantillon sanguin[56]. Une imagerie par tomodensitométrie doit ensuite être réalisée, suivie éventuellement d'examens de médecine nucléaire plus spécifiques tels que la

Tomodensitométrie par Emission de Positrons (TEP) associée au marqueur [18F]-labelled fluorodihydroxyphenylalanine (F-DOPA)[59].

Les autres formes d'HTA secondaire

Le syndrome de Cushing est lié la sécrétion inappropriée de cortisol par la glande surrénale[60]. Il a de nombreuse conséquence cliniques et biologique (favorise l'obésité, le diabète, la déminéralisation osseuse, les infections…). Comme les phéochromocytomes et paragangliomes, il peut être lié à une tumeur bénigne ou maligne (visible ou non).

D'autre causes peuvent enfin être responsables d'hypertension artérielle :

- La prise de certains toxiques : l'alcool, les œstrogènes, notamment au sein d'une contraception oestro-progestative, les antiinflammatoires non stéroïdiens, les vasoconstricteurs nasaux, le réglisse, certaines drogues de synthèse telles que la cocaïne
- les néphropathies autres qu'hypertensives
- Les dysthyroïdies
- Le syndrome d'apnée obstructif du sommeil est quant à lui associé à l'HTA
- Nous décrirons dans la partie suivante les méthodes actuelles d'aide à la décision pouvant être appliquées à l'amélioration du dépistage des HTA secondaires.

Chapitre 2 – l'intelligence artificielle et les modèles prédictifs en médecine et dans l'hypertension artérielle secondaire

Modèles prédictifs en médecine

Le développement de modèles prédictifs en médecine est en pleine expansion, et s'appuie notamment sur l'intelligence artificielle et l'apprentissage automatique.

L'objectif est l'utilisation de données cliniques pour prédire la survenue de maladies ou d'évènements. Par extension, certains modèles analysent des données de patients pour diagnostiquer une maladie. Dans le cas présent, les données de patients hypertendus sont utilisées pour analyser la probabilité que l'hypertension artérielle soit secondaire à une cause particulière.

Des modèles d'aide à la décision existent depuis longtemps et la littérature scientifique sur le sujet est abondante. Nous nous focaliserons ici sur l'HTA et les modèles pouvant être applicable pour améliorer le dépistage de l'HTA secondaire.

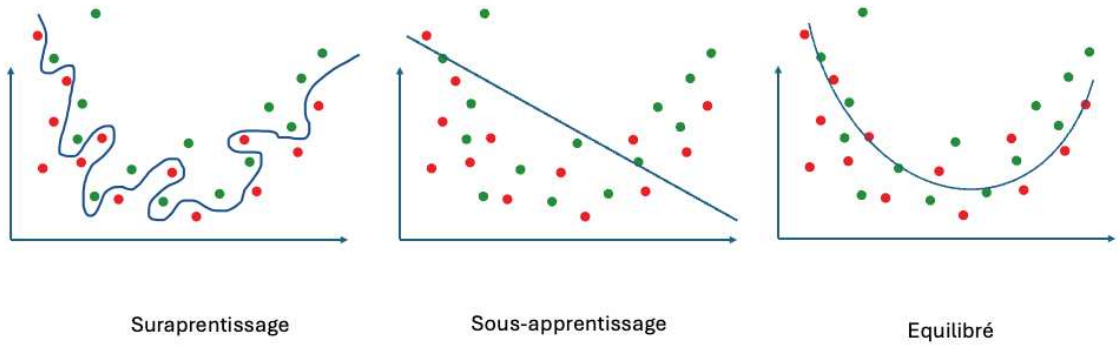
Étapes du développement d'un modèle prédictif.

Le développement d'un modèle prédictif passe par plusieurs étapes, qui sont indispensables pour assurer la validité dudit modèle et la généralisation de son application à de nouvelles données.

1. Préparation des données : il est indispensable de disposer d'une base de données nettoyée, non biaisée, avec des données fiables. La présence de biais et de données non fiables dans la base de données sur laquelle le modèle prédictif est entrainé va inévitablement conduire à une reproduction de ces biais dans le modèle[61,62].
2. Développement du modèle : différentes techniques sont utilisables, que nous détaillerons en partie par la suite. Pour la technique d'apprentissage automatique, ou d'apprentissage profond, le jeu de données initial est séparé en un jeu d'entrainement et un jeu de test, afin de ne pas évaluer le modèle entrainé sur les mêmes données. Cela conduirait à un phénomène de « surapprentissage », qui rendrait le modèle très performant sur les données d'apprentissage, mais très peu performant sur de nouvelles données, car trop spécifique[61,63].
3. Validation du modèle : il est indispensable d'évaluer ensuite le modèle sur un autre jeu de données. Cela peut être fait dans un premier temps sur une partie du jeu de données n'ayant pas servi à construire le modèle, puis cela doit idéalement être

réalisé ensuite sur un jeu de données externe, afin d'en évaluer sa validité externe et donc sa généralisabilité.

Figure 7 - exemple de surapprentissage



Métriques et évaluation des modèles

Dans le cadre du développement de modèles prédictifs, et notamment les modèles utilisant des méthodes d'intelligence artificielle, un certain nombre de métriques sont utilisées pour l'évaluation, qui sont similaires à celles utilisées classiquement dans la médecine basée sur les preuves pour l'évaluation des tests diagnostiques.

- **Précision** : proportion de prédictions correctes par rapport au nombre total de prédiction

$$Precision = \frac{VP + VN}{VP + VN + FP + FN}$$

- **Valeur prédictive positive** (VPP) : probabilité que les cas détectés par le test soient réellement positifs.

$$VPP = \frac{VP}{VP + FP}$$

- **Valeur prédictive négative** (VPN) : probabilité que les cas non détectés par le test soient réellement négatifs.

$$VPN = \frac{VN}{VN + FN}$$

- **Sensibilité** (autrement appelée **Rappel**) : capacité à identifier correctement les vrais positifs parmi tous les cas positifs.

$$Sensibilité = \frac{VP}{VP + FN}$$

- **Spécificité** : capacité à identifier correctement les vrais négatifs parmi tous les cas négatifs.

$$Sensibilité = \frac{VN}{VN + FP}$$

VP : vrais positifs ; VN : vrais négatifs ; FP : faux positifs ; FN : faux négatifs

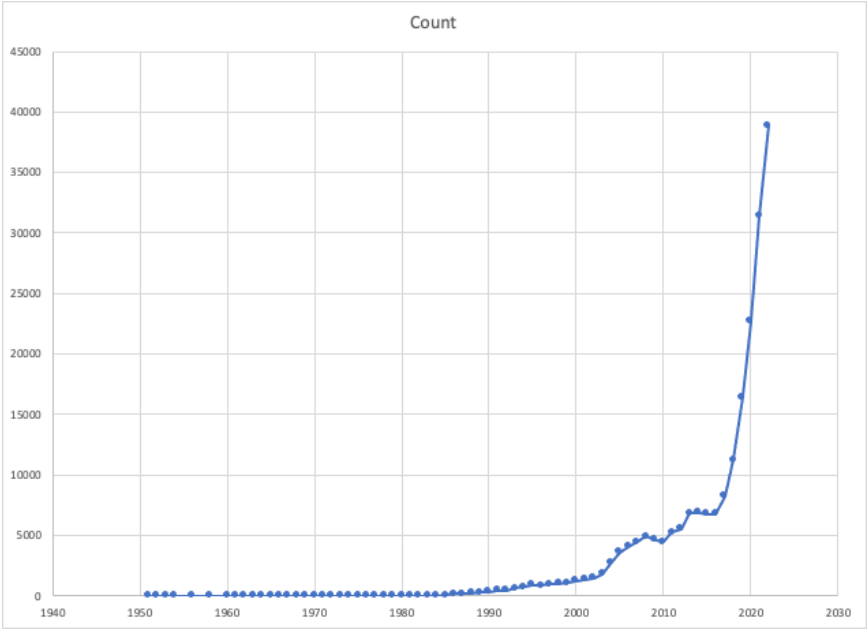
N.B : il est important de bien distinguer la précision (*accuracy* en anglais) de la VPP (*precision* en anglais)

Utilisation de l'intelligence artificielle

La recherche en médecine utilisant l'intelligence artificielle est en expansion exponentielle, comme le montre le nombre de publications traitant de modèles de ce type (Figure 8).

Les applications ne sont pas limitées aux modèles prédictifs, mais touchent également l'analyse d'image, la génomique ou l'analyse des omics en général, la communication avec les patients avec les grands modèles de langage, l'aide à la décision[64].

Figure 8 - nombre de publications (incidence) en recherchant le terme « intelligence artificielle » sur pubmed depuis 1950



Source : Pubmed

L’apprentissage automatique : grands principes et principaux algorithmes

Dans cette partie, nous décrivons les grands principes de l’apprentissage automatique, en détaillant particulièrement les modèles qui seront utilisés dans les parties suivantes sur notre base de données.

Le principe de l’apprentissage automatique

L’apprentissage automatique consiste à permettre un modèle ayant des paramètres évolutifs, d’apprendre à partir d’un jeu de données. En pratique, on soumet le modèle à un grand nombre d’exemples, et il modifie lui-même ses paramètres de manière à en extraire des règles [31,65].

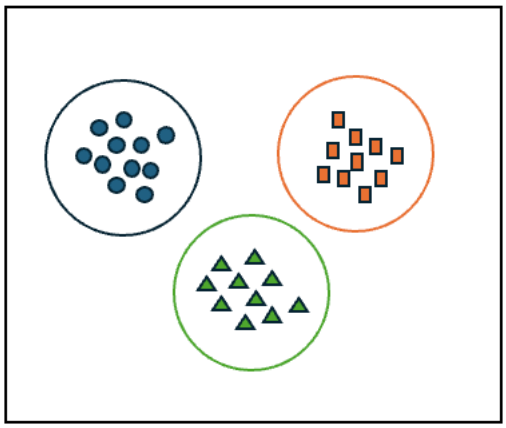
Il existe trois grands types d’apprentissage automatique :

- L’apprentissage supervisé : l’algorithme apprend à partir de données étiquetées, où la sortie attendue est connue pour chaque entrée
- L’apprentissage non supervisé : l’algorithme apprend à partir de données non étiquetées, et cherche à découvrir des structures sous-jacentes
- L’apprentissage par renforcement : les données sont créées au fur et à mesure par l’interaction de l’algorithme avec son environnement

Analyses non supervisées

Dans le cas de l’apprentissage non supervisé, l’algorithme apprend à partir de données non étiquetées, et cherche à découvrir des structures sous-jacentes. Un exemple est fourni Figure 9.

Figure 9 - exemple d'apprentissage non supervisé. L'algorithme a identifié trois types de formes différentes. Il saura ensuite les reconnaitre sur un nouveau jeu de données.



Nous allons présenter le principe de trois algorithmes d'apprentissage non supervisé : les analyses en composantes principales (ACP) ou des correspondances multiples (ACM), la méthode des k-means, et la méthode uniform manifold approximation and projection (UMAP).

Analyses en composantes principales et analyse des composantes multiples

L'ACP (Analyse en Composantes Principales, ou Principal Component Analysis) est une technique de réduction de dimensionnalité [66].

Le principe est le suivant :

- **Décomposition de la variance** : elle identifie les directions (appelées "composantes principales") dans lesquelles la variance des données est la plus élevée. Ces directions représentent les axes le long desquels les données varient le plus.
- **Transformation** : les données sont ensuite projetées sur ces nouvelles directions. La première composante principale capte le plus de variance, la deuxième capte la deuxième plus grande variance, et ainsi de suite.
- **Réduction** : on conserve les premières composantes principales qui capturent l'essentiel de la variance, tout en réduisant le nombre de dimensions, ce qui simplifie l'analyse et la visualisation des données.

L'ACM (analyse des composantes multiples), fonctionne peu ou prou comme l'ACP, mais pour des données catégorielles. Elle débute ainsi par une phase de transformation des données sous forme d'une matrice binaire. Elle se poursuit par la décomposition de la variance sur des données binaires.

K-means

K-means est une méthode de classification non supervisée qui divise un ensemble de données en un certain nombre de groupes ou "clusters". L'objectif est de regrouper des points de données similaires dans le même cluster, en minimisant la variance au sein de chaque groupe et en maximisant la différence entre les groupes [66].

Le processus de K-means fonctionne en trois étapes principales :

- **Initialisation** : on choisit aléatoirement un certain nombre de points (appelés "centroïdes") qui représenteront les centres des clusters.
- **Assignation** : chaque point de données est assigné au centroïde le plus proche, formant ainsi des clusters.
- **Mise à jour** : les centroïdes sont recalculés en prenant la moyenne des points de chaque cluster. Le processus est répété jusqu'à ce que les centroïdes ne bougent plus beaucoup ou qu'un critère de convergence soit atteint.

UMAP (uniform manifold approximation and projection)

UMAP (Uniform Manifold Approximation and Projection) est une méthode de réduction de dimensionnalité permettant la représentation des données dans un espace à faible dimension (nous avons ici représenté les données en deux dimensions)[67].

UMAP crée d'abord un graphe local des données en connectant chaque point à ses plus proches voisins. Ensuite, elle projette ces points dans un espace de plus faible dimension tout en maintenant autant que possible ces relations de voisinage.

Analyses supervisées

Le principe de l'apprentissage supervisé est que l'algorithme apprend à partir de données étiquetées, où le résultat attendu est connu pour chaque entrée.

Un exemple intuitif est celui de l'analyse d'images : on fournit à l'algorithme en entrée un grand nombre d'images de chiens et de chat, en lui donnant également la réponse (chien ou chat).

L'objectif du modèle est ensuite de pouvoir trouver le type d'animal (chien ou chat) sur de nouvelles données non étiquetées.

On distingue deux grands types d'analyses supervisées : celles pour la classification, et celles permettant la régression[66] :

- Régression : la variable à prédire est continue
- Classification : la variable à prédire est discrète ou catégorielle

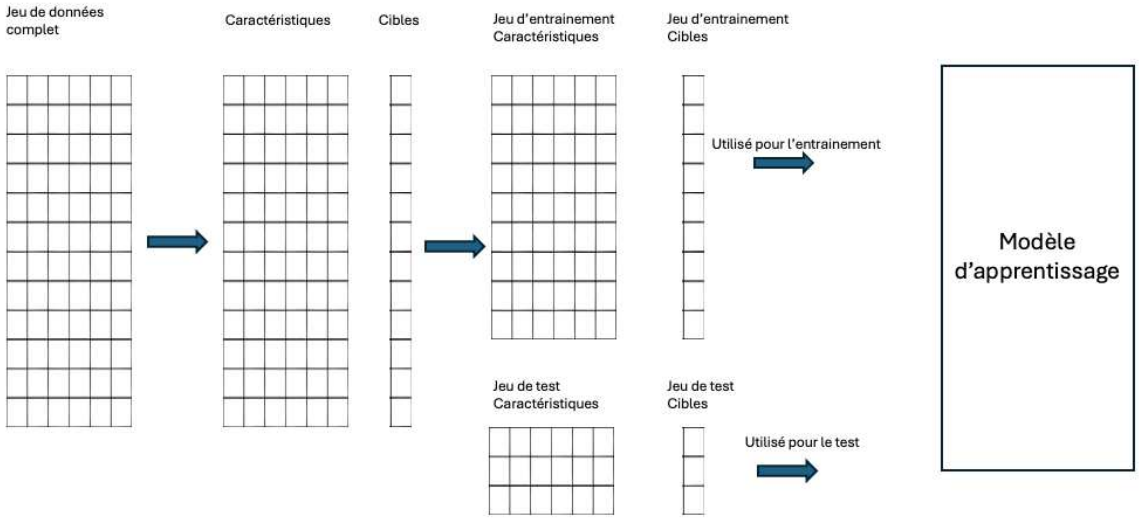
Nous allons présenter le principe des principaux algorithmes d'apprentissage supervisé, et particulièrement ceux que nous avons utilisés pour notre modèle prédictif, la régression logistique, et XGBoost.

Développement d'un algorithme d'apprentissage supervisé et principe de la validation croisée (cross-validation) et du boostrap

Développement de l'algorithme

Le développement classique comprend une phase d'entrainement, une phase de test, et une validation (Figure 10).

Figure 10 - Principe de développement d'un algorithme d'apprentissage automatique supervisé, avec séparation du jeu de données initial en un jeu d'entrainement et un jeu de test



Validation croisée

La validation croisée (ou *cross-validation*) est une technique statistique utilisée pour évaluer la performance d'un modèle d'apprentissage automatique, tout en optimisant l'utilisation des données disponibles. Elle permet de tester la robustesse d'un modèle en le validant sur plusieurs sous-ensembles de données.

Elle consiste à diviser le jeu de données en plusieurs parties, généralement appelées plis (en anglais *folds)*, puis à entrainer le modèle plusieurs fois. À chaque itération, un des plis est utilisé comme ensemble de test, et les autres sont utilisés pour l'entraînement. Les performances du modèle sont calculées pour chaque itération. En moyenne, cela donne une estimation plus fiable de la performance globale du modèle.

La validation croisée est particulièrement utile dans les situations où l'on dispose d'un petit jeu de données, car elle maximise l'utilisation des données disponibles tout en donnant une estimation plus robuste des performances du modèle[66].

Bootstrap (tirage avec remise)

Nous présentons ici la méthode du *bootstrap*, utilisée pour le recalcul de probabilités dans les modèles présentés en partie 4 et 5.

Le *bootstrap* est une méthode d'inférence statistique, utilisant le rééchantillonnage pour estimer un paramètre dans une distribution[66].

En pratique :

On génère de nouveaux échantillons de taille égale à partir de la population initiale par tirage aléatoire avec remise

Pour chaque échantillon, on calcule notre paramètre à estimer

On répète ces opération un très grand nombre de fois pour obtenir une distribution des valeurs de l'estimateur, qui va permettre d'obtenir une valeur moyenne, et des intervalles de confiance.

Classification

Régression logistique

Il s'agit d'une méthode permettant de prédire la probabilité qu'une variable (cible, ou target) appartienne à une classe particulière, en fonction d'autres variables indépendantes (caractéristiques, ou features)[66].

Le modèle mathématique sous-jacent le plus fréquent utilise la fonction sigmoïde (ou logistique), ou pour transformer une combinaison linéaire des variables d'entrée en une probabilité entre 0 et 1.

La fonction sigmoïde s'écrit :

$$f(x) = \frac{1}{1 + e^{-x}}$$

L'équation générale du modèle est :

$$y = f(w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3 + b)$$

Où y est la variable à prédire, w_i sont les coefficients de régression, b est l'ordonnée à l'origine, et x_i sont les variables indépendantes.

On obtient donc

Le développement du modèle suit le déroulement classique de l'apprentissage automatique :

- Initialisation des paramètres du modèle : ici les coefficients de régression
- Calcul de la prédiction pour chaque instance du jeu d'entrainement

Calcul de l'erreur (différence entre la prédiction et la valeur réelle) par une fonction de coût, on utilise ici la vraisemblance, et on cherche à minimiser:

$$L(\beta) = \prod \pi_i^{y_i} \times (1 - \pi_i)^{(1-y_i)}$$

Où:

- y_i est la valeur observée de la variable dépendante
- π_i est la probabilité prédite pour $y_i = 1$

- Mise à jour des paramètres pour minimiser l'erreur par la méthode de la descente de gradient : ajustement des paramètres par itérations successives en suivant la pente de la fonction de coût dans le sens qui réduit l'erreur.
- Répétition des étapes jusqu'à convergence

L'avantage de cette méthode est qu'elle est rapide et simple à implémenter.

D'autres méthodes d'optimisation sont possibles, comme la régularisation par la méthode du LASSO (Least Absolute Shrinkage and Selection Operator). Il s'agit d'ajouter une pénalité proportionnelle à la somme des valeurs absolues des coefficients, ce qui peut même entraîner la suppression de certaines caractéristiques en forçant leurs coefficients à zéro.

De la même manière la régression de crête (*ridge regression* en anglais), est une méthode de régularisation qui ajoute une pénalité proportionnelle à la somme des carrés des coefficients du modèle.

La combinaison de ces deux méthodes porte le nom *d'Elastic net*.

L'intérêt de cette méthode est de contrôler la complexité du modèle et d'éviter le surapprentissage.

La formule de la fonction de coût agrémentée de la pénalisation *Elastic Net* est la suivante :

$$\prod \pi_i^{y_i} \times (1 - \pi_i)^{(1-y_i)} + \lambda_1 \sum |w_i| + \lambda_2 \sum {w_i}^2$$

Où $\lambda_1$est le paramètre de régularisation LASSO, et $\lambda_2$est le paramètre de régularisation Ridge regression

XGBoost

XGBoost (*eXtreme Gradient Boosting*) est un algorithme d'apprentissage automatique basé sur le principe du *boosting* par gradient[66].

Le boosting est une technique d'ensemble où des modèles faibles sont ajoutés séquentiellement pour corriger les erreurs des modèles précédents. Le modèle commence par un prédicteur simple (comme un arbre de décision de petite taille), et chaque nouvel arbre construit est ajusté pour corriger les erreurs de l'ensemble des arbres précédents.

Dans le cas de XGBoost, une technique de descente de gradient est utilisée pour réduire les erreurs du modèle. À chaque étape, l'algorithme cherche à minimiser les différences entre les prédictions actuelles et les valeurs réelles, en ajustant chaque arbre pour combler les lacunes des précédents.

La fonction de coût minimisée par l'algorithme intègre deux éléments :

- L'erreur d'entraînement (log-loss pour la classification) :

  $H = -\frac{1}{N}\sum(y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \hat{y}_i))$

  où $\hat{y}$ est la valeur prédite, et y la valeur réelle
- La complexité du modèle pour éviter le surapprentissage (une régularisation de type LASSO ou Elastic Net pour limiter la complexité des arbres).

Les avantage de XGBoost sont sa vitesse d'apprentissage, la gestion automatique des valeurs manquantes, et ses performances notamment dans le contrôle du surapprentissage.

Autres algorithmes

De nombreux autres algorithmes d'apprentissage automatique ont été développés, tels que les *random forests*, les machines à vecteurs de support, ou des algorithmes faisant intervenir la théorie Bayésienne tels que Naïve Bayes.

Nous ne développerons pas ici le détail de fonctionnement de ces algorithmes.

Régression

Régression linéaire

La régression linéaire est une méthode qui permet de modéliser la relation entre une variable dépendante (ou cible) et une ou plusieurs variables indépendantes (ou caractéristiques). Le principe est de trouver une relation linéaire entre ces variables afin de prédire la valeur de la variable cible.

Elle suppose que la relation entre les variables est linéaire. Cela signifie que l'on peut approximer cette relation par une droite (ou un hyperplan dans le cas de plusieurs variables indépendantes).

L'équation générale pour une régression linéaire multiple (plusieurs variables indépendantes) est :

$$y = w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3 + b$$

Où y est la variable à prédire, w_i sont les coefficients de régression, b est l'ordonnée à l'origine, et x_i sont les variables indépendantes.

De la même manière que pour la classification, il existe de nombreux autres algorithmes utilisant par exemple les arbres de décision ou le boosting, que nous ne détaillerons pas ici.

Identification de motifs fréquents

Certains algorithmes ont pour objectif d'identifier des motifs fréquents dans une base de données. C'est le cas d'*apriori* et de *prefixpan.*

Apriori

L'algorithme Apriori est conçu pour identifier les ensembles d'éléments fréquents dans une base de données transactionnelle. Le principe de cet algorithme repose sur la génération d'éléments candidats à partir d'ensembles fréquents, puis d'élaguer en éliminant ceux ayant un support (proportion de transactions contenant cet ensemble) inférieur à un seuil. Ce processus est ensuite répété de manière itérative[68].

Prefixpan

Prefixpan est utilisé pour extraire des motifs séquentiels de bases de données[69].

Le processus est séquentiel et comprend plusieurs étapes :

- Identification des éléments fréquents dans la base de données
- Projection de la base de données pour chaque élément fréquent en ne gardant que les éléments comprenant cet élément
- Nouvelle itération de l'algorithme sur chaque base de données projetée

Traitement Automatique des Langues Naturelles (*Natural Language Processing*)

Le NLP (Natural Language Processing, ou Traitement Automatique du Langage Naturel en français) désigne un domaine de l'intelligence artificielle qui se concentre sur l'interaction entre les ordinateurs et les langues humaines. Il englobe des techniques permettant aux machines d'interpréter, de générer et de répondre au langage humain.

Le NLP, popularisé par son application dans les chatbots depuis quelques années, a des nombreuses applications.

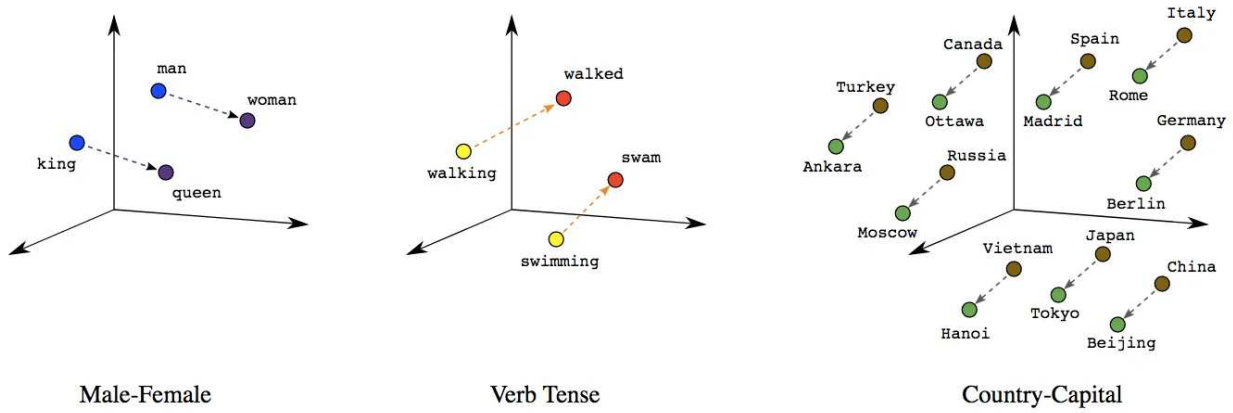
Parmi celles-ci, la reconnaissance d'entités nommées (Named Entity Recognition, ou NER), et la classification de texte, est utilisée dans le domaine médical pour extraire de l'information de documents textes, et notamment de comptes-rendus médicaux, qui sont souvent non structurés et peu standardisés, et dont l'information est par conséquent complexe à extraire.

Nous décrivons dans cette partie un concept utilisé dans le domaine du NER afin d'optimiser les performances des algorithmes de reconnaissance d'entité.

Word-Embedding

Un *embedding* consiste à vectoriser des termes dans un espace multidimensionnel afin de prendre en compte le contexte et les relations sémantiques entre les mots. Des mots différents mais de signification similaire sont ainsi proches dans cet espace vectoriel. Les opérations sur ces vecteurs peuvent également être rapprochées des relations existantes entre différents mots (Figure 11). Les *embeddings* sont construits en prenant en compte les contextes d'utilisation d'un mot dans un large corpus de textes. Le choix du corpus peut avoir une influence sur la qualité des *embeddings* construits.

Figure 11 - Relations linéaires entre les mots vectorisés par embedding. La relation entre homme et femme correspond à un certain déplacement dans l'espace, tout comme le lien entre participe présent et passé, ou entre un pays et sa capitale.



Source : developers.google.com

Réseaux de neurones et apprentissage profond (*Deep learning*)

L'apprentissage profond est une technique d'apprentissage automatique basée sur des réseaux de neurones artificiels multicouches[66].

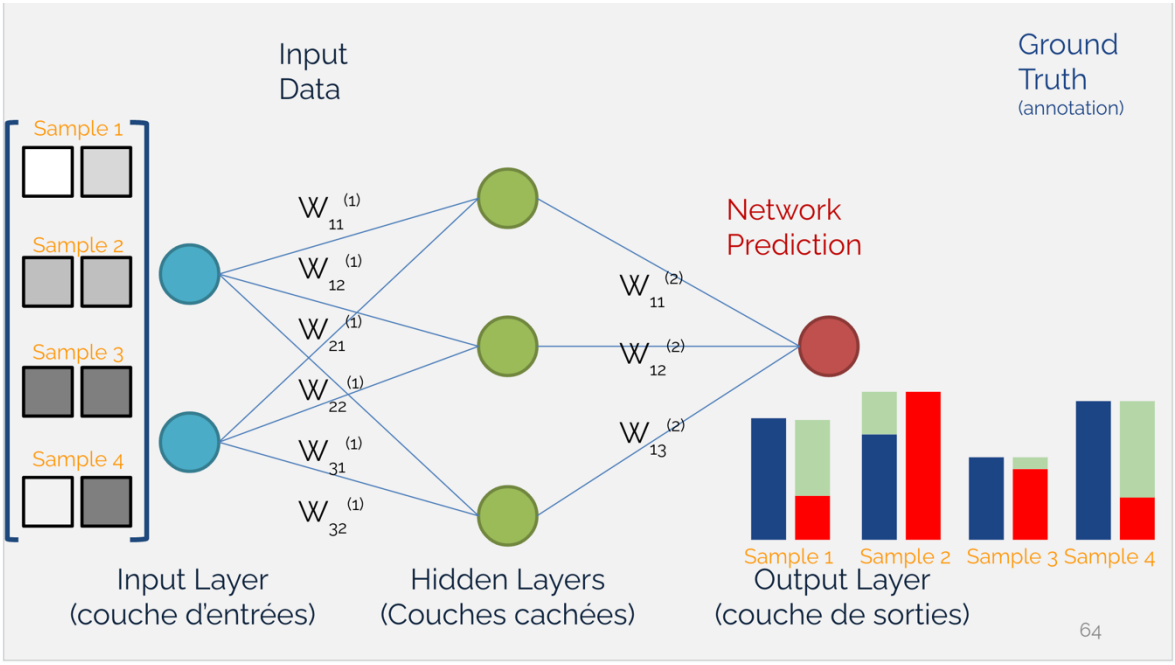
L'apprentissage profond utilise des réseaux de neurones artificiels avec de nombreuses couches cachées. Chaque couche est composée de nœuds (neurones) interconnectés. Les connexions entre les nœuds ont des poids qui sont ajustés durant l'apprentissage. Ils ont la particularité d'effectuer des opérations non linéaires à chaque couche. Dans la figure 12, W_{ii} représente les fonctions d'activation des nœuds, qui sont responsables de leurs poids.

Les paramètres du réseau sont modifiés au fur et à mesure de l'apprentissage selon un principe de rétropropagation : les données d'entrée traversent chaque couche jusqu'à atteindre la sortie, puis l'erreur par rapport à la valeur attendue est calculée, et propagée en arrière dans le réseau. Dans la figure 12, les prédictions des couches de sortie vont être évaluées par rapport aux annotations pour chaque échantillon, et les poids des connexions sont ensuite ajustés de manière à minimiser l'erreur.

L'intérêt de ces architectures est de pouvoir apprendre des représentations très abstraites des données.

Cependant, ils nécessitent un très grand nombre de données afin de pouvoir être optimisés.

Figure 12 – architecture d'un réseau de neurones.



Source : avec l'aimable autorisation de Bastien Rance

LSTM/Bi-LSTM

Nous nous attardons sur l'architecture d'un type de réseaux de neurones profonds, les Long-Short Term memory (LSTM), car ils ont été utilisés lors de l'étape de reconnaissance diagnostique lors de la constitution de notre base de données (partie 1).

Les réseaux de neurones de type *Long-Short Term memory* (LSTM), et bi-LSTM (*bi-directional LSTM*), sont des architectures de réseaux de neurones récurrents (RNN) c'est-à-dire qu'ils comprennent des couches récurrentes permettant de « se souvenir » des informations précédentes, dont la sortie est connectée à l'entrée, et des vecteurs d'état cachés, stockant l'information « mémorisée » au fur et à mesure. Ils diffèrent des RNN par le fait qu'ils contiennent des structures appelées cellules de mémoire, comprenant également une porte d'oubli (*forget gate*) permettant la mise à jour sélective de la mémoire. Les Bi-LSTM comprennent quant à eux des couches traitant la séquence d'entrée dans des directions opposées, ce qui permet à chaque cellule d'avoir l'information passée et future. Ce type d'architecture est souvent plus performant pour les taches de NLP.

Les biais liés aux algorithmes d'intelligence artificielle dans le domaine de la médecine

Les biais liés aux modèles d'intelligence artificielle en médecine sont nombreux. Certains sont plus spécifiques, bien que cette liste ne soit pas exhaustive[61,70] :

- Échantillon de taille limitée : les algorithmes d'apprentissage automatique sont sujets au surapprentissage, donc à un modèle trop adapté au jeu de données d'entrainement
- Surapprentissage : d'autres facteurs peuvent engendrer un problème de surapprentissage. La validation croisée et le *boostrap* permettent de maximiser le potentiel d'un jeu de données, mais une validation externe est trop peu souvent réalisée, ce qui est le principal élément permettant de vérifier la généralisabilité d'un modèle[63]
- Représentativité et homogénéité des données : de la même manière, les données d'entrainement ne sont souvent pas représentatives de la population, ce qui est responsable d'un biais de représentation et limite la généralisabilité[63,70]
- Données inexactes ou manquantes, responsable d'un biais de mesure
- Mauvais étiquetage des données, pouvant produire un biais d'annotation
- Biais algorithmique, lié à l'algorithme en lui-même

Il faut ajouter à cela les questions d'explicabilité et d'interprétabilité des algorithmes. La confiance d'un médecin clinicien dans un algorithme est dépendante de la compréhension qu'il a de son mécanisme et des facteurs identifiables influençant la réponse, ce qui n'est pas toujours évident avec l'apprentissage automatique[71–73].

Etat de l'art des algorithmes prédictifs dans le domaine de l'HTA secondaire

L'utilisation de l'IA dans le domaine de l'HTA est comme dans de nombreuses spécialités en plein essor [27,30,32,74–76].

En ce qui concerne le dépistage de l'HTA secondaire, plusieurs modèles ont été développés à ce jour.

Des modèles simples comme des nomogrammes intégrant des données cliniques et biologiques simples existent, comme développé par Wang et al, et par Liu et al. [77,78]. Le modèle prenait comme variables le sexe, le potassium sérique, le produit Calcium-Phosphore sérique, et le pH urinaire. Les performances étaient assez limitées. Le second modèle prenait en entrée les antécédents d'hypokaliémie, de nodules surrénaliens, le potassium urinaire des 24 heures, et l'activité rénine plasmatique. Les performances étaient bien supérieures, mais il faut noter que l'activité rénine plasmatique n'est pas facilement accessible en médecine de ville et intervient en général après l'étape de dépistage. Dans les deux cas aucune validation externe n'a été réalisée.

Rossi et al.[79] ont développé il y a quelques années un modèle pour le dépistage des adénomes de Conn, mais, utilisant la mesure de l'aldostérone plasmatique, moins accessible en médecine de ville, dans une population sélectionnée car suspecte d'HAP. Cependant une validation externe a été réalisée, avec d'excellentes performances.

Des scores numériques ont également été proposés à partir de données simples, mais avec des performances insuffisantes dans l'études de Yamashita et al [80]. Le score développé par Kietsiroje et al.[81] Démontrait quant à lui une VPN très intéressante (98,8%) en validation interne, mais a été développé sur une population sélectionnée (suspecte d'HAP) et aucune validation externe n'a été réalisée.

Diao *et al.*[82] ont développé des modèles prédictifs sur des données issues des dossiers de santé informatisés, en utilisant l'algorithme XGBoost, pour prédire quatre causes d'hypertension artérielle (rénovasculaire, HAP, dysthyroïdies, et coarctation aortique) mais sans validation externe des modèles. Il faut également noter que certaines des variables utilisées pour la prédiction entraient dans la définition du diagnostic d'HTA secondaire (par exemple le degré de sténose d'artère rénale au scanner pour l'HTA rénovasculaire), ce qui est un biais de confirmation majeur.

Buffolo et al.[28] ont développé un score conventionnel et un modèle d'apprentissage automatique pour prédire la probabilité d'HAP, sur une cohorte de plus de 4000 patients. Ces outils ont permis d'identifier des patients à risque d'HAP, et d'HAP latéralisé. Les modèles prenaient en entrée le sexe, la PAS, le nombre de traitements, l'IMC, le potassium, et les atteintes d'organe cible. Il faut noter que dans ce modèle, le diagnostic d'HAP était

confirmé suite à un test de charge en sel, et seulement suspecté en cas de ratio aldostérone/rénine élevé. Les performances étaient remarquables sur la validation interne, avec un dépistage positif de seulement 3.2% et un diagnostic confirmé dans seulement 0.6% des cas dans le groupe le moins à risque pour le modèle d'apprentissage, mais avec une perte de performance importante sur une cohorte de validation externe, avec un dépistage positif dans 25.6% des cas et un diagnostic confirmé dans 5.5% des cas.

Enfin, Reel et al.83 ont analysé des données multi-omiques (microARNs plasmatiques, métabolites plasmatiques et urinaires) provenant d'une large cohorte de patients afin de distinguer les HTA secondaires d'origine endocrine (HAP, PPGL, Cushing) de l'hypertension primaire. Ils ont utilisé différents algorithmes d'apprentissage automatique, avec une performance intéressante (précision 92 %, sensibilité 88 % et spécificité 96 % pour les random forests). Cependant l'article ne présente pas de validation externe.

Figure 13 – algorithmes prédictifs d'HTA secondaire

Etude	Population	Méthodes	Variables d'intérêt	Cible	Résultat	Limites
Rossi et al. , Clinical Endoc, 1998	574 patients suspects d'HAP : développement : 206 (Padoue, Italie) validation interne prospective : 48 (Padoue, Italie) Cohorte de validation externe: 320 patients (Reggio d'Emilia, Italie)	Analyse discriminante multivariée	Aldostérone plasmatique aldostérone post-captopril activité rénine plasmatique couchée potassium	Adénome de Conn confirmés par biochimie, TDM, KTVS, +/- anapath si surrénalectomie	validation prospective : sensibilité 100%, spécificité 80,9% validation externe : sensibilité 100%, une spécificité 89,6%	Variables plus difficilement accessibles Etape post-diagnostic d'HAP
Diao et al. JMIR Medi Inf, 2021	7532 patients EHR Fuwai, Chine	Apprentissage automatique supervisé (XGBoost)	79 variables (Imagerie, biologie, antécédents, traitements)	HAP, HTA rénovasculaire, sténose aortique, dysthyroïdie	AUC 0.924 pour le dépistage d'HTA secondaire	Biais de confirmation : Variables entrant dans la définition du diagnostic Absence de validation externe
Liu et al., Hypertens Res, 2023	404 patients hypertendus dépistés pour HAP Nanjing (Chine)	Régression logistique multivariée Construction d'un nomogramme Evaluation de l'AUC par bootstrap	antécédents d'hypokaliémie, nodules surrénaliens, potassium urinaire des 24 heures, activité rénine plasmatique	HAP diagnostiqué sur rapports rénine et aldostérone	AUC 0.89 (développement) et 0.86 (validation)	Absence de validation externe Variables non accessibles en pratique clinique simple.

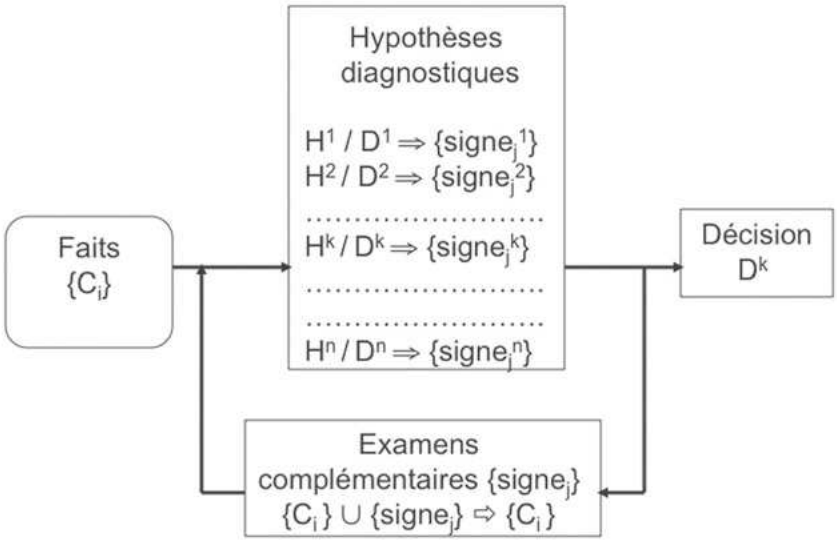
Yamashita et al., J Hypertens, 2018	165 patients : Développement : 130 Validation: 35 (Sapporo, Japon)	score numérique conventionnel par régression multivariées	Âge, sexe, sévérité HTA, potassium, sodium – Chlore, acide urique, pH urinaire	HAP confirmé par test de charge en sel	AUC 0.73, score plus performant que l'hypokaliémie seule	Absence de validation externe
Kietsiriroje et al., Clin Endoc, 2020	420 patients suspects d'HAP Songkla (Thaïlande)	score numérique conventionnel par régression multivariées	Âge, IMC, diabète, nombre d'antihypertenseurs, sodium, potassium	HAP confirmé par test de charge en sel	AUC 0.87, VPN 98,8% si score < 4	Absence de validation externe Population sélectionnée
Wang et al., Endocrine, 2021	806 patients hypertendus dépistés pour HAP Xinjiang (Chine)	Régression logistique multivariée Construction d'un nomogramme Evaluation de l'AUC par bootstrap.	sexe féminin, potassium sérique, produit Ca-P sérique, pH urinaire	HAP confirmé par test de charge en sel	AUC 0.73 (développpement) et 0.68 (validation)	Pas de validation externe
Buffolo et al. , Hypertension, 2022	4059 patients hypertendus à Turin (Italie), Validation externe 584 patient (Munich)	Score numérique conventionnel par régression multivariées et méthodes d'apprentissage automatique supervisé	sexe, pression artérielle systolique, traitement antihypertenseur, IMC, potassium le plus bas, atteintes d'organes	HAP confirmé par test de charge en sel	cf supra.	Résultats validation externe au second plan
Reel et al., eBioMed, 2022	487 patients hypertendus ou non Cohorte Européenne ENS@T	Apprentissage automatique supervisé (Random Forest)	409 caractéristiques multi-omiques (métabolites, miRNA, stéroides,…)	HTA endocrine (HAP, Cushing, PPGL)	AUC 0.95, Sensibilité 90%, Spécificité 86%. Pas d'impact de l'âge et du sexe	Absence de validation externe

Aide à la décision – aide au diagnostic

La démarche diagnostique est un processus basé sur la mise en relation de données cliniques ou paracliniques (biologie, imagerie, anatomopathologie) et des connaissances.

La démarche classique hypothético-déductive consiste à formuler des hypothèses à partir de données observées, puis à essayer de confirmer le diagnostic, éventuellement à l'aide d'examens complémentaires. Cette démarche peut par ailleurs être itérative avec l'apparition de nouvelles données.

Figure 14 - raisonnement hypothético-déductif



Source : Informatique médical, e-Santé [84]

Les systèmes d'aide à la décision diagnostique peuvent intervenir à plusieurs niveaux, comme l'accès aux données, la construction du raisonnement clinique et l'élimination des diagnostics différentiels, ou la prescription d'examens complémentaires.

Les scores cliniques sont très utilisés en médecine, et notamment dans le domaine de l'hypertension artérielle, des scores ont également été développés pour l'aide au diagnostic d'hyperaldostéronisme primaire [28,77,81,85,86].

De tels outils sont cependant à manier avec précaution. En effet, en dehors de problèmes méthodologiques récurrents comme l'absence de validation externe et donc de

généralisabilité, ils doivent être considérés comme des outils d'aide, et non des prédicteurs du diagnostic final.

Dans le domaine de l'intelligence artificielle, une des principales causes évoquée à la limitation des modèles prédictifs est le fait qu'ils se concentrent sur le diagnostic final, au lieu de guider les cliniciens dans un cheminement diagnostic en les aidant à interpréter les données et en leur proposant une marche à suivre [87]. La confiance dans un système d'aide à la décision est un facteur majeur pour l'utilisation d'un modèle par le clinicien. La fiabilité intervient, mais aussi la transparence du modèle[88]

On peut donc considérer qu'un modèle prédictif acceptable est un système d'aide à la décision, axé sur le processus plus que sur le diagnostic final, et intégré dans une démarche diagnostique globale.

Par ailleurs, les enjeux de ces modèles en médecine sont importants, et doivent tenir compte du contexte. En effet, les attentes du clinicien seront tout à fait différentes selon la gravité de la maladie, sa fréquence, et le profil du patient. Les métriques utilisées ont ainsi une importance capitale dans le développement d'un modèle[89]. Dans le domaine de l'hypertension artérielle secondaire, dans le cas du dépistage du phéochromocytome, on cherchera à maximiser la sensibilité et la valeur prédictive négative de l'algorithme, car un faux négatif peut avoir des conséquences importantes, la maladie pouvant être très grave. A l'inverse, dans le cas d'un sténose artérielle rénale, un faux négatif aura le plus souvent des conséquences moins dramatiques. Cependant il faut garder en tête que notre objectif ici est le dépistage, c'est-à-dire une première étape avant la confirmation diagnostique. Dans ce contexte, on privilégie toujours la sensibilité et la valeur prédictive négative, puisque les faux positifs seront réévalués par le test diagnostic final. Il faut cependant trouver le bon équilibre entre la minimisation des faux négatifs, et la réalisation d'une quantité importante d'examens inutiles.

L'approche que nous avons cherché à développer est donc finalement une approche globale d'aide à la décision, afin de pouvoir élaborer des stratégies globales, et de permettre au clinicien de s'approprier les outils que nous avons développé, et de l'intégrer dans son processus de décision.

Chapitre 3 : Construction d’une base de données de patients hypertendus

Dans cette partie, nous décrirons la constitution de la base de données des patients hypertendus de la base de l'HEGP. Nous commencerons par décrire l'entrepôt de données de l'HEGP, puis détaillerons les différentes étapes ayant abouti à la constitution de la base.

Description de l'entrepôt de données de l'HEGP

Données médicales à l'Hôpital Européen Georges Pompidou

L'entrepôt de données de l'HEGP[90–93] est le premier entrepôt de données cliniques conçu en France. Sa conception a démarré en 2007, et les premières utilisations datent de 2010. L'entrepôt intègre une variété des données collectées dans le cadre du soin à partir de 2000 (date de l'ouverture de l'HEGP). L'entrepôt intègre également des données collectées avant l'ouverture de l'HEGP dans des bases spécialisées. Les données médicales à l'HEGP sont stockées au sein d'un entrepôt de données de type I2B2.

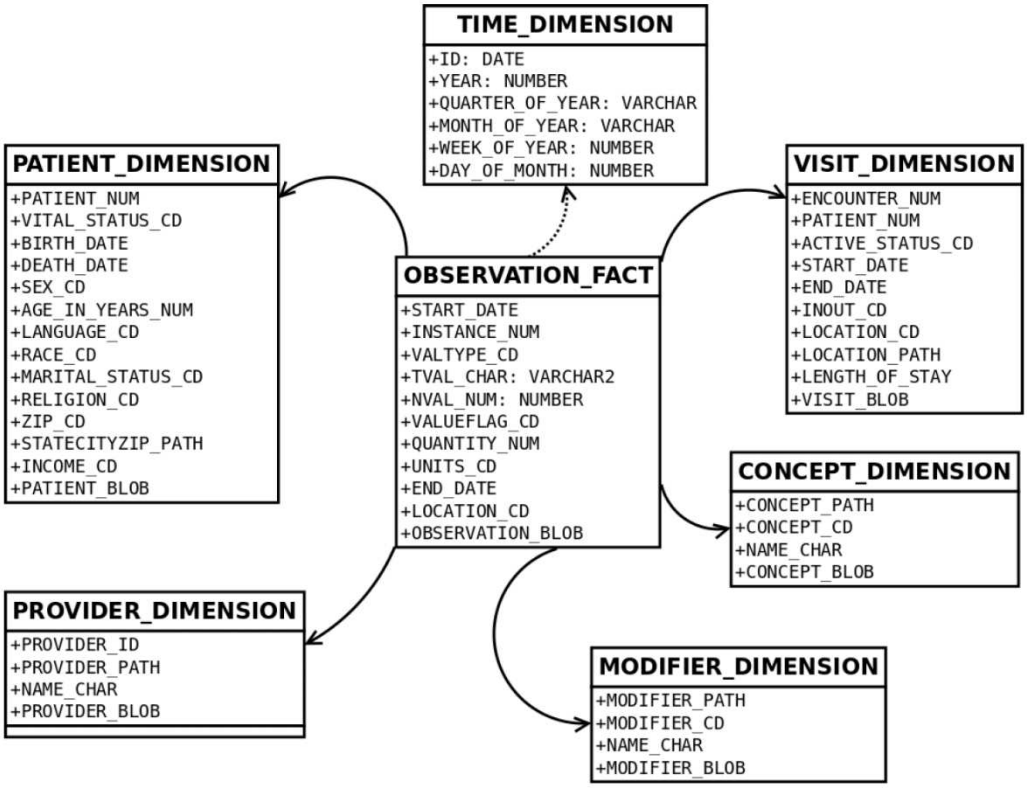
Les données accessibles sont de plusieurs types, et stockées dans un schéma en étoile caractéristique des entrepôts utilisant la modélisation de Kimball[94] (Figure 15) :

- Comptes-rendus d'hospitalisation, de consultation, d'examens
- Codes CIM-10 (Classification Internationale des Maladies - 10ème version) et CCAM (Classification Commune des actes Médicaux), issus du PMSI (Programme de Médicalisation des Systèmes d'Information)
- Questionnaires structurés
- Données démographiques
- Prescriptions de médicaments

L'accès aux données se fait via un serveur Oracle, sous la forme de requête SQL.

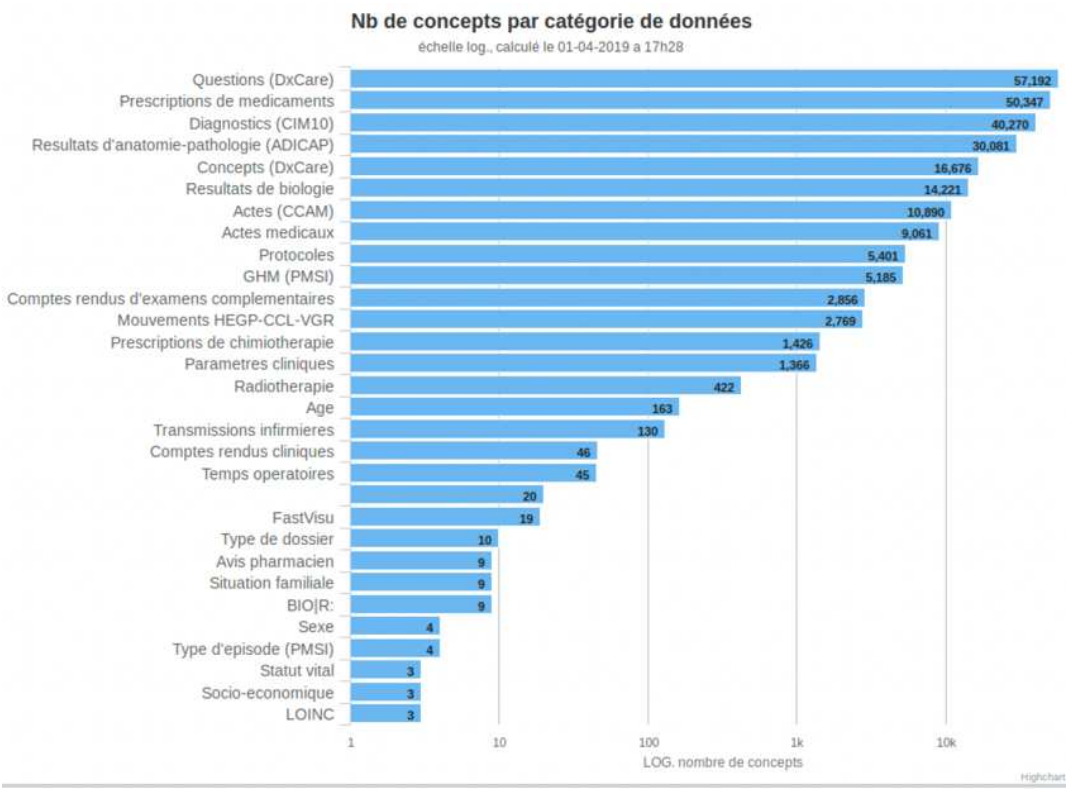
Les données sont déidentifiées, les identifiants étant cryptés.

Figure 15 - données stockées dans l'entrepôt de l'HEGP



Source : document de formation interne

Figure 16 - quantité de données disponible selon le type



Source : document de formation interne

Ci-après deux exemples de requête SQL :

- Extraction des noms et numéros de question composant le questionnaire de consultation du service d'HTA :

```
SELECT concept_cd, name_char
FROM concept_dimension
WHERE concept_path like '\\i2b2\\80. Questions (DxCare)\\2. Type Bilans Medicaux\\ 08. OBSERVATION : Méd Vasculaire HTA\\ 01. (389) VISITE\\%
```

- Extraction des données sélectionnées (`tval_char` pour les données catégorielles et `nval_num` pour les données numériques) d'une liste de patients (1ère venue), pour une liste de questions (QST_RESTREINT), entre 2007 et 2020 :

```
SELECT concept_cd, patient_num, tval_char, nval_num, start_date
```

```
FROM observation_fact
WHERE patient_num in (select * from
PATIENT_1ERE_VENUE)
AND start_date > TO_DATE('20070101','yyyymmdd')
AND start_date < TO_DATE('20200501','yyyymmdd')
and concept_cd in (select* from QST_RESTREINT)
```

Questionnaires d'hypertension artérielle

Le service d'hypertension artérielle a implémenté depuis les années 1975 à l'hôpital Broussais et à l'hôpital St Joseph à Paris un questionnaire structuré permettant de renseigner les données des patients de consultation ou hospitalisés[95,96]. Les données issues de ce questionnaire sont implémentées dans l'entrepôt I2B2 depuis le transfert à l'HEGP en 2000.

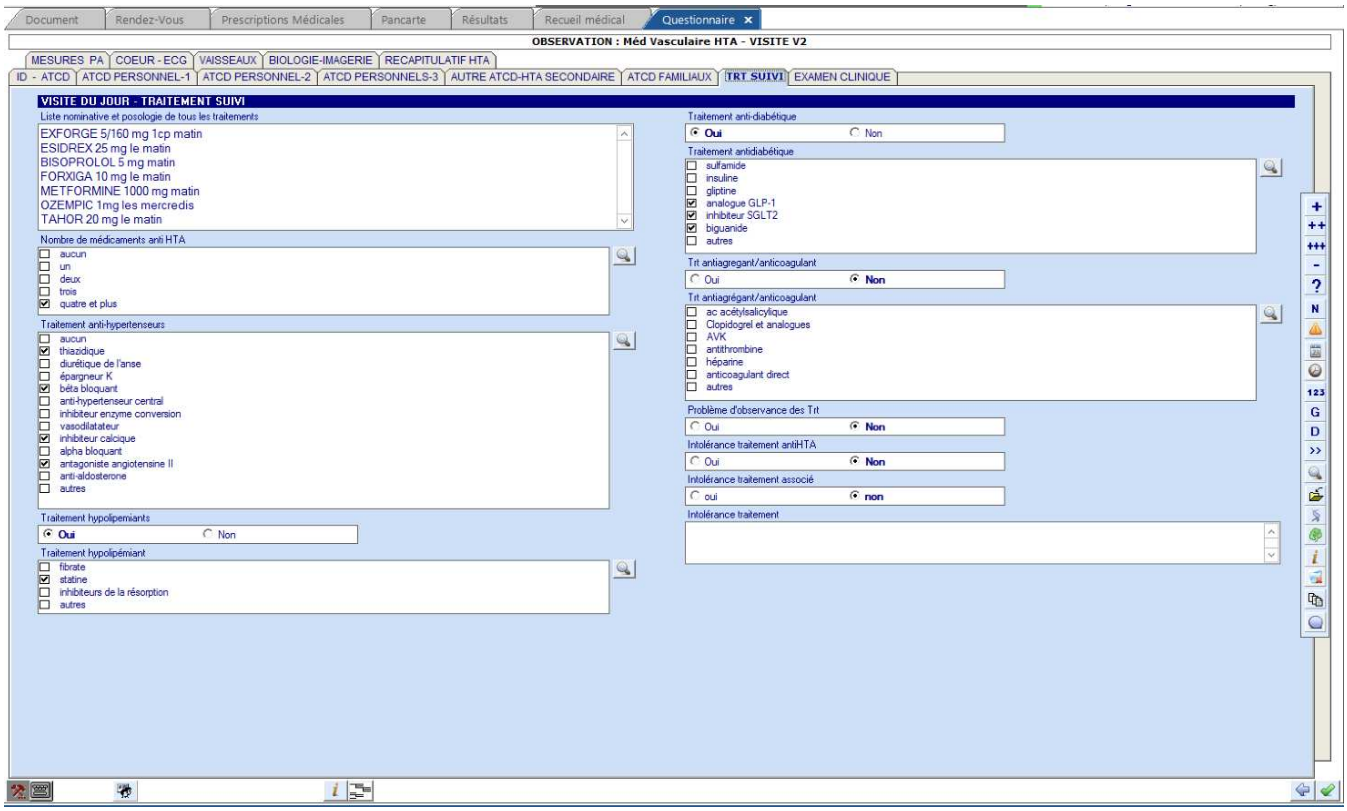
Pour des raisons de qualité et de cohérence des données (le questionnaire a changé plusieurs fois de forme), et comme détaillé dans les parties suivantes, nous n'avons utilisé les données de l'HEGP qu'à partir de 2005.

Les données sont majoritairement structurées :

- Booléens (oui/non)
- Choix multiples
- Nombres

Ils contiennent néanmoins parfois du texte libre (Figure 17).

Figure 17 - questionnaire d'hypertension artérielle



Méthodes d'identification des diagnostics d'hypertension artérielle secondaire

Identification à partir de données textuelles

L'objectif de ce travail était le développement d'une méthode d'identification, des diagnostics d'hypertension artérielle, des patients hypertendus de l'HEGP à partir de comptes-rendus textuels.

Ce travail a été réalisé par Madame Violaine Courrier, dans le cadre d'un stage de science des données de l'ENSAE, effectué dans le service d'informatique médical de l'HEGP en 2024, et que j'ai encadrée.

Deux méthodes ont principalement été explorées, une première basée sur une reconnaissance d'entité classique, utilisant les expressions régulières (RegEx) et un dictionnaire regroupant les termes associés aux différents diagnostics, et une seconde utilisant des méthodes d'apprentissage profond, en l'occurrence des réseaux de neurones.

La reconnaissance d'entité textuelle est appelée reconnaissance d'entités nommées, ou Named-Entity Recognition (NER)

Les modèles ont été évalués sur un échantillon de comptes-rendus annotés manuellement.

Exemple de compte-rendu textuel avec annotation :

> « Cette patiente de 69 ans a déjà été staffée à plusieurs reprises pour sténose athéromateuse de l'artère rénale gauche traitée par angioplastie sans pose de stent en Septembre 2016, puis par angioplastie avec pose de 2 stents en Juillet 2017. Elle a une insuffisance rénale chronique avec un DFG à 25 mL/min, une hypertension connue depuis 2004, une exposition au tabac à 80 PA sevré depuis 2016, un diabète de type 2, une dyslipidémie.
>
> Ces nuisances se sont compliquées d'AVC ischémiques multiples et d'AOMI ayant nécessité une angioplastie de l'aorte abdominale. Elle est revue dans le service pour surveillance de la sténose artérielle rénale. Sous Metoprolol 200, Indapamide LP 1.5, Amlodipine 10, Ramipril 10, Rilmenidine 1, Eurelix 6 sa MAPA diurne est à 168/50 mmHg. Sa kaliémie est à 5.1 mmol/L, son DFG à 23 mL/min (créatinine à 186 µmol/L) avec protéinurie à 0.16 g/mmol. Le LDL et la glycémie a jeun sont à 2.15 et

6.4 mmol/L, respectivement. L'écho doppler mesure le rein gauche à 78 mm versus 95 mm

En aout 2017 avec un stent partiellement ou totalement occlus et un retentissement d'aval sur la moitié supérieure du rein qui est atrophié tandis que la moitié inférieure ne présente pas de signe de mauvaise vascularisation. Le rein droit est mesuré à 95 mm, stable.

Cette patiente a une maladie cardiovasculaire avancée et diffuse. Elle a été dilatée à 2 reprises d'une sténose artérielle gauche qui n'a permis d'améliorer ni le contrôle tensionnel, ni la fonction rénale. Le bénéfice attendu d'une troisième dilatation est faible compte tenu de la taille du rein inférieure à 80 mm et de la protéinurie supérieure à 0.1 g/mmol. Dans ce contexte on ne prévoit pas de nouvelle dilatation. »

Création d'un dictionnaire

La première étape a été la création d'un répertoire, regroupant les diagnostics d'intérêt, puis la détermination des termes d'intérêt pouvant être associés.

Ci-après un tableau récapitulatif des termes associés aux différents diagnostics, élaborés à partir de l'UMLS (Unified Medical Language System) français, et de l'expérience des médecins du service d'HTA.

La correspondance entre les termes retrouvées et les diagnostics utilisait les expressions régulières (Regex) afin de prendre en compte des variations de ses différents termes (pluriel, accentuation, construction multiple) autour de ces termes, et des annotations complémentaires et le contexte syntaxique (Négation, hypothèse, antécédents familiaux,…).

Ci-après, voici quelques exemples de fonctionnement de la notation :

Correspondance exacte : sténose artérielle rénale, SAR

Variations de terme : sténose des artères, rénales, sténose bilatérale des artères rénales

Annotateurs complémentaires :

- Pas de SAR (négation)
- SAR suspectée (hypothèse)
- SAR chez la mère (antécédent familial)
- Regex : [^.\?\!\;]*?art[eè]res? r[eé]nales?,

Figure 18 - Dictionnaire

Concept	stenose_artere_renale	stenose_arterielle_renale_atheromateuse	Dyspl	infarctus_renal	hyperaldosteronisme_primaire	pheochromocytome_non_secretant	pheochromocy
Terme	\bSt[ée]noses? [^.\?\!\;]*art[ée]res? R[ée]nales?	\bst[ée]nose arterielle r[ée]nale ath[ée]romat...	\bDFM\b	\binfarctus r[ée]nal	\bHyperaldost[ée]ronisme primaire	\bPh[ée]ochromocytomes?[^.;!?]*(non|peu|pas|pa...	\bPh[ée]ochromocytomes?
NaN	\bSAR\b	\bSAAR\b	\bdysplasies? fibromusculaires? Art[ée]rielles?	\bisch[ée]mies? R[ée]nales?	\bHAP\b	\bParagangliomes?[^.;!?]*(non|peu|pas|pauci)[...	\bParagangliomes?
NaN	\bHTA rénovasculaire	\bst[ée]noses? [^.\?\!\;]*art[ée]res? R[ée]nal...	\bdysplasies? fibromusculaires?	\bencoches? R[ée]nales?	\bsyndromes? de Conn	\bPHEO[^.;!?]*(non|peu|pas|pauci)[-]s[ée]cr[é...	\bPHEO\b
NaN	NaN	\bSt[ée]nose ath[ée]romateuse de l'art[ée]re r...	\bfibrodysplasie	\bocclusions? [^.\?\!\;]*art[ée]rer[ée]nale	\bad[ée]nomes? de Conn	\bPGL[^.;!?]*(non|peu|pas|pauci)[-]s[ée]cr[ée...	\bPGL\b
NaN	NaN	\bath[ée]rome st[ée]nosant	\bDysplasies?	NaN	\bHypermin[ée]ralo primaire	\bPPGL[^.;!?]*(non|peu|pas|pauci)[-]s[ée]cr[é...	\bPPGL\b
NaN	NaN	\bSARA\b	NaN	NaN	\bHypermin[ée]ralocorticisme primaire	\bChromaffinomes?[^.;!?]*(non|peu|pas|pauci)[...	\bChromaffinomes?
NaN	NaN	NaN	NaN	NaN	NaN	\bChemodectomes?[^.;!?]*(non|peu|pas|pauci)[-...	\bChemodectomes?

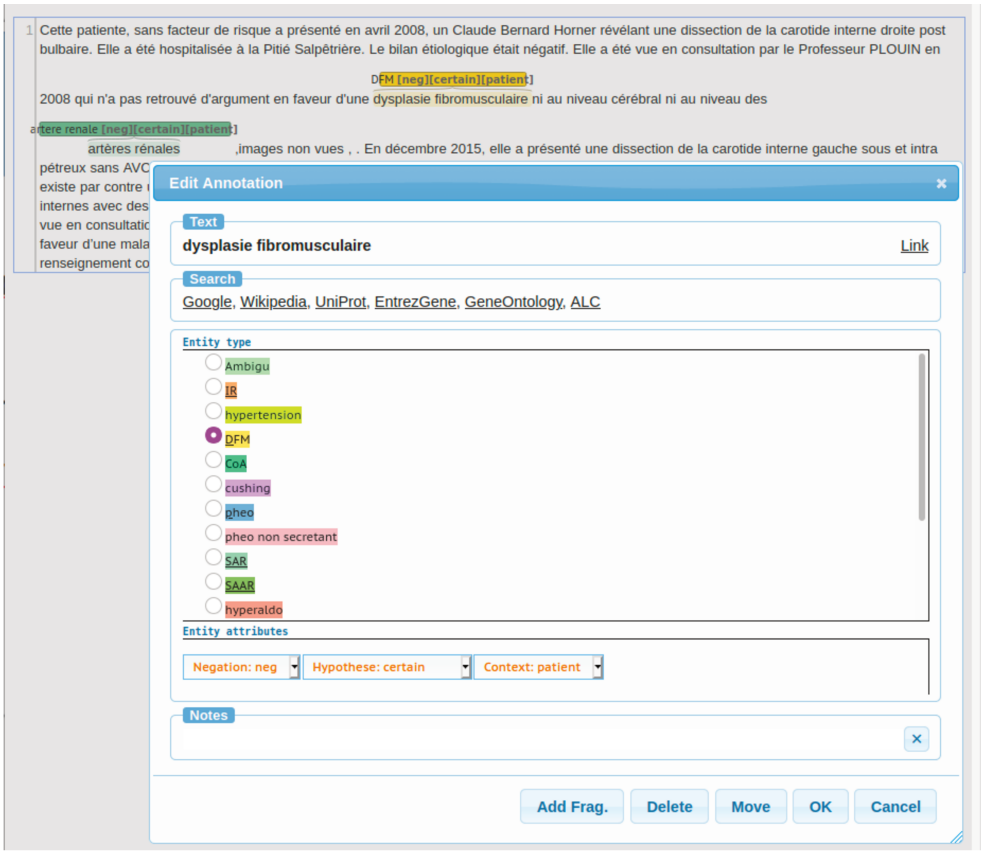
Les textes des différents comptes-rendus sont ensuite annotés grâce à l'annotateur RegexMatcher de Pymedext[97], prenant en entrée le dictionnaire et le texte à traiter, et attribuant en sortie un label (diagnostic) à un *token (*entité lexicale).

Annotation Manuelle

Un corpus de comptes-rendus a été pré-annoté puis vérifié manuellement par des médecins du service d'hypertension artérielle, afin d'avoir à disposition un jeu de données supervisées, pour l'entrainement et la validation des outils de détermination diagnostics.

La pré-annotation a été réalisée à l'aide de l'outil Pymedext (maintenant medkit), puis la vérification manuelle a été réalisée grâce à l'outil BRAT (Figure 19). On retrouve dans cet outil un document entier, avec les entités d'intérêt surlignées et affectées à une catégorie de diagnostic, et des valeurs d'attributs (négation, hypothèse, familial). On voit sur la figure 21 que les mots « dysplasie fibromusculaire » ont été annoté dans la catégorie « DFM » et que ses attributs sont « neg », « certain » et « patient ». On considère donc ici que le document évoque la négation certaine d'un diagnostic de DFM qui concerne le patient.

Figure 19 – outil BRAT.



Reconnaissance d'entités nommées (NER)

Dictionnaire

La première approche était une approche « basée sur les règles », et a consisté en la reconnaissance directe d'expressions régulières, à l'aide du dictionnaire et de l'utilisation de Regex, comme décrit dans le paragraphe précédent « dictionnaire ».

NER spécialisé

Pré-processing

Le jeu de données d'entrainement a été constitué des phrases de comptes-rendus comprenant au moins un diagnostic annoté par et vérifié par les médecins experts.

Chaque phrase a été tokenisée en utilisant le format BIO (Beginning, Inside, Outside), Le préfixe B- devant un token indique qu'il est le début d'une entité d'intérêt, et le préfixe I- indique qu'il est à l'intérieur d'une entité d'intérêt. Enfin, O- indique qu'il n'appartient à aucune entité d'intérêt.

Par exemple, pour la phrase : "Le patient est atteint d'une sténose des artères rénales." :

- Tokenisation : [Le, patient, est, atteint, d, ', une, sténose, des, artères, rénales, .]
- Format BIO : [O, O, O, O, O, O, O, B-SAR, I-SAR, I-SAR, I-SAR, O], où "SAR" est la catégorie du diagnostic et sténose de l'artère rénale l'entité d'intérêt.

Chaque phrase tokenisée sous format bio a été ensuite été normalisée en longueur par une méthode de padding : la longueur des séquences a été fixée à au 95 percentile, les séquences les plus courtes ont été rallongées allongées avec des 'O', et les plus longues ont été tronquées.

Word-Embedding

L'embedding utilisé a été fastText (https://github.com/facebookresearch/fastText)

L'embedding améliore significativement les performances des modèles de NER, tels que BERT ou BioBERT[98].

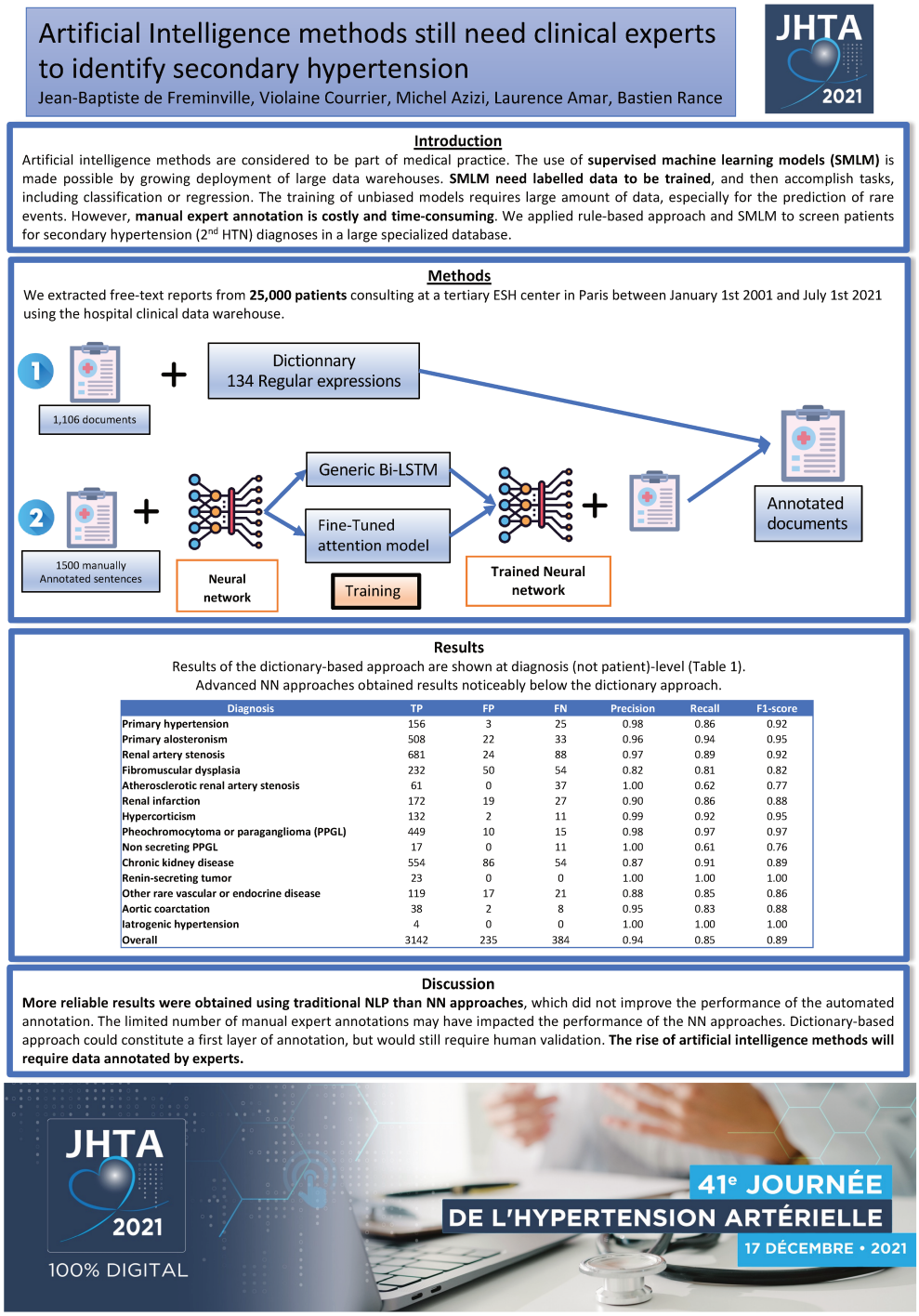
Réseaux de neurones

Les réseaux de neurones développés étaient de type LSTM et bi-LSTM.

Résultats

Les résultats de la reconnaissance par NLP ont fait l'objet d'un poster aux journées de l'Hypertension Artérielle à Paris (Figure 20).

Figure 20 - poster présentant les résultats de la reconnaissance diagnostique par NLP



Dictionnaire

Par l'approche "basée sur les règles" utilisant le dictionnaire, 1106 documents ont été annotés, dont 125 sans diagnostic détecté, soit 981 documents contenant des annotations.

Les résultats sont décrits sur la figure 21.

Figure 21 - performance par catégorie de diagnostic

	TP	FP	FN	Precision	Recall	F1
Apnee_du_sommeil	172	21	27	0.8912	0.8643	0.8776
Coarctation_de_l_aorte	38	2	8	0.9500	0.8261	0.8837
Dysplasie	232	50	54	0.8227	0.8112	0.8169
Hyperaldosteronisme_primaire	508	22	33	0.9585	0.9390	0.9486
Hypercorticisme	132	2	11	0.9851	0.9231	0.9531
Hyperplasie_des_surrenales	119	7	8	0.9444	0.9370	0.9407
Hypertension_iatrogene	4	0	0	1.0000	1.0000	1.0000
Hypertension_primaire	156	3	25	0.9811	0.8619	0.9176
Infarctus_renal	172	19	27	0.9005	0.8643	0.8821
Maladie_rare	119	17	21	0.8750	0.8500	0.8623
Maladie_renale_parenchymateuse	554	86	54	0.8656	0.9112	0.8878
Pheochromocytome	449	10	15	0.9782	0.9677	0.9729
Pheochromocytome_non_secretant	17	0	11	1.0000	0.6071	0.7556
Stenose_artere_renale	681	24	88	0.9660	0.8856	0.9240
Stenose_artere_renale_atheromateuse	61	0	37	1.0000	0.6224	0.7673
Tumeurs_a_renine	23	0	0	1.0000	1.0000	1.0000
Overall	3437	263	630	0.9289	0.8913	0.9097

NER spécialisé

Un corpus de 2977 phrases a été utilisé, dont 63% pour l'entrainement des modèles, 7% pour la validation et 30%, pour le test.

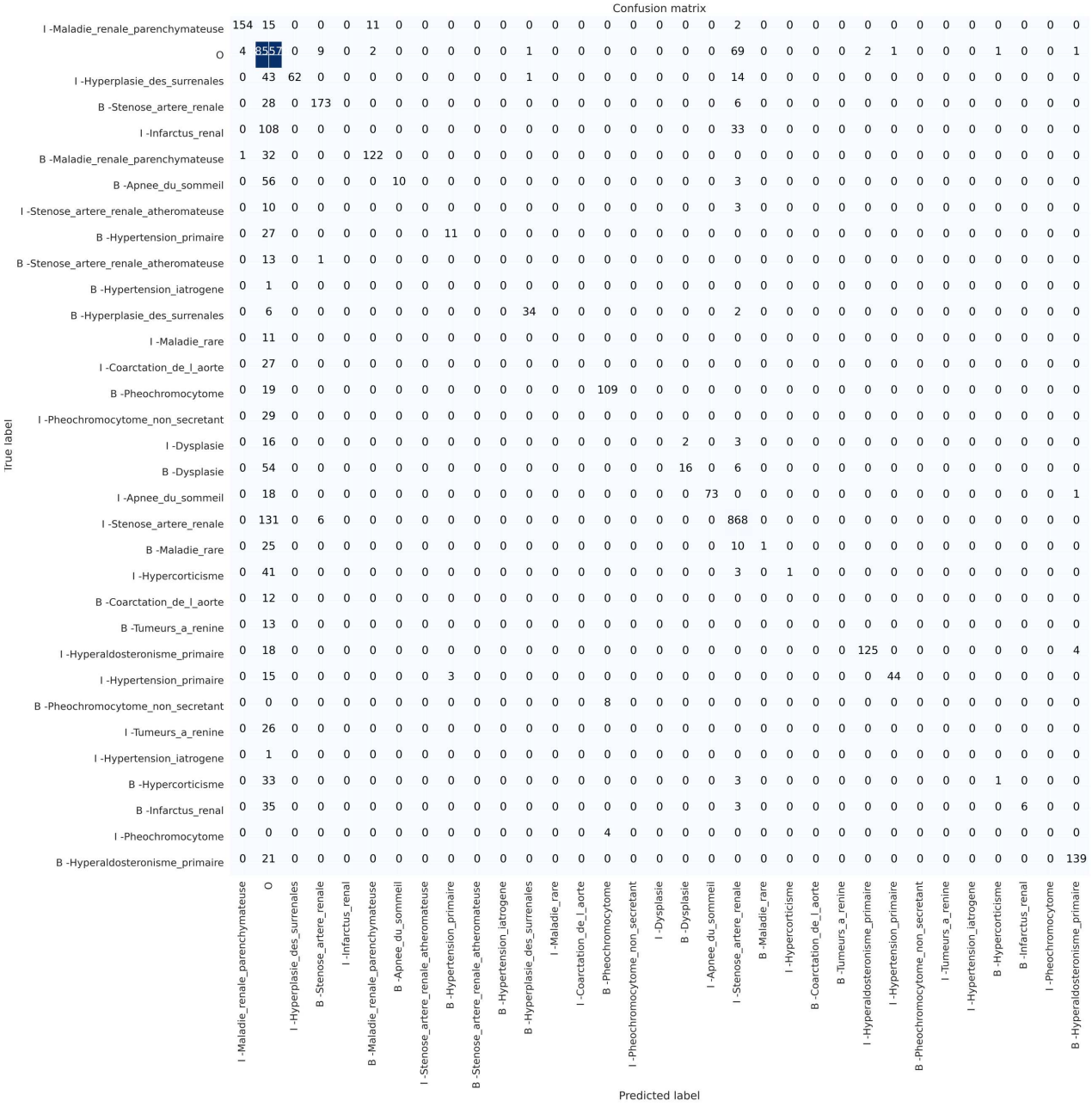
Bi-LSTM Fasttext

Un embedding fastText de dimension 300 pré-entrainé sur des articles provenant de wikipédia. On utilise alors la même architecture que précédemment en ajoutant un couche d'embedding. On teste ensuite la même architecture en bidirectionnelle, c'est-à-dire avec un Bi-LSTM.

La matrice de confusion des résultats est décrite Figure 22.

On note que certaines catégories de diagnostics ne sont pas apprises et sont associées à la catégorie « O ».

Figure 22 - matrice de confusion des résultats du NER spécialisé



Identification à partir du PMSI

Les codes CIM-10 et CCAM ont été utilisé pour retrouver des diagnostics d'HTA secondaire. Pour les patients hospitalisés, les informations du PMSI comprennent le diagnostics principaux et secondaires selon la classification CIM-10, et les actes médicaux

réalisés selon la classification CCAM. Le PMSI a été utilisé à de nombreuses reprises pour la recherche de diagnostics ou pour l'étude des maladies cardiovasculaires ou rénales[99–101].

Une limite de cette approche était la disponibilité de ces données uniquement pour les patients ayant été hospitalisés. Cependant, les bilans d'hypertension artérielle secondaire sont en grande majorité réalisés au cours d'une hospitalisation de semaine ou de jour.

Une autre limite est la possibilité de faux positifs. En effet, les diagnostics sont codés par les médecins au cours de l'hospitalisation, et dans notre expérience certains diagnostics sont codés par excès, alors que le diagnostic est suspecté, mais non encore prouvé. Cela est plutôt à contre-courant de ce qui est habituellement décrit (on observe dans la littérature une tendance au sous-codage[102]). Ce biais non conservateur est probablement lié au centre expert et au fait que le diagnostic soit codé par des médecins particulièrement sensibilisés à ces recherches diagnostiques.

Les codes utilisés pour les diagnostics sont décrits dans les tableaux 1 et 2.

Tableau 1 - Codes CIM-10 associés aux diagnostics

CIM10	DIAGNOSTIC
I10	Hypertension artérielle primaire
E260	Hyperaldostéronisme primaire
I771	Sténose d'une artère rénale
I773	Dysplasie fibromusculaire
N280	Infarctus rénal
E275	Phéochromocytome ou paragangliome
E270	Hypercorticisme
N18	Maladie rénale chronique

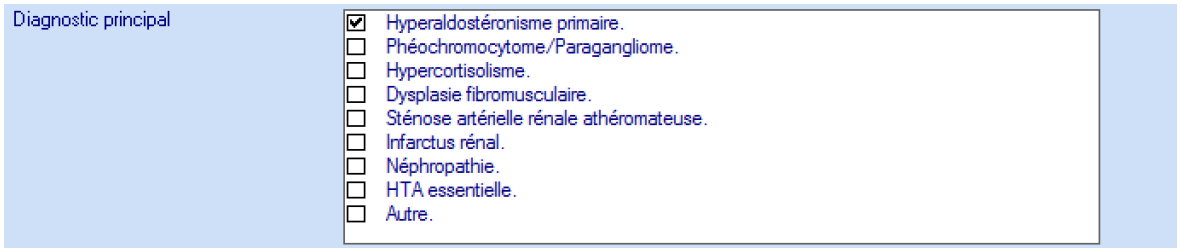
Tableau 2 - codes CCAM associés aux diagnostics

CCAM	ACTE	DIAGNOSTIC ASSOCIE
EGHF003	Prélèvements sanguins bilatéraux étagés ou simultanés des veines surrénales, par voie veineuse transcutanée	Hyperaldostéronisme primaire
EDAF010	Dilatation intraluminale suprasélective de l'artère rénale avec pose d'endoprothèse, par voie artérielle transcutanée	Sténose artérielle rénale
EDAF001	Dilatation intraluminale sélective ou hypersélective de l'artère rénale avec pose d'endoprothèse, par voie artérielle transcutanée	Sténose artérielle rénale
EDAF009	Dilatation intraluminale suprasélective de l'artère rénale sans pose d'endoprothèse, par voie artérielle transcutanée	Sténose artérielle rénale

Identification à partir des questionnaires

Le diagnostic suspecté ou affirmé du patient est renseigné lors de chaque visite du patient (Figure 23).

Figure 23 – diagnostic renseigné sur le questionnaire d'hypertension artérielle



Bases de données existantes

Les bases de données COMETE (Cortex-MEdullary Tumor Environment, base de et ENSAT-HT (European Network for the Study of Adrenal Tumors - Hypertension)[103,104], ont été utilisées pour compléter les diagnostics des patients.

COMETE est une base de données dédiée à l'étude des tumeurs endocrines, en particulier celles de la médullosurrénale (phéochromocytomes) et des paranganglionmes.

ENSAT-HT est une base de données européenne dédiée aux tumeurs des glandes surrénales.

Algorithme retenu pour l'identification du diagnostic final

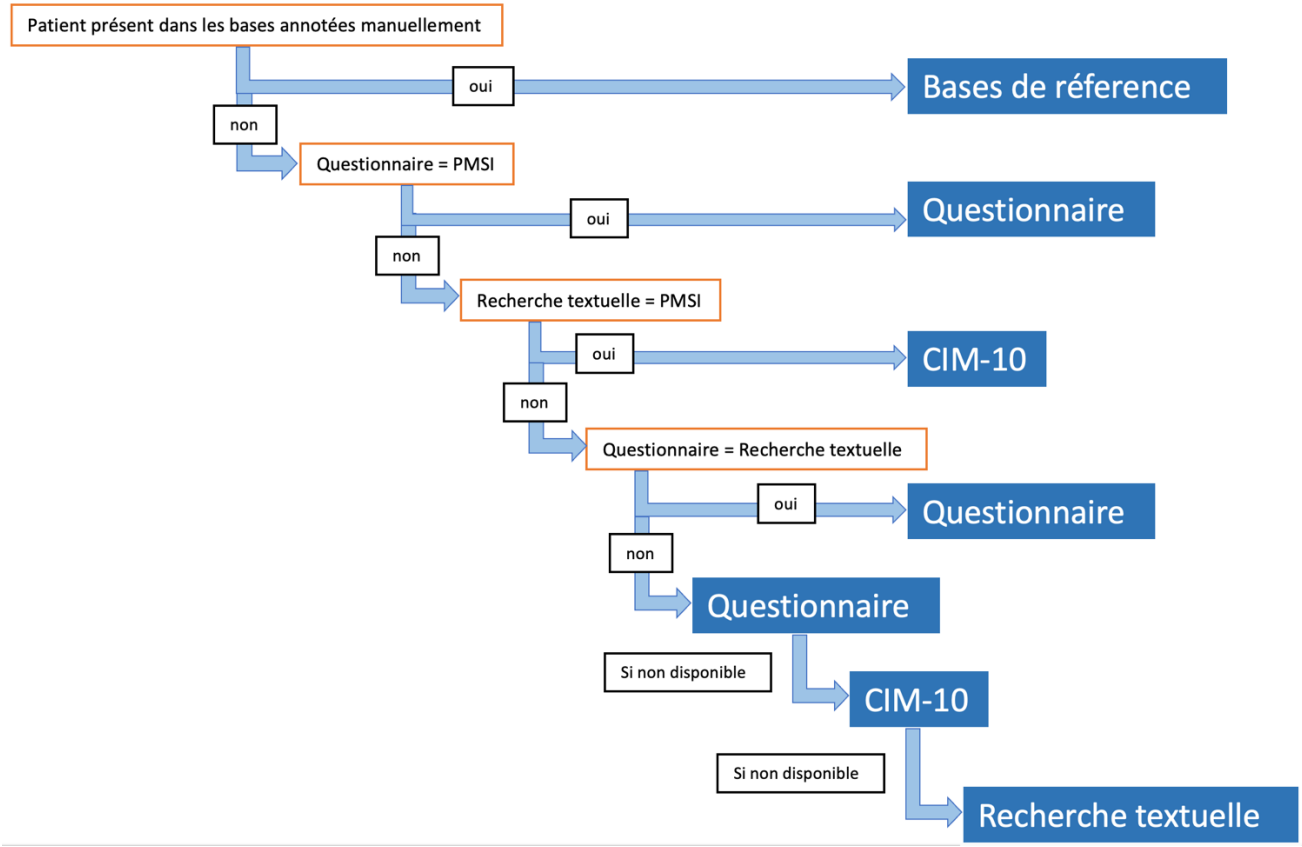
Le diagnostic final retenu pour chaque patient a été défini à partir d'un algorithme favorisant :

- Les bases de données les plus fiables (COMETE et ENSAT-HT)
- Les diagnostics vérifiés manuellement par les médecins experts
- La cohérence entre les différentes sources

Concernant les diagnostics retrouvés à partir des données textuelles, les annotations « basées sur les règles », à partir du dictionnaire et de la recherche textuelle simple à base d'expression régulières ont été retenues.

Le choix du diagnostic final retenu est décrit dans la figure 24.

Figure 24 - algorithme de choix du diagnostic selon les sources disponibles et leur concordance



Mise en forme des bases de données finales

Formatage des données

Les données issues du questionnaire, après extraction, sont présentées sous une forme non exploitable, à savoir une ligne par patient, date, et valeur (Figure 25).

Figure 25 - exemple de présentation des données extraites

	A	B	C	D	E
1	CONCEPT_CD	PATIENT_NUM	NVAL_NUM	TVAL_CHAR	START_DATE
2	QST:QN|2478	xxx		thiazidique	24/10/2016 09:32
3	QST:QN|2411	xxx	46	E	24/10/2016 09:32
4	QST:QN|7579	xxx	135	E	24/10/2016 09:32
5	QST:QN|7797	xxx	49	E	25/07/2016 15:11
6	QST:QN|2449	xxx	11	E	25/07/2016 15:11
7	QST:QN|2509	xxx		Non	19/09/2016 13:37
8	QST:QN|2427	xxx	84	E	19/09/2016 13:37
9	QST:QN|12350	xxx		Non	19/09/2016 13:37
10	QST:QN|2571	xxx	0	E	19/09/2016 13:37
11	QST:QN|2533	xxx			25/07/2016 15:11
12	QST:QN|2407	xxx		Non	24/10/2016 09:32
13	QST:QN|2406	xxx		Non	25/07/2016 15:11
14	QST:QN|2536	xxx		aucun	25/07/2016 15:11
15	QST:QN|2589	xxx		25/07/2016	25/07/2016 15:11
16	QST:QN|2504	xxx		2016	19/09/2016 13:37
17	QST:QN|2406	xxx		Non	19/09/2016 13:37
18	QST:QN|2542	xxx		aucun	25/07/2016 15:11
19	QST:QN|10579	xxx	0	E	19/09/2016 13:37
20	QST:QN|2478	xxx		antagoniste angiotensine II	19/09/2016 13:37
21	QST:QN|10560	xxx			25/07/2016 15:11

Les étapes initiales sont :

- Correspondance entre les numéros de question (`concept_cd`) et leur nom
- Elimination des doublons
- Jonction avec une table comprenant les dates de première venue, et sélection des données correspondant à cette date pour chaque patient
- Séparation des différents types de questions
- One-hot encoding pour les variables binaires
- Pivotement de la table de sorte à obtenir une table avec une ligne par patient, et une colonne par variable

Sélection des patients

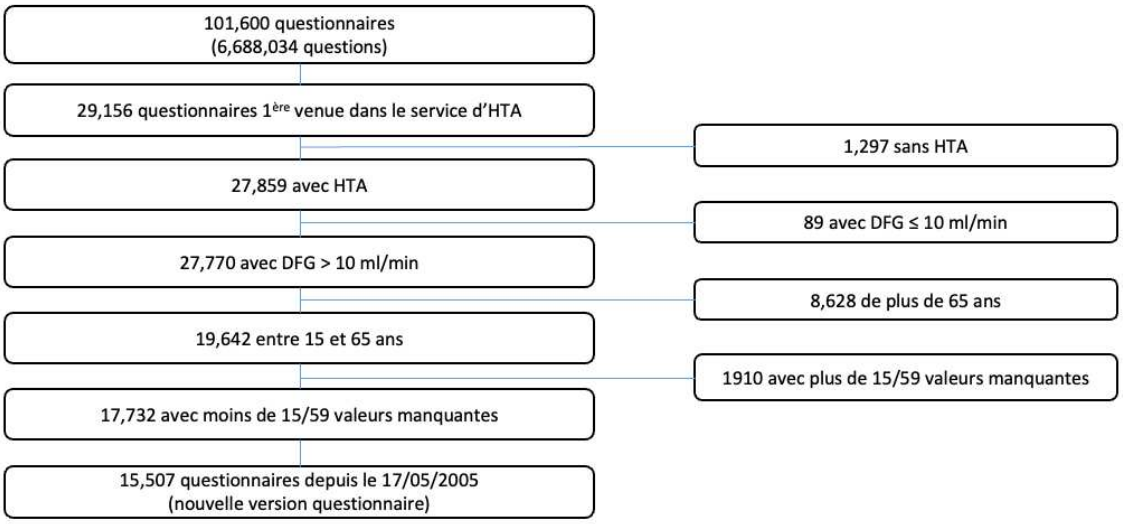
Différentes étapes ont été réalisée avant de pouvoir exploiter la base de données finale.

Les patients exclus étaient :

- Les patients non hypertendus (pression artérielle < 140/90 mmHg et absence de traitements antihypertenseurs)

- Les patients insuffisants rénaux dialysés. Nous avons fixé un seuil arbitraire de débit de filtration glomérulaire (DFG) à 10 ml/min CKD-EPI pour exclure les patients dialysés
- Les patients âgés de plus de 65 ans : nous avons considéré qu'au-delà de cet âge, la recherche d'hyperaldostéronisme primaire et plus largement la recherche d'hypertension artérielle secondaire était rarement réalisée
- Les patients ayant un trop grand nombre de valeurs manquantes : nous avons arbitrairement fixé un seuil de 15 données manquantes sur les 59 variables recueillies initialement
- Les patients ayant consulté avant 2005, du fait de la qualité insuffisante des questionnaires recueillis avant cette date

Figure 26 - Flow Chart patients final des patients inclus dans la base HEGP



Qualité des données

La recherche de valeurs aberrantes a été réalisée à l'aide de diverses représentations graphiques selon la distribution de chaque variable.
Ci-après sur les Figures 26, 27 et 28 des exemples de représentation permettant l'exclusion de ces valeurs aberrantes.

Figure 27 - représentation de l'ancienneté de l'HTA en fonction de l'âge après suppression des valeurs aberrantes. Tous les points au-dessus de la droite identité (ligne rouge) et les valeurs aberrantes ont été exclus

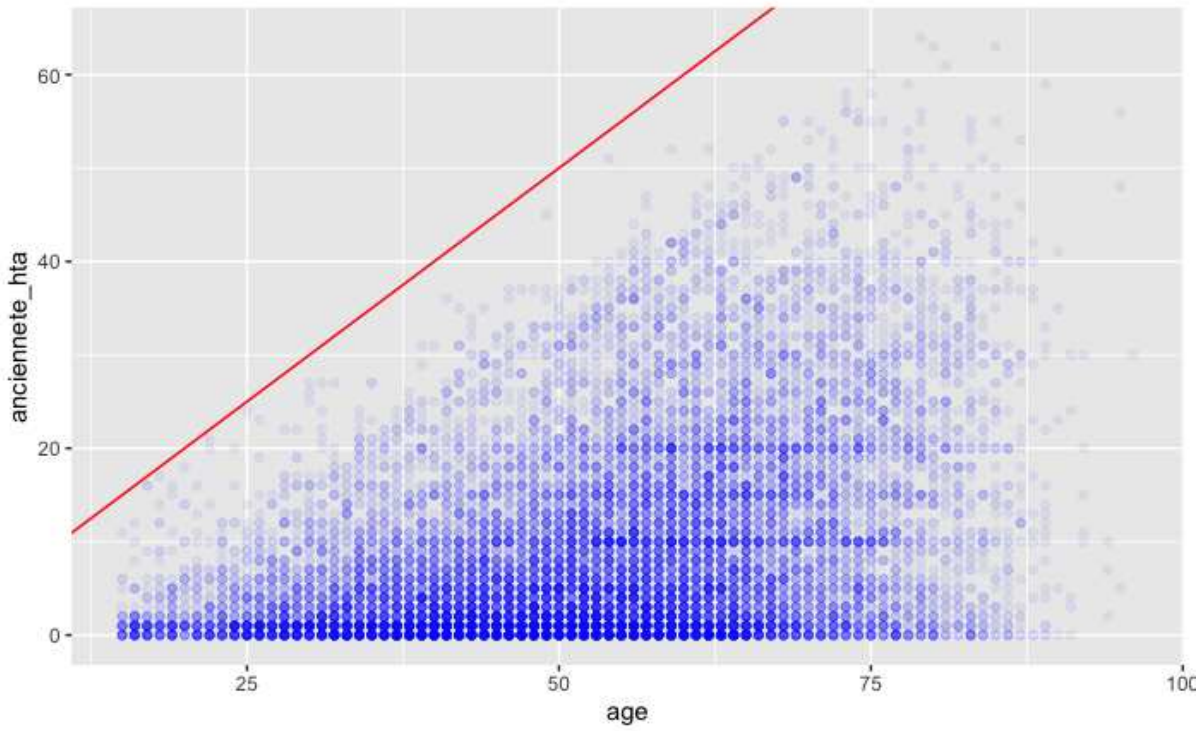


Figure 28 - violin plot de la taille après suppression des valeurs aberrantes

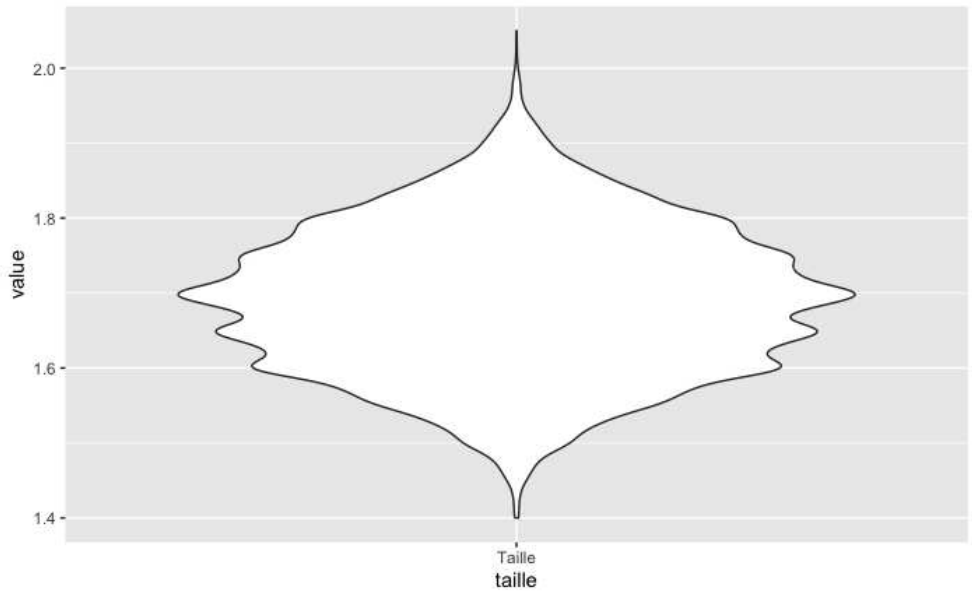
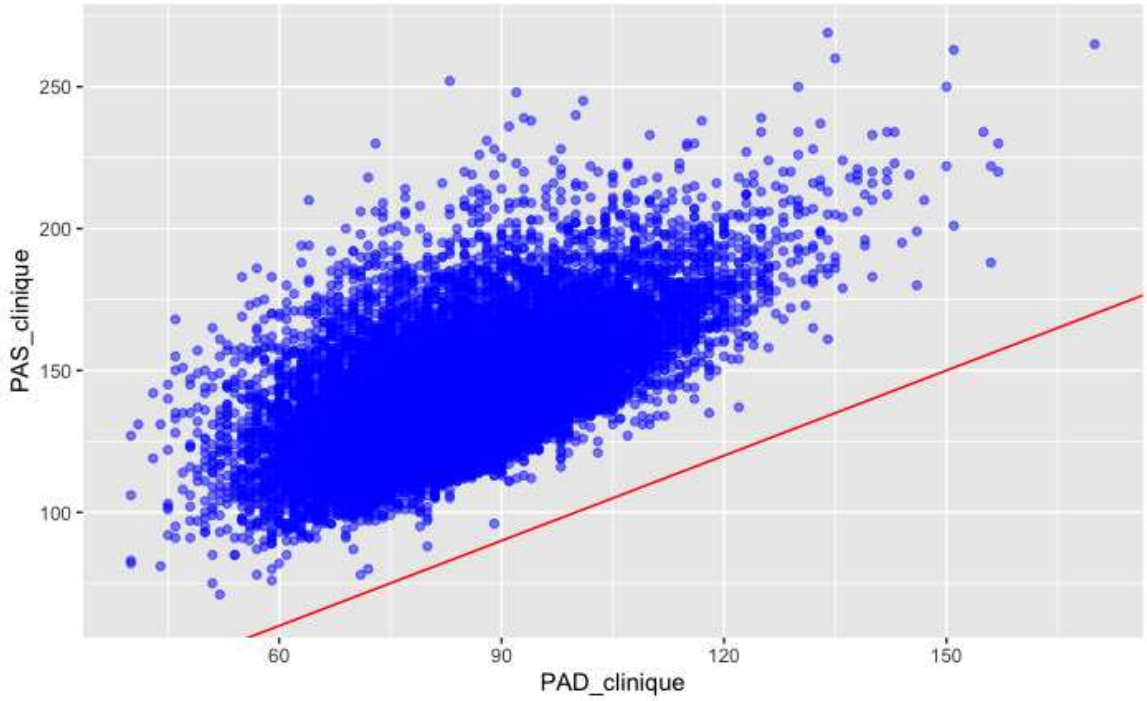


Figure 29 - Représentation de la PA systolique en fonction de la PA diastolique. Tous les points au-dessous de la droite identité (ligne rouge) et les valeurs aberrantes ont été exclus



Sélection des variables

37 variables ont finalement été retenues (tableau 3). Les valeurs manquantes par variables sont représentées sur la Figure 30.

Figure 30 - représentation des données manquantes par variable après binarisation avant sélection de la base finale

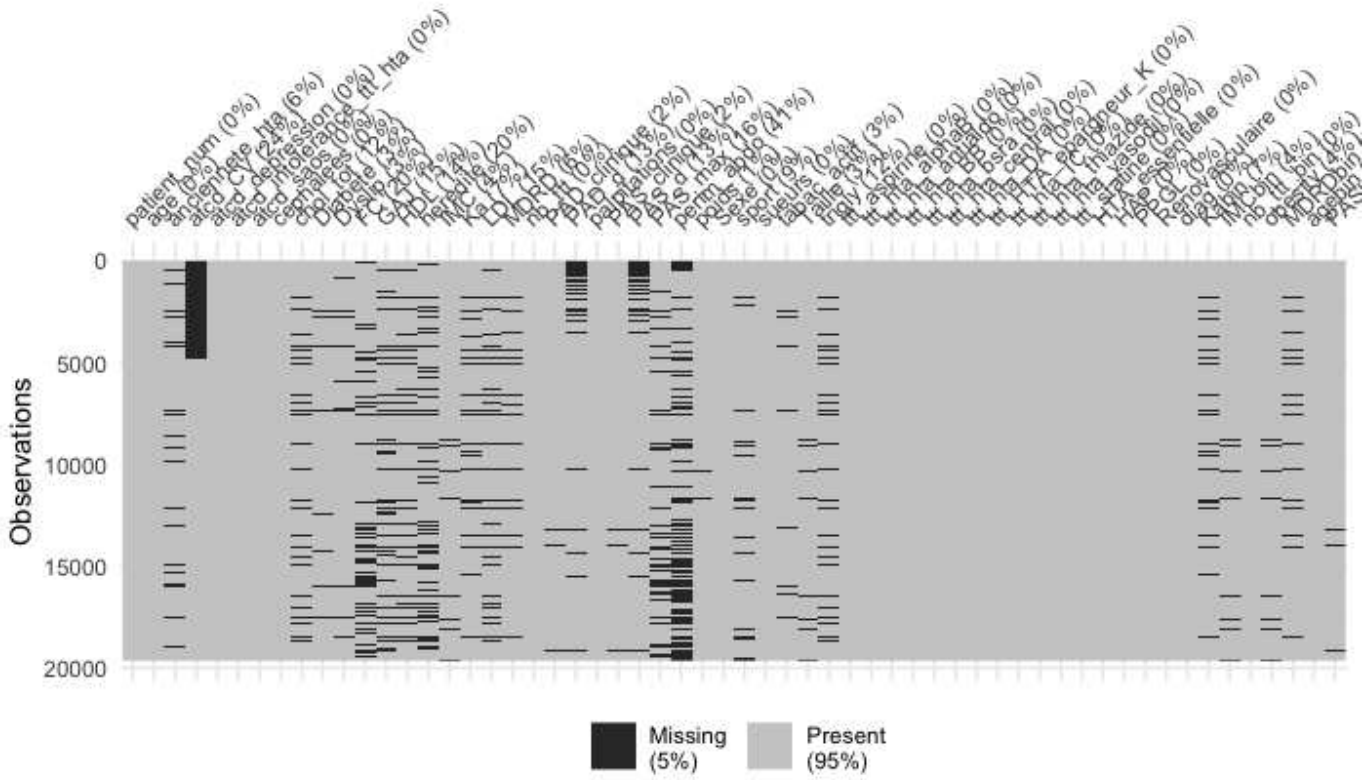


Tableau 3 - variables sélectionnées et correspondance des noms dans la base de de données

VARIABLES	NOM DANS LA BASE DE
AGE	age
SEXE	sexe
ANCIENNETE DE L'HTA EN ANNEES	anciennete_hta
ANTECEDENT FAMILIAL D'HTA	heredite
ANTECEDENT DE MALADIES	atcd_cv
ANTECEDENT DE DEPRESSION	atcd_depression
ANTECEDENT DE SAOS	atcd_saos
DIABETE	diabete
DYSLIPIDEMIE	dyslip
TABAGISME ACTIF	tabac_actif
CEPHALEES	cephalees

SUEURS	sueurs
PALPITATIONS	palpitations
FREQUENCE CARDIAQUE	fc
PRESSION ARTERIELLE	pas_clinique
PRESSION ARTERIELLE	pad_clinique
HTA RESISTANTE	htar
HTA GRADE 3	htagrade3
IMC	imc
TAILLE	taille
KALIEMIE	ka
DFG ESTIME	dfg
CHOLESTEROL TOTAL	chol_tot
HDL CHOLESTEROL	hdl
LDL CHOLESTEROL	ldl
NOMBRE DE MEDICAMENTS	nb_ttt
INTOLERANCE TRAITEMENT	atcd_intolerance_ttt_hta
TRAITEMENT ALPHA-BLOQUANT	ttt_hta_alphab
TRAITEMENT ANTI-	ttt_hta_antialdo
TRAITEMENT BETA-BLOQUANT	ttt_hta_bb
TRAITEMENT IEC OU ARA2	ttt_hta_bsra
TRAITEMENT	ttt_hta_central
TRAITEMENT INHIBITEUR	ttt_hta_ic
TRAITEMENT EPARGNEUR DE	ttt_hta_epargneur_k
TRAITEMENT DIURETIQUE DE	ttt_hta_da
TRAITEMENT DIURETIQUE	ttt_hta_thiazide
TRAITEMENT STATINE	ttt_statine

Le périmètre abdominal a été retenu pour l'algorithme de stratification, mais pas pour les modèles d'apprentissage automatique, du fait de trop nombreuses valeurs manquantes.

Binarisation des variables

Les variables continues ont été binarisées pour le développement de l'algorithme de stratification.

Plusieurs seuils ont été choisis. Les seuils correspondaient selon les variables a des valeurs pertinentes cliniquement, ou aux tertiles des variables dans la base de données de l'HEGP.

Tableau 4 - variables binarisées et seuils

VARIABLES BINARISÉES	SEUIL	SEUIL 2	SEUIL 3
ABDOMINAL PERIMETER IN WOMEN	88.0		
ABDOMINAL PERIMETER IN MEN	102.0		
AGE	40.0	50.0	66.0
BODY MASS INDEX (KG/M2)	25.0	30.0	
DIASTOLIC BLOOD PRESSURE (MMHG)	80.0	90.0	100.0
GLYCEMIA (MMOL/L)	6.1		
HDL CHOLESTEROL IN MEN (MMOL/L)	1.3		
HDL CHOLESTEROL IN WOMEN (MMOL/L)	1.0		
HEART RATE	68.0	79.0	
HEIGHT (M)	1.65	1.74	
HTN DURATION	2.0	8.0	
LDL CHOLESTEROL (MMOL/L)	2.6	4.9	
MAXIMUM SYSTOLIC		200.0	
NUMBER OF TREATMENTS	2.0	3.0	
POTASSIUM (MMOL/L)	3.5	4.0	4.5
SYSTOLIC BLOOD PRESSURE (MMHG)	130.0	140.0	160.0
TRIGLYCERIDES	1.69		

Base de données Bordeaux

La base de données du centre d'excellence en HTA du CHU de Bordeaux, utilisée pour la validation externe de nos modèles, a été implémentée de manière prospective avec les données de patients ayant consulté en hôpital de jour. L'ensemble des patients avaient eu un bilan d'hypertension artérielle complet et l'ensemble des valeurs ont été revues manuellement par l'équipe de Bordeaux.

Aspects technico-règlementaires

Notre étude a été approuvée par le le CERAPHP (Comité d'Ethique de la Recherche APHP Centre).

L'étude étant hors Loi Jardé (recherche sur données rétrospectives), elle n'a pas nécessité l'autorisation d'un Comité de Protection de Personnes (CPP).

Cependant, ce type de recherche nécessite théoriquement une information individuelle des patients inclus afin de rentrer dans la méthodologie de référence MR004[105,106]. Nous avons donc demandé et obtenu une dérogation à l'information individuelle à la CNIL (Commission Nationale de l'Informatique et de Libertés), après avis du CESREES (Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé) et réalisé des affichages des informations de l'étude sur le site internet du service (https://centre-hypertension.org/recherche/etudes-et-protocoles-en-cours/).

L'étude a également été inscrite au registre des traitements de l'APHP (registre recensant les traitements de données prévu par le RGPD (règlement général de protection des données).

Chapitre 4 : Hypertension artérielle secondaire chez le sujet jeune

Introduction

Selon certaines études, la prévalence de l'HTA secondaire serait inversement corrélée à l'âge, atteignant entre 15 et 30 % des jeunes adultes[14,42,107–109].

L'âge en dessous duquel il est recommandé de dépister un HTA secondaire a varié dans les dernières recommandations. Le consensus est actuellement à 40 ans, mais les dernières recommandations proposent de rechercher un HAP chez tout hypertendu, quel que soit l'âge [16–18]. Cela repose en grande partie sur des études rétrospective, avec une hétérogénéité importante dans les populations de ses études. En dehors du fait que la prévalence est plus importante chez les patients jeunes, l'enjeu est également plus important, car ces patients seront exposés plus longuement à des causes pouvant avoir des conséquences chroniques importantes, indépendamment du niveau de pression artérielle. Ainsi, l'hyperaldostéronisme primaire, principal cause d'hypertension artérielle, secondaire (dont il représente environ 10 %) est responsable de toxicités rénale, cardiaque, et vasculaire, liées à l'exposition prolongée à une hypersécrétion d'aldostérone, et ceux indépendamment du contrôle de la pression artérielle[13,47,110].

Cette étude a été réalisée dans des centres experts, ce qui, comme nous le discuterons, peut avoir des conséquences sur la population étudiée du fait d'un biais de recrutement, mais ce qui nous a également permis d'analyser les données d'une population parfaitement phénotypée, ayant eu un bilan complet à la recherche d'une hypertension artérielle secondaire.

Objectifs

L'objectif de ce travail était d'étudier les caractéristiques des hypertensions artérielles secondaires chez les sujets jeunes. Nous avons pour cela analysé les données de patient ayant eu un bilan complet d'hypertension artérielle secondaire à l'hôpital Européen Georges Pompidou.
Le deuxième objectif de l'étude était d'analyser les facteurs associés à l'HTA secondaire autres que l'âge des patients, afin de mieux cibler les examens réalisés dans cette population, à l'aide de méthodes statistiques robustes basées sur le bootstrap. En effet, et comme nous l'avons vu en introduction, la recherche de causes secondaires d'hypertension artérielle peut être coûteuse, et difficile à mettre en place, par exemple pour la réalisation de bilan hormonaux nécessitant des interruptions et modifications des traitements antihypertenseurs.

Présentation du manuscrit

Ce manuscrit a été publié dans la revue *Hypertension :*

Jean-Baptiste de Freminville, Margherita Gardini, Antoine Cremer, Scarlett Camelli, Stephanie Baron, Guillaume Bobrie, Philippe Gosse, et al. « Prevalence and Risk Factors for Secondary Hypertension in Young Adults ». Hypertension 81, nº 11 (novembre 2024): 2340-49. https://doi.org/10.1161/HYPERTENSIONAHA.124.22753.

Prevalence and risk factors for secondary Hypertension in young adults

Jean-Baptiste de Freminville* [a,b]; MD; Margherita Gardini* [a]; Antoine Cremer, MD, PhD [c]; Scarlett Camelli [a,d]; MD; Stephanie Baron [e,f]; Guillaume Bobrie [a,g]; Philippe Gosse [c]; Romain Boulestreau [h]; Nicole Gebara [a]; Julien Doublet [c]; Thomas Dussartre [c]; MD; Christine Grataloup [i]; Aurélien Lorthioir [a]; Christine Massien [a]; Anne-Marie Madjalian [a]; Julien Riancho [a]; Gilles Soulat [i]; Nicolas Postel-Vinay [a]; Michel Azizi [a,j]; Bastien Rance [b,k,l]; Laurence Amar [a,m,n]

*contributed equally to this work

a Hypertension Unit, Université Paris-Cité, AP-HP, Hôpital Européen Georges Pompidou, Paris, France

b INSERM, UMRS 1138, CRC, Université Sorbonne-Paris Cité, Paris, France

c Hypertension Excellence Center, University Hospital of Bordeaux, Hopital Saint André, Bordeaux

d Hopital Privé Seine Saint Denis, Drancy

e Physiology department, AP-HP, Hôpital Européen Georges Pompidou, Université Paris-Cité, Paris, France

f Inserm,. France

g Service de Néphrologie Hémodialyse – Hôpital Privé du Vert Galant – Ramsay Générale de Santé – Tremblay en France, France

h Service des Maladies Coronaires et Vasculaires, CHU de Bordeaux

i Radiology department, AP-HP, Hôpital Européen Georges Pompidou, Université Paris-Cité, Paris, France

j Inserm CIC 1418, F 75015, Paris, France

k INRIA France

l Medical Informatics, Biostatistics and Public health department, AP-HP, Hôpital Européen Georges Pompidou, Université Paris-Cité, Paris, France

m Adrenal referral center, Université Paris Cité, AP-HP, France

n Inserm Paris Cardiovascular research center, 75015 Paris, France

Short Title: Secondary hypertension in young adults

Corresponding author

Laurence AMAR, MD,PHD

Hypertension Unit, Hôpital Européen Georges Pompidou, 20 rue Leblanc, 75908 Paris cedex 15, France

Fax 33156093791, Tel 33156093771

email address: laurence.amar@aphp.fr

Word count: 6104

ABSTRACT

Background: The prevalence of secondary causes of hypertension (HTN) in young adults is unknown, and therefore there is no consensus regarding the indication of screening of secondary hypertension (2HTN) in this population. The objective was to report the prevalence and the causes of 2HTN in young subjects.

Methods: 2090 patients with confirmed hypertension aged 18 to 40 years with full work-up for 2HTN screening, were included in this Cross-sectional study. We assessed the prevalence of secondary hypertension and analyzed factors associated.

Results: 619/2090 patients (29.6 %) had a 2HTN. The most frequent diagnoses of 2HTN in descending order were PA (n = 339, 54.8%), renovascular HTN (n = 114, 18.4%), primary kidney disease (n = 80, 12.9%), pheochromocytoma/functional paraganglioma (n = 37, 5.9%), HTN caused by drugs or substances (n = 32, 6.0%) and other diagnoses (n = 17, 2.7%). Patients with BP < 160/100 mmHg did not have a lower prevalence of 2HTN regardless of the number of treatments. The prevalence of 2HTN was higher in the decade between 30 and 40 years old than between 18 and 30 years old (p=0.024). Female sex, hypokalemia, a treatment with at least two medications, no familial history of hypertension, BMI < 25kg/m², and diabetes, were associated with a higher prevalence of 2HTN.

Conclusions: Prevalence of 2HTN is high among young hypertensive patients (29.6% in our cohort), regardless of the age and of BP level. All hypertensive patients under 40 years old should be screened for secondary causes.

Keywords : Hypertension, Hypokalemia, Body Mass Index, Pheochromocytoma, Hyperaldosteronism, Diabetes Mellitus, Renovascular Hypertension

ABBRÉEVIATIONS

ABPM : Ambulatory Blood Pressure Monitoring
AIC : Akaike Information Criteria
ARR : Aldosterone-to-Renin Ratio
BP : Blood Pressure
CHU BDX : Bordeaux University hospital
CKD : Chronic Kidney Disease
CV : Cardiovascular
eGFR : Estimated glomerular filtration rate
HBPM : Home Blood Pressure Monitoring
HTN : Hypertension
2HTN : Secondary causes of HTN
HEGP : Georges Pompidou Hospital
LCMSMS : Liquid Chromatography with tandem Mass Spectrometry
CKD-EPI : Chronic Kidney Disease Epidemiology Collaboration
OBP : Office BP
OOBP : Out of Office BP
OSAS : Obstructive Sleep Apnea Syndrome
PAC : Plasma Aldosterone Concentration
PRC : Plasma Renin Concentration
RIA : Radioimmunoassay

INTRODUCTION

Hypertension (HTN) affects more than one billion people worldwide [6–8] and represents one of the main risk factors contributing to cardiovascular (CV) morbidity and mortality [9]. Its prevalence increases with age but it is also common among younger populations affecting approximately 1 in 8 adults aged between 20 and 40 years with a range between 2 and 40% [7,10,11]. Despite the availability of multiple classes of antihypertensive treatments, blood pressure (BP) control rates remain poor worldwide for various reasons including non-adherence to treatment, therapeutic inertia, low socioeconomic status, and poor access to medications or healthcare [12]. Besides, undetected secondary causes of HTN (2HTN) may also contribute to uncontrolled and resistant hypertension. They also may be responsible for increased organ damages independently of blood pressure level [13].

However, the true prevalence of 2HTN remains not well defined and depends on the population studied. It is estimated to be around 5-10% in the general hypertensive population, which is probably underestimated [14]. Regarding young hypertensive patients, it is commonly accepted that the prevalence of 2HTN might be higher but the real prevalence is unknown as studies on 2HTN prevalence in younger hypertensive adults remain scarce and mainly focused on children and adolescents [1,4,15–17].

The 2018 ESH/ESC[16] guidelines considered that a patient with grade 2 HTN younger than 40 years should "raise the suspicion" of a 2HTN. Regarding the 2020 ISH guidelines [18], the authors suggested to "consider" 2HTN in patients younger than 30 years old in the absence of obesity or familial history of hypertension. The recent 2023 ESH guidelines [17], endorsed by the ISH, confirmed that patients younger than 40 years should raise the suspicion of 2HTA, but only when they have grade 2 HTN. The underdiagnosis of 2HTN is an important health issue, even more in young patients, because specific treatment or cures can avoid a long exposure to HTN. However, the benefit of a generalized screening for 2HTN in young patients is yet to be proven.

We report in this study the prevalence and the underlying causes of 2HTN in young hypertensive subjects using computerized electronic health records. We also looked for predictors of 2HTN diagnoses in young adults.

METHODS

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Patient selection and data retrieval

We performed a cross-sectional study on patients consecutively referred to the Georges Pompidou Hospital (HEGP) Hypertension Unit and to the Bordeaux University hospital (CHU BDX) Hypertension Unit. Main reasons for referral are either uncontrolled HTN in primary health care, suspicion of 2HTN or initial workup after HTN diagnosis.

In HEGP, a standardized, computerized medical record has been used in daily practice for both clinical care and research purposes since 1975. It is interfaced with the patients' hospital electronic medical record, and completed by physicians during each consultation. It is divided into blocks of questions (medical history, biochemical parameters, etc.) [19–21]. These dedicated medical records are stored on an i2b2 platform in a clinical data warehouse together with all other hospital health records. Data were extracted regarding the first out-patient and following visits from all patients with confirmed HTN aged 18 to 40 years between 01/01/2007 to 31/05/2020 who underwent a full work-up (including standardized hormone determinations and abdominal CT- or MR- angiograms) to look for 2HTN from this dedicated computerized medical record using the clinical data warehouse interface.

In CHU BDX, the clinical data of interest are collected by the Clinical Research Associates from the patients' electronic medical record via a form containing various sections (demographics, medical history, treatment, etc.). The form then implements a password-protected ACCESS database. The data collected concerns all patients consulting for an etiological assessment of hypertension and aged between 18 and 40 years between 01/01/2019 and 31/01/2023. The etiological work-up includes biochemical parameter, hormonal assessment and abdominal CT scan.

The current study was approved by the institutional review board (Regional Ethics Committee) no. 00001072; project no. CDW_2014_0014). French legislation does not request written consent from patients, given the retrospective nature of the data collected. To ensure complete reporting of our routinely collected health data, we followed the REporting of studies Conducted using Observational Routinely-collected health Data statement [22]. We also followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) recommendations [23].

Data extraction:

Different types of variables were extracted in HEGP, including those self-reported by the patient, those retrieved from records provided by the patients' referring physicians (e.g. medical history, ongoing treatments, clinical signs), those measured during the out-patient visit (e.g. office BP, out-of-office BP, weight, height, clinical examination), and biochemical parameters (plasma creatinine, potassium, cholesterol, blood glucose, proteinuria/albuminuria).

In CHU BDX the variables reported were age, medical history (e.g. cardiovascular risk factors, duration of hypertension), current treatment (including contraception if any), weight, height, office BP, laboratory work-up (creatinine, eGFR, potassium, aldosterone, plasma renin, LDL, HDL, total cholesterol, TSH, PTH, proteinuria) and out-of-office BP.

Estimated glomerular filtration rate (eGFR) was calculated using the Chronic Kidney Disease Epidemiology Collaboration formula (CKD-EPI). Chronic Kidney Disease (CKD) was defined by eGFR of less than 60 mL/min/1.73m² and/or albuminuria/creatininuria > 30 mg/g. Hypokalemia was defined by a plasma potassium concentration of less than 3.5 mmol/L.

The antihypertensive treatments were stratified into different categories, including renin-angiotensin system blockers (angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers), calcium channel blockers, beta-receptor blockers; diuretics (thiazide or thiazide-like diuretics, loop diuretics), potassium sparing diuretics, ~~and~~ mineralocorticoid receptor antagonists, alpha-receptor blockers, centrally acting antihypertensive drugs, and direct vasodilators. Regarding the number of antihypertensive drugs, for patients taking single-pill combinations of multiple drugs, those drugs were counted separately.

We obtained age at diagnosis of HTN, family history of HTN, current smoking status, type I or type II diabetes, dyslipidemia, overweight (BMI ≥ 25 kg/m2), and diagnosis of apparent resistant HTN from the extracted variables of the medical record.

a) Blood pressure

In both centres, office BP (OBP) was recorded as the average of three measurements taken by trained nurses after 5-minute rest in the lying position with a validated semiautomatic oscillometric device [24]. Out of office BP (OOBP) was defined as either daytime Ambulatory Blood Pressure Monitoring (ABPM) if available, or Home Blood Pressure Monitoring (HBPM). HBPM consisted in twice daily BP measurements at home during 3 to 7 consecutive days with three measurements 1-minute apart in the sitting position after a 5-minutes rest in the morning before drug intake and breakfast and in the evening

after dinner. For 24-hour ABPM, BP was measured every 15 minutes during daytime (from 07AM to 10PM), and every 30 minutes during night-time (from 10PM to 07AM) [25]. HTN was defined as OBP above or equal 140/90mmHg or HBPM/daytime ABPM above or equal 135/85 mmHg or by regular prescription of antihypertensive medication(s) [17]. Resistant HTN was defined by inadequate control of BP on HBPM or ABPM in patients with optimal or best-tolerated doses of three or more drugs (at best including a diuretic, typically an ACE inhibitor or an ARB, and a CCB [17]).

b) Diagnostic work-up for 2HTN

We measured plasma hormone for endocrine HTN after discontinuation of antihypertensive medication interfering with the renin-angiotensin system and correction of hypokalemia. Diuretics, ACE inhibitors, ARB, beta blockers were withdrawn for 2 weeks at least and MR blockers for 6 weeks. When withdrawal of all antihypertensive drugs was not possible, patients received drugs with limited ~~the~~ interactions with the concentrations of renin and aldosterone [26]. Plasma aldosterone concentration (PAC) and plasma renin concentration (PRC) were measured in standardized conditions with blood collected in the morning after 30 minutes rest in semi-recumbent position.

The diagnostic criteria for primary aldosteronism (PA) follow the current recommendations [27,28]. As we published previously, it was based on the combination of (1) an Aldosteron-to-Renin Ratio (ARR) above the threshold on at least 2 occasions plus (2) an elevation of aldosterone secretion based on a basal PAC, post-saline infusion test PAC or an elevated urinary aldosterone concentration [29]. The Plasma Renin Concentration (PRC) threshold was set to a minimum of 5 mUI/L for the ARR calculation to avoid overestimation of the ARR in subjects with an undetectable PRC. In HEGP, PAC and free urine aldosterone concentration at pH1 were measured by radioimmunoassay (RIA) (Coat-A-Count; Siemens Medical Solutions Diagnostics, Saint-Denis, France) until 2016-03-15 and subsequently after 2016 by Liquid Chromatography coupled to tandem Mass Spectrometry (LCMSMS) since then [30]. The ARR and the PAC cut off changed after 2016 with the use of Liquid chromatography-tandem mass spectrometry as we reported previously [30]. PRC was measured by chemiluminescent immunoassay (LIAISON; Diasorin, Sallugia, Italia). In CHU BDX, PAC and PRC were also measured by chemiluminescent immunoassay (LIAISON; Diasorin, Sallugia, Italia) and the cut-offs were the same than used in HEGP before 2016. The cut-off used for the saline infusion test was 140 pmol/l following the guidelines.

Pheochromocytoma and functional paraganglioma were diagnosed by the presence of a tumor associated with an excess of urinary fractionated metanephrines and/or plasma free metanephrines measured by LCMSMS, as recommended by the current guidelines [31,32]. The diagnosis was confirmed by functional imaging (^{18}F-DOPA PET CT) and/or pathology for all patients.

Cushing syndrome was diagnosed by an increase in 24-hour urine cortisol excretion, high midnight cortisol concentration and post-dexamethasone (1mg) plasma cortisol concentration above 50 nmol/L [26]. Plasma and urine cortisol was determined by chemiluminescent immunoassay (LIAISON; Diasorin, Salugia, Italia).

Renovascular causes included hemodynamic renal artery stenosis (mostly fibromuscular dysplasia), renal infarction, and renal artery dissections detected by non-invasive imaging [33]. The stenosis was considered significant in the presence of hemodynamic stenosis on renal artery Doppler, or improvement of blood pressure control following vascular intervention.

HTN caused by drugs or substances was systematically investigated by interrogation.

Renal disease even if it cannot be cured was also part of the screening for secondary hypertension. It was defined as CKD not defined as caused by HTN [34].

Less common etiologies of 2HTN, defined as "other diagnoses", were renin tumor, aortic coarctation or monogenic arterial disease with mid-aortic syndrome.

With the exception of CKD and HTN caused by drugs or substances, the formal diagnosis of all forms of 2HTN was systematically ascertained during a multidisciplinary concertation meeting. For the purpose of this study, the medical files of all retrieved patients were checked individually by two of the authors (MG and SC) for final diagnosis of primary or secondary hypertension.

Statistical analysis

Results are expressed as percentages or means ± standard deviations. Groups (2HTN vs. primary HTN) were compared, using unpaired Student's t tests or Wilcoxon nonparametric test for continuous variables and chi-squared tests or Fisher's exact test for qualitative variables.

To analyze the impact of each variable on the probability of having 2HTN, we performed univariate and multivariate analysis using logistic regressions. For the multivariate analysis, we used stepwise regression based on Akaike Information Criteria (AIC) analysis.

Odds ratios (ORs) with their 95% confidence interval (95% CI) were calculated for each variable included in the multivariate model.

We performed a bootstrap resampling method to examine the distribution of the probability for the absence of 2HTN, according to the presence of the previously identified predictors, based on multivariate logistic regressions on 1000 bootstrapped samples of our population.

P-values < 0.05 were considered significant.

Analysis were performed using the statistical software RStudio (Version 4.0.3)

RESULTS

Characteristics of the population

The flow chart is presented in Figure 1. Included patients were consecutively referred. After exclusion of patients with incomplete screening, with HTN diagnosed during childhood and of patients without a confirmation of HTN, 2040 patients were included in the final analysis.

Mean age at diagnosis of HTN was 29.8 years. A majority of patients had a familial history of HTN (59.6%). Few patients had diabetes (4.7%), but most patients had BMI > 25 kg/m² (66.9%) (Table 1).

619 patients (29.6 %) had confirmed diagnosis of 2HTN. Of note, 218/619 (35.2%) with 2HTN, had no clinical or biochemical criteria for 2HTN screening according to the current guidelines (non-resistant HTN, no grade 3 HTN, normal plasma potassium). Besides, among patients with no other indication to screen for 2HTN than the age of diagnosis of HTN before 40 years, the prevalence of 2HTN was still high (22.9%).

The most frequent diagnoses of 2HTN in descending order were PA (n = 339, 54.8%), - among which 104 were aldosterone-producing adenomas (30.7%), and 154 (45.4%) were bilateral hyper aldosteronisms -, renovascular HTN (n = 114, 18.4%), - including fibromuscular dysplasia (n = 85, 4.1%), atherosclerotic renal artery stenosis (n=17, 2.7%), and rare causes (Takayasu disease, renal artery hematoma…) - primary kidney disease (n = 80, 12.9%), pheochromocytoma/functional paraganglioma (n = 37, 6.0%), HTN caused by drugs or substances (n = 32, 5.2%), which were mostly oestroprogestative contraceptive (N = 18), and other diagnoses (n = 17, 2.7%) (Figure 2).

Prevalence of 2HTN in different subgroups

The prevalence of 2HTN was higher in the decade between 30 and 40 years old than between 18 and 30 years old (32.4 vs 27.2%, p=0.024) (Figure 3). This was mainly due to the higher prevalence of primary aldosteronism (Figure 3, Supplemental figure 1).

Regarding the level of hypertension, as shown in Figure 4, the prevalence of 2HTN remained high in all patients, even those with BP < 160/100 mmHg, regardless of the number of treatments.

Regarding the sex, 2HTN was more frequent in women than in men (34.4 vs 24.3%, p < 0.001) (Supplemental table 1). PA and renovascular causes mostly were more frequent in women (Supplemental figure 2). Hypokalemia was more frequent in women (21.3 vs 15.3%, p = 0.008). Otherwise, women were thinner, had less dyslipidemia and tobacco consumption, and lower office and ambulatory systolic blood pressure, while diastolic blood pressure was comparable. Resistant HTN was more frequent in men.

Comparison between patients with 2HTN and those with primary HTN

The prescription of more antihypertensive medications and the presence of hypokalemia at the first visit were positively and significantly associated with the diagnosis of 2HTN in univariate and multivariate analysis (Table 2). Conversely, male sex, BMI ≥ 25 kg/m² and family history of HTN were negatively associated with the diagnosis of 2HTN. Resistant HTN and eGFR < 60 ml/min/1.73m^2 at the first visit were positively associated with the diagnosis of 2HTN in univariate, but not in multivariate analysis. Of note, diabetes was associated with a higher risk of 2HTN in multivariate analysis. The presence of other cardiovascular risk factors or the OBP levels at the first visit, were not associated with the diagnosis of 2HTN (Table 2).

We estimated the probability of having 2HTN with bootstrapped logistic regressions (Supplemental figure 3), according to 6 predictors (plasma potassium concentration at the first visit, sex, number of antihypertensive medications, BMI, familial history of HTN, and diabetes).

DISCUSSION

We found that the prevalence of 2HTN is higher among young hypertensive patients aged less than 40 years than estimated in the whole population (29.6% vs around 10% [14]). Moreover, the prevalence of 2 HTN in patients aged from 30 to 40 years is most likely at least as high as in patients aged from 18 to 30 years. 2HTN prevalence was even more frequent in women treated with a higher number of antihypertensive drugs, with hypokalemia, diabetes, and normal BMI.

These results suggest that all patients with hypertension diagnosed between 18 and 40 years old should be screened for secondary HTN. The 2018 ESH and 2020 ISH guidelines [16,18] were conflicting regarding

the age cut-off. The recent 2023 ESH guidelines for the management of HTN [17], endorsed by the ISH, settled this argument by suggesting that patients under 40 years old with grade 2 or 3 HTN should be screened for 2HTN but without any reference to support this statement. In our cohort, 29.9% of patients with secondary hypertension would not have been screened according to those guidelines, because they had no other indication than the age of diagnosis of HTN to screen for secondary hypertension (i.e. no resistant hypertension, no grade 3 hypertension, no hypokalemia), and HTN grade 1. Our study therefore goes further and suggests that all patients with HTN under 40 years old should be screened regardless of the blood pressure level. This study also shows that diabetes is associated with 2HTN. One explanation might be the known association of diabetes with PPGL as catecholamines do increase blood glucose[35] . And PA has often been associated with a metabolic profile and diabetes[36].

2HTN is of major importance in young patients. Indeed, early treatment of an underlying cause can normalize BP, and limit the exposure to HTN complications, and specific organ damages linked to 2HTN. This is even more important in young women of childbearing age. We confirmed in our study that 2HTN is more frequent in women, mostly from PA or renovascular causes, and the consequences on the pregnancy can be dramatic [37–39].

Comparison to other studies

To our knowledge, our study comprises the largest cohort of hypertensive young adults who underwent detailed screening for the most common causes of secondary hypertension reported in the literature.

It confirmed the result of previous studies. Trifanescu et al. report a series of 80 patients younger than 40 years screened for endocrine hypertension and the prevalence of endocrine etiology identified was 20% [40]. Lower rates of 2HTN among children and younger adults were published by Yoon on 653 patients aged 18 to 21, in which 2HTN prevalence was 16% [41]. Another study reported a 2HTN prevalence of 33% but among a smaller sample (148 hypertensive patients younger than 40 years referred to their centre). As in our study, PA was the most common etiology and 2HTN patients were mainly women. However, a higher rate of HTN caused by drugs or substances was reported (mainly oral contraceptives) [2].

A large retrospective Korean study on young hypertensive men (19-29 years old) and reported a very low prevalence of 2HTN of 2.2%. However, all patients did not have a complete screening, and secondary hypertension was identified using ICD-10 codes [42].

In a belgian study, focusing on fibromuscular dysplasia, the prevalence of FMD was 5.7% among the 35 hypertensive women younger than 40 years old who underwent a computed tomography angiography or a magnetic resonance angiography[43]. The results are similar in our cohort as the prevalence of FMD is 4.1%.

Other studies did not specifically evaluated the prevalence of secondary hypertension. In the RESIST-POL, cohort among resistant hypertensive patients, the prevalence was 24% with a repartition equivalent to our study among young patients, with more than 50% of the patients with PA[44]. However in a study from Monticone et al. on the prevalence of PA in primary care practice, the prevalence was still elevated, but lower than in patients referred to the hospital. The age of patients with PA was older than the mean age of the “non-PA” patients[45]. Even if it did not focus on young hypertensive patients, it might underline the risk of bias due to the fact that in our cohort patients are seen in tertiary centers.

Strengths and limitations

The number of patients, the complete and systematic screening for each patient, and the multicentric design represent three points of strength for this study.

The high prevalence reported should be considered in the context of a tertiary centre where patients are more likely to have secondary hypertension. However, among patients with no other indication than the age of diagnosis of HTN to screen for secondary hypertension (i.e. no resistant hypertension, no grade 3 hypertension, no hypokalemia), the prevalence was still high (23.4%). Unfortunately the referral mode was not collected in our medical report and therefore could not be analyzed.

Some causes were not ruled out but it would only have increased the prevalence of 2HTN. For instance, obstructive sleep apnea syndrome (OSAS) was not systematically ruled out. Besides, we know that incidence of OSAS raises with age [46], and is less common in young patients. Moreover, treatment of OSAS showed limited or no effect on BP [47].

Hyperparathyroidism was not evaluated in this analysis because its implication as a cause of hypertension is debated. Regarding thyroid dysfunction a clinical exam was performed but it was not considered for the analysis.

A control group of patients older than 40 years old would have further strengthened the study, but such control group was difficult to set up mainly because patients older than 40 years without resistant or grade 3 HTN and with normokalemia are seldom screened for 2HTN in accordance with current guidelines.

This study was performed in tertiary centres. In order to limit the risk of potential bias, inclusions were strictly limited to patients first screened in the two centres included and aged between 18 and 40. Patients diagnosed with HTN before 18 years old were not included in the study. Correspondingly, patients diagnosed with HTN before 40 but referred after 40 were not included either because these patients were usually referred with a suspicion of 2HTN or with a resistant or complicated HTN and therefore presented higher likelihood of 2HTN : we considered that the inclusion of these patients in the study could have overestimated the prevalence of secondary hypertension.

Perspectives

According to our study, secondary hypertension is very frequent and significantly more common in hypertensive patients aged 18 to 40 years old than the prevalence reported in the general hypertensive population.

We failed to identify a population with very low risk of 2HTN, and studies analyzing precisely the determinants of each cause of 2HTN are needed in order to narrow the population of patients at risk for 2HTN.

This suggests the interest of performing a systematic screening in patients in this age range, regardless of BP levels, due to the positive impact 2HTN diagnosis can have in young patients in limiting BP exposure and preventing target organ damages, especially in women. We propose that future guidelines recommend screening of 2HTN in all adults under 40 years old.

ACKNOWLEDGEMENTS

We would like to thank Louise Marie Mesples, who participated to data collection and Lauryne Debois, who helped with final data analysis.

SOURCES OF FUNDING

This study was not financially supported.

DISCLOSURES

None

REFERENCES

1. Viera AJ, Neutze DM. Diagnosis of secondary hypertension: an age-based approach. *American Family Physician*. 2010;82(12):1471–1478.

2. Noilhan C, Barigou M, Bieler L, Amar J, Chamontin B, Bouhanick B. Causes of secondary hypertension in the young population: A monocentric study. *Annales De Cardiologie Et D'angeiologie*. 2016;65(3):159–164.

3. Anderson GH, Blakeman N, Streeten DH. The effect of age on prevalence of secondary forms of hypertension in 4429 consecutively referred patients. *Journal of Hypertension*. 1994;12(5):609–615.

4. Gupta-Malhotra M, Banker A, Shete S, Hashmi SS, Tyson JE, Barratt MS, Hecht JT, Milewicz DM, Boerwinkle E. Essential hypertension vs. secondary hypertension among children. *American Journal of Hypertension*. 2015;28(1):73–80.

5. Omura M, Saito J, Yamaguchi K, Kakuta Y, Nishikawa T. Prospective study on the prevalence of secondary hypertension among hypertensive patients visiting a general outpatient clinic in Japan. *Hypertension Research: Official Journal of the Japanese Society of Hypertension*. 2004;27(3):193–202.

6. Forouzanfar MH, Liu P, Roth GA, Ng M, Biryukov S, Marczak L, Alexander L, Estep K, Hassen Abate K, Akinyemiju TF, Ali R, Alvis-Guzman N, Azzopardi P, Banerjee A, Bärnighausen T, et al. Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990-2015. *JAMA*. 2017;317(2):165–182.

7. Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: analysis of worldwide data. *Lancet (London, England)*. 2005;365(9455):217–223.

8. Zhou B, Bentham J, Cesare MD, Bixby H, Danaei G, Cowan MJ, Paciorek CJ, Singh G, Hajifathalian K, Bennett JE, Taddei C, Bilano V, Carrillo-Larco RM, Djalalinia S, Khatibzadeh S, et al. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19·1 million participants. *The Lancet*. 2017;389(10064):37–55.

9. Rahimi K, Duncan M, Pitcher A, Emdin CA, Goldacre MJ. Mortality from heart failure, acute myocardial infarction and other ischaemic heart disease in England and Oxford: a trend study of multiple-cause-coded death certification. *Journal of Epidemiology and Community Health*. 2015;69(10):1000–1005.

10. Hinton TC, Adams ZH, Baker RP, Hope KA, Paton JFR, Hart EC, Nightingale AK. Investigation and Treatment of High Blood Pressure in Young People. *Hypertension*. 2020;75(1):16–22.

11. Battistoni A, Canichella F, Pignatelli G, Ferrucci A, Tocci G, Volpe M. Hypertension in Young People: Epidemiology, Diagnostic Assessment and Therapeutic Approach. *High Blood Pressure & Cardiovascular Prevention: The Official Journal of the Italian Society of Hypertension*. 2015;22(4):381–388.

12. Mills KT, Bundy JD, Kelly TN, Reed JE, Kearney PM, Reynolds K, Chen J, He J. Global Burden of Hypertension: Analysis of Population-based Studies from 89 Countries. *Journal of Hypertension*. 2015;33:e2.

13. Savard S, Amar L, Plouin P-F, Steichen O. Cardiovascular complications associated with primary aldosteronism: a controlled cross-sectional study. *Hypertension (Dallas, Tex.: 1979)*. 2013;62(2):331–336.

14. Rimoldi SF, Scherrer U, Messerli FH. Secondary arterial hypertension: when, who, and how to screen? *European Heart Journal*. 2014;35(19):1245–1254.

15. Spagnolo A, Giussani M, Ambruzzi AM, Bianchetti M, Maringhini S, Matteucci MC, Menghetti E, Salice P, Simionato L, Strambi M, Virdis R, Genovesi S. Focus on prevention, diagnosis and treatment of hypertension in children and adolescents. *Italian Journal of Pediatrics*. 2013;39:20.

16. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, Clement DL, Coca A, de Simone G, Dominiczak A, Kahan T, Mahfoud F, Redon J, Ruilope L, Zanchetti A, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertensionThe Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). *European Heart Journal*. 2018;39(33):3021–3104.

17. Mancia Chairperson G, Kreutz Co-Chair R, Brunström M, Burnier M, Grassi G, Januszewicz A, Muiesan ML, Tsioufis K, Agabiti-Rosei E, Algharably EAE, Azizi M, Benetos A, Borghi C, Hitij JB, Cifkova R, et al. 2023 ESH Guidelines for the management of arterial hypertension The Task Force for the management of arterial hypertension of the European Society of Hypertension Endorsed by the International Society of Hypertension (ISH) and the European Renal Association (ERA). *Journal of Hypertension*. 2023.

18. Unger T, Borghi C, Charchar F, Khan NA, Poulter NR, Prabhakaran D, Ramirez A, Schlaich M, Stergiou GS, Tomaszewski M, Wainford RD, Williams B, Schutte AE. 2020 International Society of Hypertension Global Hypertension Practice Guidelines. :24.

19. Degoulet P, Chatellier G, Devriès C, Lavril M, Ménard J. Computer-assisted techniques for evaluation and treatment of hypertensive patients. *American Journal of Hypertension*. 1990;3(2):156–163.

20. Degoulet P, Marin L, Lavril M, Le Bozec C, Delbecke E, Meaux J-J, Rose L. The HEGP component-based clinical information system. *International Journal of Medical Informatics*. 2003;69(2–3):115–126.

21. Jeunemaitre X, Degoulet P, Morice V, Chatellier G, Devries C, Plouin PF, Boisvieux JF, Ménard J. [Testing an expert system for hypertension]. *Archives Des Maladies Du Coeur Et Des Vaisseaux*. 1986;79(6):808–812.

22. Benchimol EI, Smeeth L, Guttmann A, Harron K, Hemkens LG, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM, RECORD Working Committee. [The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement]. *Zeitschrift Fur Evidenz, Fortbildung Und Qualitat Im Gesundheitswesen*. 2016;115–116:33–48.

23. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M, STROBE initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Annals of Internal Medicine*. 2007;147(8):W163-194.

24. La Batide-Alanore A, Chatellier G, Bobrie G, Fofol I, Plouin PF. Comparison of nurse- and physician-determined clinic blood pressure levels in patients referred to a hypertension clinic: implications for subsequent management. *Journal of Hypertension*. 2000;18(4):391–398.

25. Parati G, Stergiou G, O'Brien E, Asmar R, Beilin L, Bilo G, Clement D, de la Sierra A, de Leeuw P, Dolan E, Fagard R, Graves J, Head GA, Imai Y, Kario K, et al. European Society of Hypertension practice guidelines for ambulatory blood pressure monitoring. *Journal of Hypertension*. 2014;32(7):1359–1366.

26. Veldhuizen GP, Alnazer RM, de Leeuw PW, Kroon AA. The Effects of Verapamil, Hydralazine, and Doxazosin on Renin, Aldosterone, and the Ratio Thereof. *Cardiovascular Drugs and Therapy*. 2023;37(2):283–289.

27. Funder JW, Carey RM, Mantero F, Murad MH, Reincke M, Shibata H, Stowasser M, Young WF. The Management of Primary Aldosteronism: Case Detection, Diagnosis, and Treatment: An Endocrine Society Clinical Practice Guideline. *The Journal of Clinical Endocrinology and Metabolism*. 2016;101(5):1889–1916.

28. Mulatero P, Monticone S, Deinum J, Amar L, Prejbisz A, Zennaro M-C, Beuschlein F, Rossi GP, Nishikawa T, Morganti A, Seccia TM, Lin Y-H, Fallo F, Widimsky J. Genetics, prevalence, screening and confirmation of primary aldosteronism: a position statement and consensus of the Working Group on Endocrine Hypertension of The European Society of Hypertension ∗. *Journal of Hypertension*. 2020;38(10):1919–1928.

29. Cornu E, Steichen O, Nogueira-Silva L, Küpers E, Pagny J-Y, Grataloup C, Baron S, Zinzindohoue F, Plouin P-F, Amar L. Suppression of Aldosterone Secretion After Recumbent Saline Infusion Does Not Exclude Lateralized Primary Aldosteronism. *Hypertension (Dallas, Tex.: 1979)*. 2016;68(4):989–994.

30. Baron S, Amar L, Faucon A-L, Blanchard A, Baffalie L, Faucard C, Travers S, Pagny J-Y, Azizi M, Houillier P. Criteria for diagnosing primary aldosteronism on the basis of liquid chromatography-tandem mass spectrometry determinations of plasma aldosterone concentration. *Journal of Hypertension*. 2018;36(7):1592–1601.

31. Lenders JWM, Kerstens MN, Amar L, Prejbisz A, Robledo M, Taieb D, Pacak K, Crona J, Zelinka T, Mannelli M, Deutschbein T, Timmers HJLM, Castinetti F, Dralle H, Widimský J, et al. Genetics, diagnosis, management and future directions of research of phaeochromocytoma and paraganglioma: a position statement and consensus of the Working Group on Endocrine Hypertension of the European Society of Hypertension. *Journal of Hypertension*. 2020;38(8):1443–1456.

32. Lenders JWM, Duh Q-Y, Eisenhofer G, Gimenez-Roqueplo A-P, Grebe SKG, Murad MH, Naruse M, Pacak K, Young WF, Endocrine Society. Pheochromocytoma and paraganglioma: an endocrine society clinical practice guideline. *The Journal of Clinical Endocrinology and Metabolism*. 2014;99(6):1915–1942.

33. Gornik HL, Persu A, Adlam D, Aparicio LS, Azizi M, Boulanger M, Bruno RM, De Leeuw P, Fendrikova-Mahlay N, Froehlich J, Ganesh SK, Gray BH, Jamison C, Januszewicz A, Jeunemaitre X,

et al. First international consensus on the diagnosis and management of fibromuscular dysplasia: *Journal of Hypertension*. 2019;37(2):229–252.

34. Levin A, Stevens PE, Bilous RW, Coresh J, Francisco ALMD, Jong PED, Griffith KE, Hemmelgarn BR, Iseki K, Lamb EJ, Levey AS, Riella MC, Shlipak MG, Wang H, White CT, et al. Kidney disease: Improving global outcomes (KDIGO) CKD work group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International Supplements*. 2013;3(1):1–150.

35. La Batide-Alanore A, Chatellier G, Plouin P-F. Diabetes as a marker of pheochromocytoma in hypertensive patients. *Journal of Hypertension*. 2003;21(9):1703–1707.

36. Hanslik G, Wallaschofski H, Dietz A, Riester A, Reincke M, Allolio B, Lang K, Quack I, Rump LC, Willenberg HS, Beuschlein F, Quinkler M, Hannemann A, for the participants of the German Conn's Registry. Increased prevalence of diabetes mellitus and the metabolic syndrome in patients with primary aldosteronism of the German Conn's Registry. *European Journal of Endocrinology*. 2015;173(5):665–675.

37. Pappaccogli M, Prejbisz A, Ciurică S, Bruno RM, Aniszczuk-Hybiak A, Bracalente I, De Backer T, Debiève F, Delmotte P, Di Monaco S, Jarraya F, Gordin D, Kosiński P, Kroon AA, Maas AHEM, et al. Pregnancy-Related Complications in Patients With Fibromuscular Dysplasia: A Report From the European/International Fibromuscular Dysplasia Registry. *Hypertension (Dallas, Tex.: 1979)*. 2020;76(2):545–553.

38. Amar L, Azizi M. HTA secondaires chez la femme. *La Presse Médicale*. 2019;48(11, Part 1):1265–1268.

39. Quartermaine G, Lambert K, Rees K, Seed PT, Dhanjal MK, Knight M, McCance DR, Williamson C. Hormone-secreting adrenal tumours cause severe hypertension and high rates of poor pregnancy outcome; a UK Obstetric Surveillance System study with case control comparisons. *BJOG: an international journal of obstetrics and gynaecology*. 2018;125(6):719–727.

40. Trifanescu R, Carsote M, Caragheorgheopol A, Hortopan D, Dumitrascu A, Dobrescu M, Poiana C. Screening for secondary endocrine hypertension in young patients. *Maedica*. 2013;8(2):108–115.

41. Yoon EY, Cohn L, Freed G, Rocchini A, Kershaw D, Ascione F, Clark S. Use of antihypertensive medications and diagnostic tests among privately-insured adolescents and young adults with primary vs. secondary hypertension. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*. 2014;55(1):73–78.

42. Kim K, Kim J-Y, Choi E-Y, Kwon H-M, Rim S-J. Prevalence and risk factors for secondary hypertension among young Korean men. *Reviews in Cardiovascular Medicine*. 2020;21(4):627–634.

43. Di Monaco S, Lengelé J-P, Heenaye S, Danse E, Hammer F, Lopez-Sublet M, Toubiana L, Rabbia F, Persu A. Prevalence and characteristics of renal artery fibromuscular dysplasia in hypertensive women below 50 years old. *European Journal of Clinical Investigation*. 2019;49(10):e13166.

44. Florczak E, Prejbisz A, Szwench-Pietrasz E, Sliwiński P, Bieleń P, Klisiewicz A, Michałowska I, Warchoł E, Januszewicz M, Kała M, Witkowski A, Więcek A, Narkiewicz K, Somers VK, Januszewicz A. Clinical characteristics of patients with resistant hypertension: the RESIST-POL study. *Journal of Human Hypertension*. 2013;27(11):678–685.

45. Monticone S, Burrello J, Tizzani D, Bertello C, Viola A, Buffolo F, Gabetti L, Mengozzi G, Williams TA, Rabbia F, Veglio F, Mulatero P. Prevalence and Clinical Manifestations of Primary Aldosteronism Encountered in Primary Care Practice. *Journal of the American College of Cardiology*. 2017;69(14):1811–1820.

46. Fietze I, Laharnar N, Obst A, Ewert R, Felix SB, Garcia C, Gläser S, Glos M, Schmidt CO, Stubbe B, Völzke H, Zimmermann S, Penzel T. Prevalence and association analysis of obstructive sleep apnea with gender and age differences - Results of SHIP-Trend. *Journal of Sleep Research*. 2019;28(5):e12770.

47. Somers VK, White DP, Amin R, Abraham WT, Costa F, Culebras A, Daniels S, Floras JS, Hunt CE, Olson LJ, Pickering TG, Russell R, Woo M, Young T. Sleep Apnea and Cardiovascular Disease: An American Heart Association/American College of Cardiology Foundation Scientific Statement From the American Heart Association Council for High Blood Pressure Research Professional Education Committee, Council on Clinical Cardiology, Stroke Council, and Council on Cardiovascular Nursing In Collaboration With the National Heart, Lung, and Blood Institute National Center on Sleep Disorders Research (National Institutes of Health). *Journal of the American College of Cardiology*. 2008;52(8):686–717.

PATHOPHYSIOLOGICAL NOVELTY AND RELEVANCE

What is new?

The prevalence of secondary hypertension in young adults is not well defined. We report in this study the prevalence and the underlying causes of 2HTN in young hypertensive subjects in a large bicentric cohort using computerized electronic health records.

What is relevant?

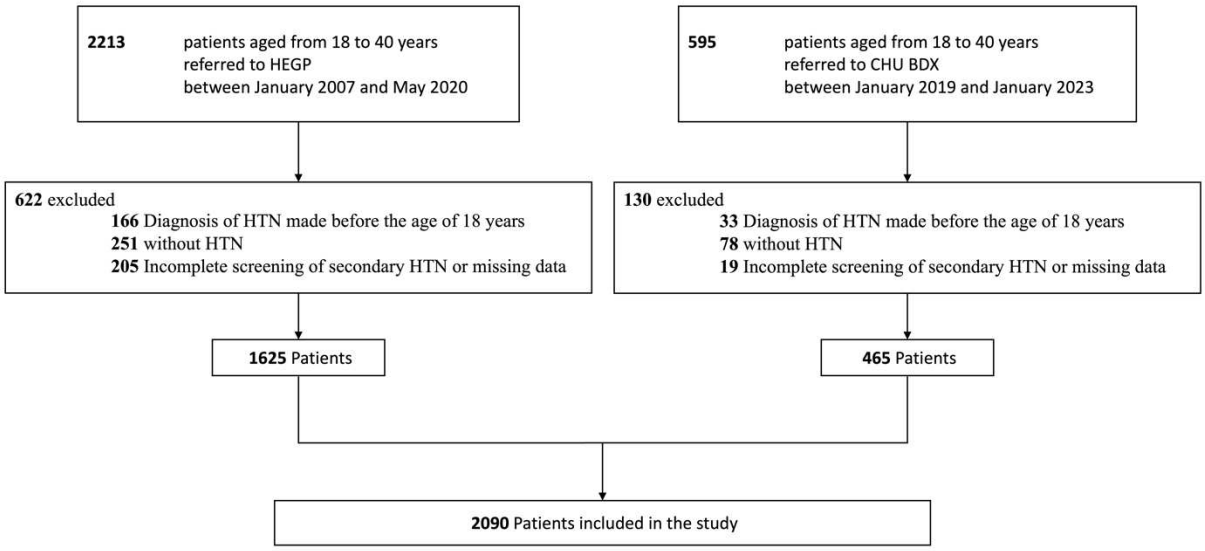
The prevalence of secondary hypertension is higher among young hypertensive patients aged less than 40 years than estimated in the whole population (29.6% vs around 10%).

The prevalence of secondary hypertension in patients aged from 30 to 40 years is most likely at least as high as in patients aged from 18 to 30 years. Secondary hypertension prevalence was even more frequent in women treated with a higher number of antihypertensive drugs, with hypokalemia, diabetes, and normal BMI.

Clinical/Pathophysiological implications?

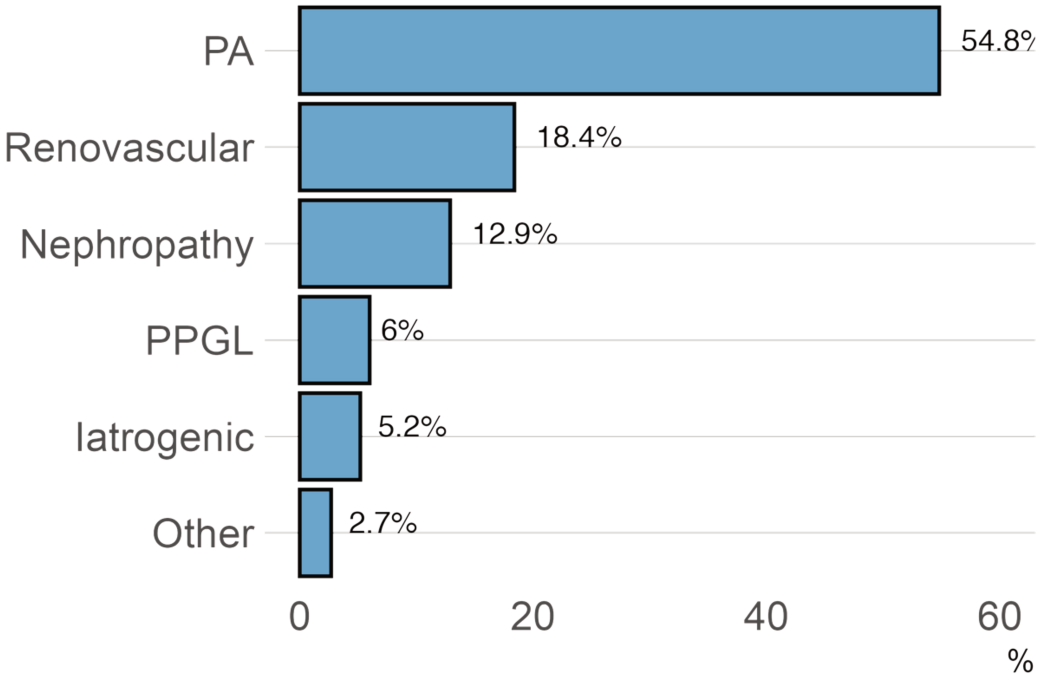
This suggests the interest of performing a systematic screening in patients in this age range, regardless of BP levels, due to the positive impact 2HTN diagnosis can have in young patients in limiting BP exposure and preventing target organ damages, especially in women. We propose that future guidelines recommend screening of 2HTN in all adults under 40 years old.

<u>Figure 1: Flow chart -</u> Patients seen in HEGP and CHU BDX, aged 18 to 40 years and with a diagnosis of hypertension after 18 years old were included in this study.

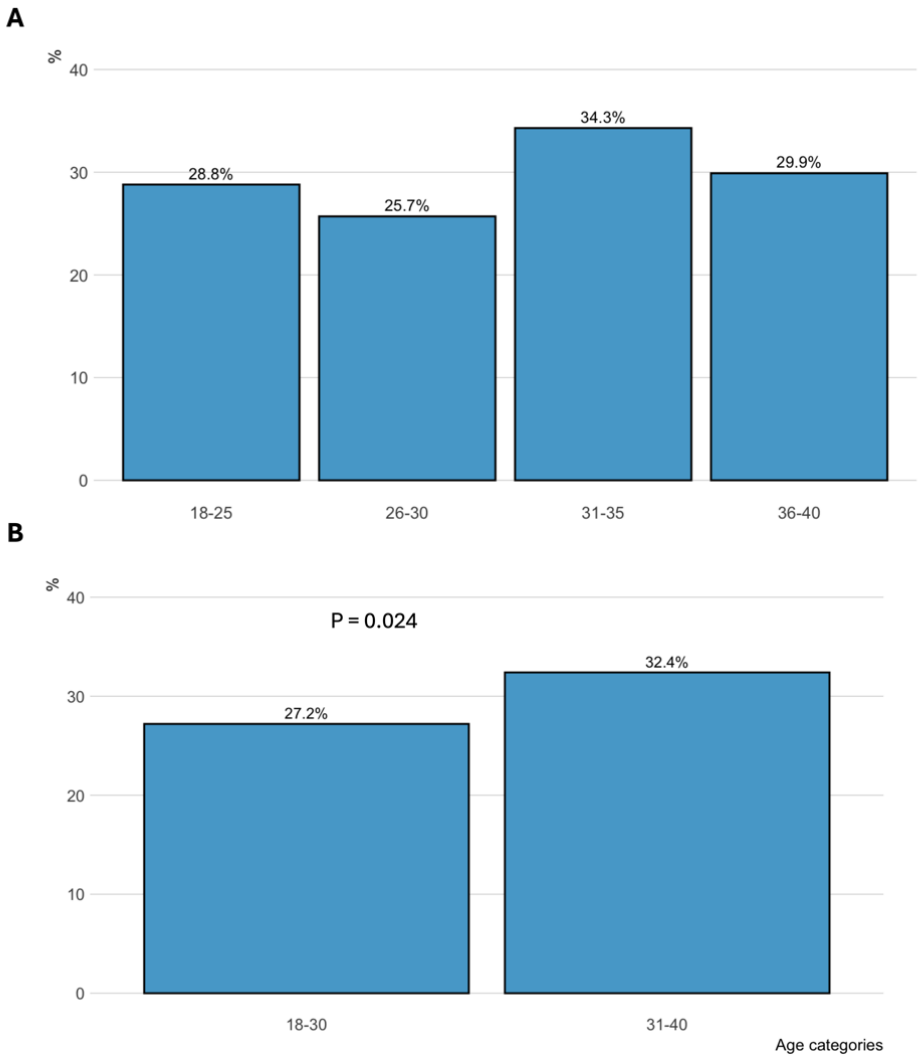


<u>Figure 2: Secondary HTN etiologies</u> - Presentation of the etiologies of hypertension.

*** PA : Primary Aldosteronism; PPGL : Pheochromocytoma or Paraganglioma;**

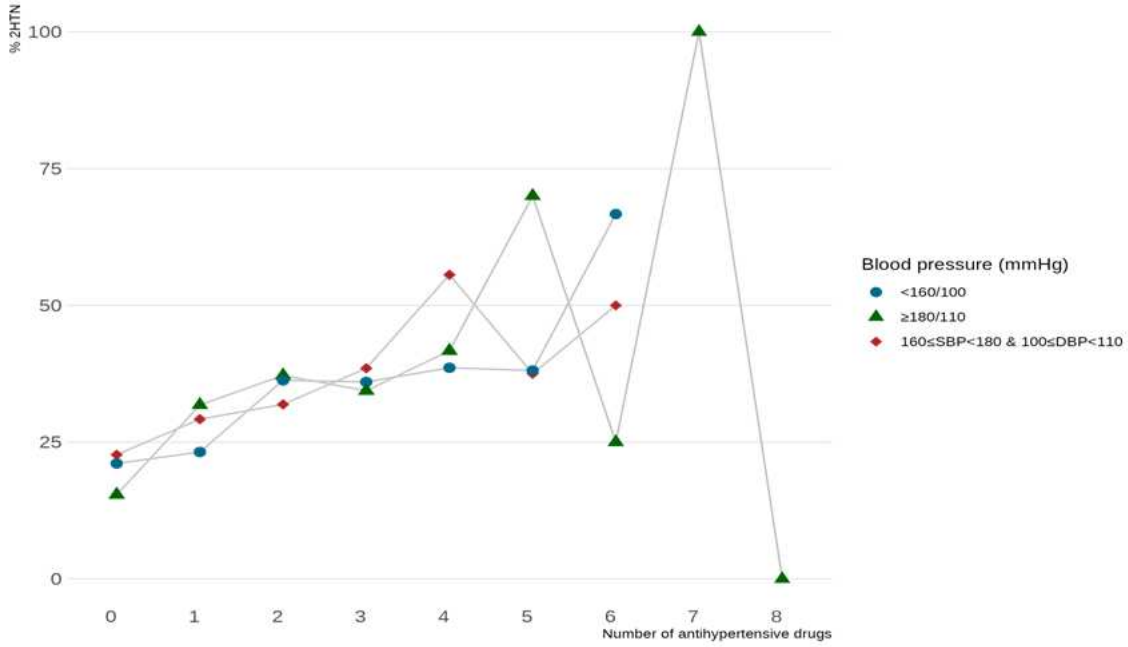


<u>Figure 3: Prevalence of secondary hypertension according to age of onset</u> - Presentation of the proportion of secondary hypertension depending on the age of onset A: four categories; B : two categories



<u>Figure 4 : Percentage of secondary hypertension according to blood pressure and number of treatments</u> - Presentation of the percentage of secondary hypertension according to blood pressure and number of treatments

*** HTN : Hypertension; 2HTN: secondary hypertension**



N tot	352	820	488	273	87	41	12	1	1
N 2HTN	74	204	175	99	37	18	6	1	0

Table 1: Clinical characteristics of the patients

	Total	Number	PHTN	2HTN	p
	2090		1471	619	
Male sex, No. (%)	987 (47.2)	2090	747 (50.8)	240 (38.8)	<0.001
Age at diagnosis of HTN, mean (SD), y	29.8 (5.9)	2090	29.65 (5.94)	30.11 (5.92)	0.105
Familial history of HTN, No. (%)	1206 (58.8)	2052	909 (62.6)	297 (49.6)	<0.001
Current smoker, No. (%)	577 (27.7)	2083	406 (27.7)	171 (27.8)	1.000
Diabetes, No. (%)	98 (4.7)	2082	60 (4.1)	38 (6.2)	0.052
Dyslipidemia, No. (%)	349 (17.1)	2045	247 (17.2)	102 (16.7)	0.854
BMI, mean (SD), kg/m²	27.8 (5.8)	2063	28.1 (5.8)	27.2 (5.6)	<0.001
BMI ≥ 25, No. (%)	1380 (66.9)	2063	1013 (69.2)	367 (61.2)	<0.001
Office systolic BP, mean (SD), mmHg	140.9 (21.1)	2071	140.6 (20.5)	141.7 (22.6)	0.678
Office diastolic BP, mean (SD), mmHg	89.4 (14.8)	2071	89.3 (14.8)	89.8 (15.0)	0.436
Out-of-office systolic BP, mean (SD), mmHg	139.8 (15.2)	1459	139.4 (14.5)	140.8 (16.7)	0.105
Out-of-office diastolic BP, mean (SD), mmHg	91.5 (11.7)	1459	91.2 (11.6)	92.4 (12.0)	0.067
Number of antihypertensive drugs, median (IQR)	1 [1-2]	2075	1 [1-2]	2 [1-3]	<0.001
Number of antihypertensive drugs ≥ 2, No. (%)	903 (43.5)	2075	567 (38.8)	336 (54.7)	<0.001
BP < 160/100 mmHg, No. (%)	1516 (73.2)	2071	1085 (74.3)	431 (70.7)	0.064
BP ≥ 180/110 mmHg, No. (%)	174 (8.4)	2071	114 (7.8)	60 (9.8)	0.152
Resistant HTN, No. (%)	235 (16.2)	1454	145 (14.0)	90 (21.6)	<0.001
Plasma potassium concentration, mean (SD), mmol/L	3.80 (0.41)	2090	3.85 (0.36)	3.66 (0.47)	<0.001
Potassium < 3,5 mmol/L, No. (%)	393 (18.8)	2090	198 (13.4)	195 (31.5)	<0.001
eGFR, mean (SD), ml/min/1.73m2	95.8 (24.7)	2079	104.8 (22.4)	99.1 (25.7)	<0.001
eGFR ≥ 60 ml/min/1.73m2, No. (%)	1941 (93.4)	2079	1413 (96.6)	567 (92.2)	<0.001
ACR ≥ 3 mg/mmol, No. (%)	189 (17.8)	1062	114 (15.1)	75 (24.5)	<0.001

ACR : Urine Albumin to Creatinine Ratio; BP : Blood Pressure; BMI: Body Mass Index; eGFR : estimated Glomerular Filtration Rate; HTN : Hypertension; PHTN: primary hypertension, 2HTN: secondary Hypertension

Table 2: Predictive factors for 2HTN

	Univariate		Multivariate	
	OR	p	OR	p
Male sex	0.61 [0.51;0.74]	<0.001	0.50 [0.35;0.70]	< 0.001
Age at diagnosis of HTN (years)	1.01 [1.00;1.03]	0.105		
Familial history of HTN	0.59 [0.49;0.71]	<0.001	0.57 [0.41;0.81]	0.003
Current smoker	1.02 [0.83;1.25]	0.878		
Diabetes	1.55 [1.01;2.34]	0.045	2.13 [1.02,4.48]	0.045
Dyslipidemia	0.97 [0.75;1.25]	0.808		
BMI ≥ 25 kg/m²	0.70 [0.57;0.85]	<0.001	0.74 [0.52;1.05]	0.097
Office systolic BP	1.00 [1.00;1.01]	0.105		
Office diastolic BP	1.00 [1.00;1.01]	0.067		
Out-of-office systolic BP	1.00 [1.00;1.01]	0.126		
Out-of-office diastolic BP	1.00 [1.00;1.02]	0.209		
Number of antihypertensive drugs ≥ 2	1.90 [1.57;2.31]	<0.001	2.21 [1.58;3.08]	< 0.001
BP < 160/100 mmHg, No. (%)	0.83 [0.68;1.03]	0.093		
BP ≥ 180/110 mmHg, No. (%)	1.29 [0.92;1.78]	0.133		
Resistant HTN	1.70 [1.27;2.27]	< 0.001		
Potassium < 3,5 mmol/L	0.32 [0.26;0.41]	<0.001	3.57 [2.50;5.26]	< 0.001
eGFR < 60 ml/min	2.44 [1.61;3.70]	<0.001		
ACR ≥ 30 mg/mmol	1.07 [1.03;1.11]	<0.001		

Univariate and multivariate analysis to identify predictive factors for secondary hypertension

*** ACR : Urine Albumin to Creatinine Ratio; BP : Blood Pressure; BMI: Body Mass Index; eGFR : estimated Glomerular Filtration Rate; HTN : Hypertension**

Table supp 1 : clinical characteristics of the patients according to the gender

	Total	Number	Men	Women	p
	2090		987	1103	
Age at diagnosis of HTN, mean (SD), y	29.8 (5.9)	2090	30.2 (6.0)	29.4 (5.9)	<0.001
Familial history of HTN, No. (%)	1206 (58.8)	2052	571 (59.0)	635 (58.6)	0.851
Current smoker, No. (%)	577 (27.7)	2083	322 (32.7)	255 (23.2)	<0.001
Diabetes, No. (%)	98 (4.7)	2082	47 (4.7)	51 (4.6)	0.879
Dyslipidemia, No. (%)	349 (17.1)	2045	218 (22.5)	131 (12.2)	<0.001
BMI, mean (SD), kg/m²	27.8 (5.8)	2063	28.1 (5.0)	27.6 (6.4)	0.031
BMI ≥ 25, No. (%)	1380 (66.9)	2063	707 (72.7)	673 (61.7)	<0.001
Office systolic BP, mean (SD), mmHg	140.9 (21.1)	2071	143.1 (20.3)	138.9 (21.6)	<0.001
Office diastolic BP, mean (SD), mmHg	89.4 (14.8)	2071	89.5 (15.1)	89.5 (14.6)	0.969
Out-of-office systolic BP, mean (SD), mmHg	139.8 (15.2)	1459	142.3 (14.5)	137.6 (15.4)	<0.001
Out-of-office diastolic BP, mean (SD), mmHg	91.5 (11.7)	1459	91.0 (12.4)	92.0 (11.1)	0.969
Number of antihypertensive drugs, median (IQR)	1 [1-2]	2075	1 [1-2]	1 [1-2]	
Number of antihypertensive drugs ≥ 2, No. (%)	903 (43.5)	2075	449 (45.8)	454 (41.5)	0.046
~~PA~~ BP < 160/100 mmHg, No. (%)	1516 (73.2)	2071	705 (72.1)	811 (74.2)	0.279
~~PA~~ BP ≥ 180/110 mmHg, No. (%)	174 (8.4)	2071	90 (9.2)	84 (7.7)	0.216
Resistant HTN, No. (%)	235 (16.2)	1454	130 (19.0)	105 (13.6)	0.006
Plasma potassium concentration, mean (SD), mmol/L	3.80 (0.41)	2090	3.83 (0.41)	3.77 (0.40)	<0.001
Potassium < 3,5 mmol/L, No. (%)	393 (18.8)	2090	157 (15.9)	236 (21.3)	0.008
eGFR, mean (SD), ml/min	95.8 (24.7)	2079	94.8 (24.0)	104.4 (25.2)	0.097
eGFR ≥ 60 ml/min, No. (%)	1941 (93.4)	2079	920 (94.1)	1060 (96.5)	0.014
ACR ≥ 3 mg/mmol, No. (%)	189 (17.8)	1062	68 (14.0)	121 (21.0)	0.003
Secondary HTN	619 (29.6%)	2090	240 (24.3)	379 (34.4)	<0.001

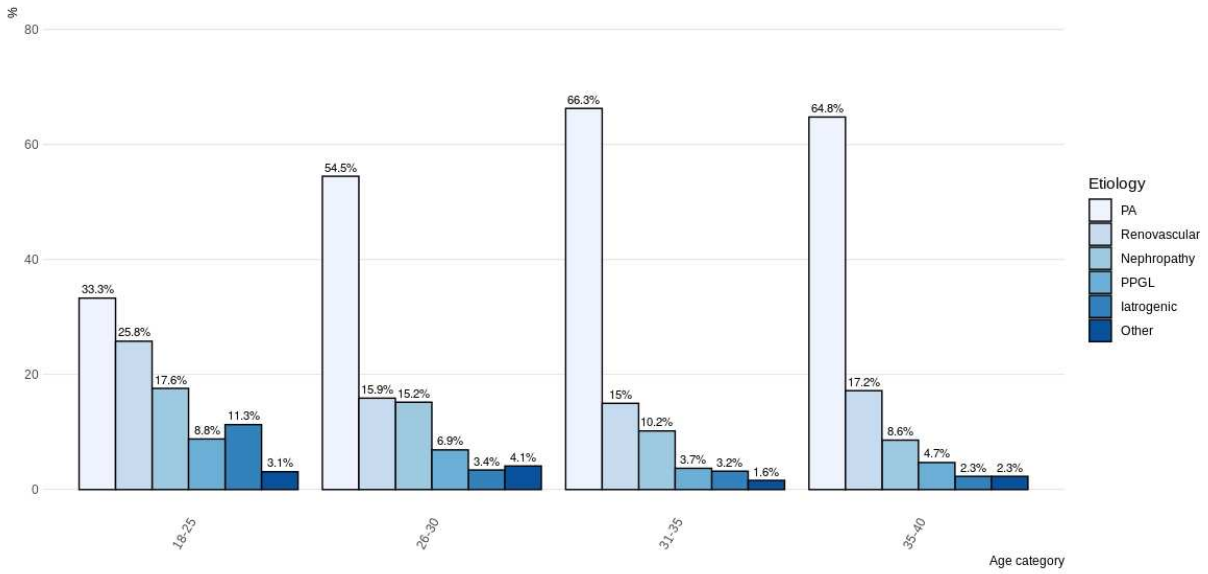
* ACR : Urine Albumin to Creatinin Ratio; BP : Blood Pressure; BMI: Body Mass Index;
eGFR : estimated Glomerular Filtration Rate; HTN : Hypertension;

Table supp 2 : characteristics of the patients according to the center

	Total	HEGP	Number	Bordeaux	Number
	2090	1625		465	
Male sex, No. (%)	987 (47.2)	743 (45.7)	1625	244 (52.5)	465
Age at diagnosis of HTN, mean (SD), y	29.8 (5.9)	29.4 (5.9)	1625	31.1 (5.9)	465
Familial history of HTN, No. (%)	1206 (58.8)	914 (67.2)	1587	173 (37.2)	465
Current smoker, No. (%)	577 (27.7)	424 (26.2)	1618	153 (32.9)	465
Diabetes, No. (%)	98 (4.7)	68 (4.2)	1617	30 (6.5)	465
Dyslipidemia, No. (%)	349 (17.1)	277 (17.5)	1580	72 (15.5)	465
BMI, mean (SD), kg/m²	27.8 (5.8)	27.5 (5.7)	1599	29.1 (5.9)	464
BMI ≥ 25, No. (%)	1380 (66.9)	1036 (64.8)	1599	344 (74.1)	464
Office systolic BP, mean (SD), mmHg	140.9 (21.1)	139.2 (18.8)	1606	146.8 (27.0)	465
Office diastolic BP, mean (SD), mmHg	89.4 (14.8)	89.8 (13.7)	1606	88.2 (18.2)	465
Out-of-office systolic BP, mean (SD), mmHg	139.8 (15.2)	139.3 (15.2)	1277	143.1 (14.7)	182
Out-of-office diastolic BP, mean (SD), mmHg	91.5 (11.7)	91.0 (11.7)	1277	94.9 (11.1)	182
Number of antihypertensive drugs, median (IQR)	1 [1-2]	1 [1-2]	1625	1 [1-2]	450
Number of antihypertensive drugs ≥ 2, No. (%)	903 (43.5)	730 (44.9)	1625	173 (38.4)	450
BP < 160/100 mmHg, No. (%)	1516 (73.2)	1205 (75.0)	1606	311 (66.9)	465
BP ≥ 180/110 mmHg, No. (%)	174 (8.4)	126 (7.8)	1606	48 (10.3)	465
Resistant HTN, No. (%)	235 (16.2)	217 (17.0)	1260	18 (10.2)	177
Plasma potassium concentration, mean (SD), mmol/L	3.80 (0.41)	3.79 (0.42)	1625	3.83 (0.35)	462
Potassium < 3,5 mmol/L, No. (%)	393 (18.8)	319 (19.2)	1625	74 (15.4)	462
eGFR, mean (SD), ml/min	95.8 (24.7)	101.8 (22.2)	1615	107.4 (23.5)	464
eGFR ≥ 60 ml/min, No. (%)	1941 (93.4)	1520 (94.2)	1615	460 (99.1)	464
ACR ≥ 3 mg/mmol, No. (%)	189 (17.8)	114 (15.1)	1062	NA	0
Secondary HTN	619 (29.6%)	492 (30.3)	1625	127 (27.3)	465

* ACR : Urine Albumin to Creatinin Ratio; BP : Blood Pressure; BMI: Body Mass Index;
eGFR : estimated Glomerular Filtration Rate; HTN : Hypertension;
PHTN: primary hypertension; 2HTN: secondary Hypertension

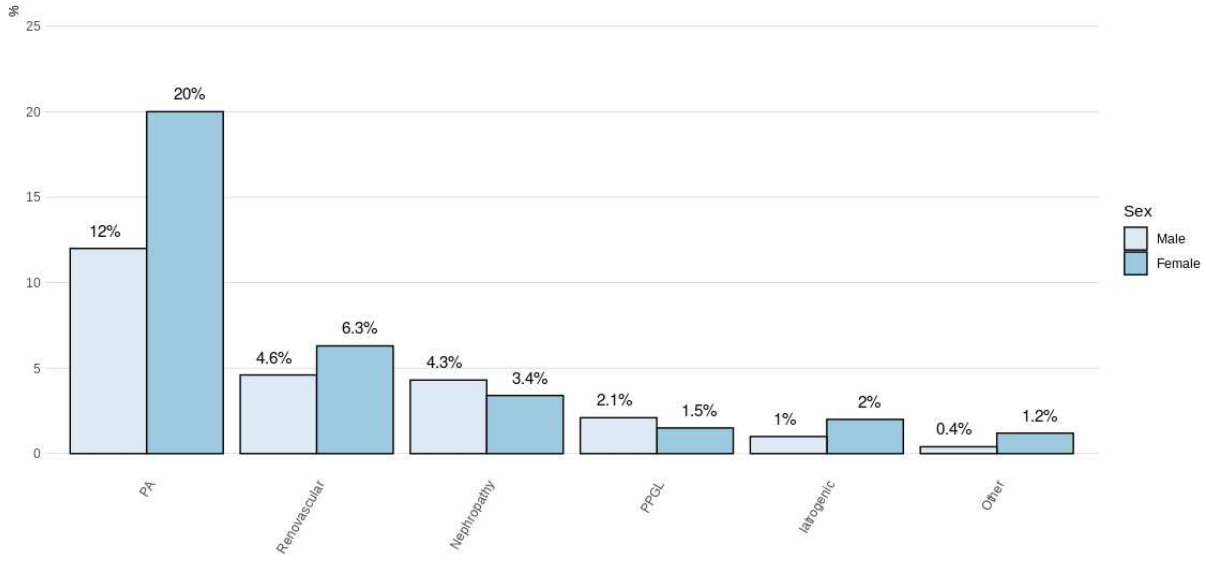
Figure supp 1: Prevalence of causes of secondary hypertension according to age of onset



Repartition of the proportion of different causes of secondary hypertension depending on the age of onset

*** PA: Primary Aldosteronism; PPGL: Pheochromocytoma or Paraganglioma**

Figure supp 2: Prevalence of secondary hypertension according to sex



Repartition of the proportion of secondary hypertension depending on sex

*** PA: Primary Aldosteronism; PPGL: Pheochromocytoma or Paraganglioma**

Figure supp 3: Estimated probabilities of secondary HTN using Bootstrap resampling

			Male				Female			
			No familial history of HTN		Familial history of HTN		No familial history of HTN		Familial history of HTN	
			BMI ≥ 25	BMI < 25	BMI ≥ 25	BMI < 25	BMI ≥ 25	BMI < 25	BMI ≥ 25	BMI < 25
Number of treatments ≥2	Potassium < 3,5	Diabetes	50.3 [40.6, 59.8]	60.6 [50.5, 70.5]	64.9 [55.9, 73.3]	73.8 [64.9, 82]	62.7 [53.3, 71.9]	71.9 [63.3, 80.2]	75.4 [67.3, 82.4]	82.4 [75.5, 88.4]
		No diabetes	37.4 [32.2, 43.1]	47.8 [40.8, 54.8]	52.3 [46.4, 58.5]	62.7 [55.8, 69.4]	50.0 [44, 55.3]	60.5 [54, 66.5]	64.7 [59.2, 70.1]	73.7 [68.4, 79.1]
	Potassium ≥ 3,5	Diabetes	25.8 [19.1, 32.7]	34.7 [25.9, 43.8]	38.8 [30.2, 47.4]	49.2 [38.8, 59.5]	36.6 [28.2, 45.4]	46.9 [36.8, 56.9]	51.4 [42, 60.4]	61.7 [51.6, 71.2]
		No diabetes	16.9 [14.3, 19.6]	23.8 [19.4, 28.4]	27.2 [23.3, 31.3]	36.5 [30.8, 42.2]	25.4 [22, 28.9]	34.3 [29.2, 39.1]	38.5 [33.9, 43]	48.9 [43.1, 54.5]
Number of treatments <2	Potassium < 3,5	Diabetes	35.8 [26.7, 45.6]	46.0 [35.3, 57.6]	50.5 [40.2, 60.8]	60.8 [50, 71.2]	48.2 [37.9, 58.5]	58.6 [47.8, 69.2]	62.9 [53, 72.2]	72.1 [62.9, 80.4]
		No diabetes	24.7 [20.2, 29.5]	33.5 [27.2, 39.9]	37.6 [31.5, 43.9]	48.0 [40.8, 55.2]	35.4 [30.1, 41]	45.7 [39.3, 52]	50.2 [44.1, 56.3]	60.7 [54.5, 67.2]
	Potassium ≥ 3,5	Diabetes	16.1 [11.3, 21.5]	22.7 [16.1, 30.1]	25.9 [19.3, 33.4]	34.9 [26.2, 44.3]	24.2 [17.6, 31.5]	32.8 [24.4, 41.6]	36.8 [28.3, 45.7]	47.1 [37.5, 56.9]
		No diabetes	10.0 [8.26, 11.9]	14.6 [11.9, 17.6]	17.0 [14.4, 19.9]	23.9 [19.9, 28.2]	15.7 [13.3, 18.4]	22.2 [19, 25.6]	25.5 [22.1, 29.1]	34.5 [30.2, 39.1]

Probability of 2HTN

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Complément de méthodes et résultats

Méthodes : estimation du risque d'HTA secondaire par méthode régression logistique et bootstrap

Nous avons réalisé une régression logistique multiple pour estimer le risque d'HTA secondaire en fonction de différents prédicteurs.

Les prédicteurs choisis l'ont été en utilisant une méthode stepwise, basée sur le critère d'Akaike (AIC – Akaike Information Criterion)

Régression logistique stepwise

Le principe est d'automatiser la sélection des variables les plus pertinentes dans un modèle. Plutôt que de choisir manuellement quelles variables inclure, cette méthode ajoute ou retire des variables en fonction de leur impact sur la qualité du modèle, évaluée par le critère d'Akaike.

Le critère d'Akaike mesure la qualité d'un modèle statistique tout en pénalisant la complexité (nombre de variables) pour éviter d'ajouter des variables inutiles. Un AIC plus bas indique un meilleur équilibre entre précision et simplicité.

Formellement, l'AIC est défini par :

$$AIC = 2k - \ln L$$

Où :

- k est le nombre de paramètres dans le modèle,
- L est la vraisemblance maximale du modèle.

La fonction de vraisemblance d'une régression logistique étant :

$$L(\beta) = \prod \pi_i^{y_i} \times (1 - \pi_i)^{(1-y_i)}$$

Où:

- y_i est la valeur observée de la variable dépendante
- π_i est la probabilité prédite pour $y_i = 1$

Il existe deux approches principales pour la sélection stepwise : avant (forward) et arrière (backward). La méthode stepwise consiste à combiner ces deux méthodes. Les étapes en sont les suivantes :

- Étape initiale : on commence avec l'ensemble des variables d'intérêt.
- Ajout et retrait : à chaque étape, on teste l'ajout et le retrait de variables et on effectue l'action qui améliore le plus l'AIC
- Arrêt : quand aucune action ne peut améliorer l'AIC.

Estimation de la probabilité par inférence statistique et bootstrap

La probabilité pour chaque combinaison de variable a été estimée à partir de modèle de régression logistique.

En effet, comme nous l'avons vu dans la partie 1 :

$$p = \frac{1}{1 + e^{-(w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3 + b)}}$$

Avec p la probabilité de Y = 1 pour la variable X (x_1, x_2, x_3).

Cette probabilité a été estimée pour chaque combinaison de variables, sur 1000 échantillons bootstrap, d'où on a tiré une valeur moyenne et un intervalle de confiance pour la probabilité d'HTA secondaire.

Complément de discussion : apport de l'algorithme dans la stratégie de dépistage des patients jeunes

Le premier résultat de cette analyse est la confirmation d'une prévalence élevée d'HTA secondaire chez les patients jeunes. Ce résultat est d'importance pour plusieurs raisons. Tout d'abord, les études épidémiologiques concernant l'HTA secondaire sont difficiles car malgré les recommandations préconisant un bilan exhaustif dans cette catégorie d'âge [17], il existe très probablement un sous-diagnostic important. L'ensemble des patients de moins adressés à l'HEGP pour une HTA ayant débuté avant l'âge de 40 ans bénéficient d'un bilan d'HTA secondaire complet, ce qui nous a permis de décrire une cohorte parfaitement phénotypée. Cela confirme la prévalence élevée d'HTA secondaire dans cette tranche d'âge et la nécessité de réaliser un bilan exhaustif chez tous ces patients. Depuis la publication de cet article, de nouvelles recommandations de la société européenne de cardiologie[18] ont proposé que certaines causes comme l'HAP soient recherchées chez tout patient hypertendu, quel que soit l'âge.

Le second résultat d'importance est la caractérisation du risque d'HTA secondaire en fonction de certains paramètres, premier pas vers un modèle prédictif plus individualisé, développé dans la partie suivante. Les paramètres retrouvés sont pour la plupart déjà connus. Ainsi l'hypokaliémie et le nombre de traitements antihypertenseurs (et par extension l'HTA résistante) sont connus pour être associés à l'HTA secondaire[10,111]. De même, des antécédents familiaux d'HTA, et le sexe masculin, sont connus pour être associés à l'HTA primaire[43]. En revanche, l'obésité et le surpoids, tout comme le diabète et le syndrome métabolique, ont été associés à la fois à l'HTA primaire[112,113], mais aussi à certaines causes d'HTA secondaires comme l'HAP[114]. De même, le diabète est associé aux PPGL[115]. Notre analyse prédictive propose une probabilité d'HTA secondaire en considérant la combinaison de ces différents prédicteurs, sachant leur équivocité.

Le *bootstrap* a déjà été utilisé dans des modèle prédictifs, que ce soit pour le calcul de coefficients de régression logistiques, ou dans le cadre d'un score[116–118]. Dans notre cas, il permet une amélioration de la robustesse du calcul de probabilités, et ainsi une interprétation plus fine de l'impact de la présence ou non de chaque prédicteur.

Compte tenu de l'importance du diagnostic de l'HTA secondaire, nous avons conclu au fait qu'il était difficile de considérer que certains patients pouvaient être exclus du dépistage de l'HTA secondaire. En effet, la probabilité est au plus bas autour de 10%, avec un intervalle de confiance assez rapproché. Il appartient à chacun, et notamment au patient, qui doit être intégré aux décisions concernant son parcours

de soin, de mesurer l'impact de ce risque, mais ce chiffre nous a semblé trop élevé pour envisager de ne pas recommander le dépistage systématique dans cette population de patient. Dans la discussion avec le patient, l'estimation de cette probabilité peut cependant apporter des éléments concrets pour nourrir la discussion.

Cette analyse présente évidemment des limites, avec en premier lieu la sélection préalable des patients, puisque ces patients ont été adressés dans un centre expert d'HTA, limite qui est bien décrite comme une des principales difficultés dans les études sur la prévalence des maladies51. La population étudiée n'est donc pas représentative de la population générale. Cependant, elle présente l'avantage d'inclure des sujets affectés et non affectés, et d'être la plus importante décrite à ce jour avec un tel niveau de phénotypage. Enfin, une grande partie de la population a été adressée uniquement sur le critère de l'âge.

Chapitre 5 : modèle prédictif d’hypertension artérielle secondaire

Introduction

La recherche d'hypertension artérielle secondaire est une question récurrente dans la prise en charge de l'hypertension artérielle, comme nous l'avons vu dans les précédentes parties.

Parmi les causes d'hypertension artérielle secondaires, la plus fréquentes et l'hyperaldostéronisme primaire, probablement responsable d'environ 10 % des cas, bien que cela soit possiblement sous-estimé. Nous avons d'ailleurs constaté que dans la population des patients de moins de 40 ans, la prévalence de cette maladie en centre expert est plutôt autour de 30 %[44]. Il est difficile d'estimer correctement cette prévalence en population générale, car cette recherche est réalisée la plupart du temps dans des centres experts, avec par conséquent un biais de sélection possible lors de l'analyse de ses populations[51].

Le dépistage de l'hyperaldostéronisme primaire est particulièrement important car il a été démontré que l'hypersécrétion d'aldostérone est responsable de dommages sur différents organes, tels que le rein et le cœur, indépendamment du niveau de pression artérielle[12,13,110]. Par ailleurs, son diagnostic est délicat, puisqu'il nécessite la modification du traitement antihypertenseur pour des molécules « neutres », c'est-à-dire qui n'interagissent pas avec le système rénine en en aldostérone et ne faussent donc pas les dosages[119].

Certains facteurs associés à l'hyperaldostéronisme primaire sont bien connus, et sont évoqués dans les recommandations pour réaliser ce dépistage[17]. Il s'agit de l'âge inférieur à 40 ans, de l'hypokaliémie, des dommages d'organes disproportionnés par rapport au niveau de pression artérielle, de la présence d'une fibrillation atriale, ou de la découverte d'un incidentalome surrénalien. Une histoire familiale d'hyperaldostéronisme primaire doit aussi faire évoquer le diagnostic[120]. Cependant, dans une proportion non négligeable de cas, il n'y a pas de symptômes particuliers pouvant faire évoquer ce diagnostic[25].

Les dernières recommandations de l'ESC[18] proposent le dépistage de l'HAP chez tous les adultes hypertendus. Cependant, cela semble difficile à mettre en place pour les raisons évoquées, ci-dessus. Des stratégies de stratification du risque d'hyperaldostéronisme primaire serait donc intéressantes afin de d'affiner ces recommandations et d'identifier des populations plus à risque. Dans cette optique, un algorithme de dépistage aurait son intérêt en maximisant sa sensibilité et sa valeur prédictive négative, et en limitant donc au maximum le risque de faux négatifs. Cela pourrait donc permettre de dépister le maximum de patients

hypertendus, tout en éliminant le diagnostic d'HAP primaire chez quelques patients, et donc quelques examens inutiles.

Nous nous sommes également intéressés à deux autres diagnostics d'HTA secondaire : l'HTA rénovasculaire, et les phéochromocytomes et paragangliomes sécrétants (PPGL). En effet, l'HTA rénovasculaire est une cause très fréquente d'HTA secondaire, pouvant avoir des conséquences significatives, et disposant parfois d'un traitement efficace[121–123]. Quant aux PPGL, il s'agit d'une cause rare, mais potentiellement grave d'HTA secondaire, dont la mise en place rapide du traitement est essentielle[56,124]. Nous n'avons pas étudié les autres causes du fait de leur rareté, de leur prévalence faible dans notre cohorte, et l'impact clinique que nous avons estimé moindre.

Les pistes évoquées pour développer un algorithme de dépistage intègrent l'intelligence artificielle, dont le développement extrêmement rapide actuellement permet des applications dans des domaines très variés[64]. Nous avons donc envisagé d'utiliser l'apprentissage automatique et l'apprentissage profond, avec des algorithmes entraînés sur notre base de données de l'HEGP et validé de manière externe sur les données du CHU de Bordeaux. Ce type d'algorithme ou de modèle est utilisable avec des données structurées, telles que les nôtres, et il existe plusieurs exemples à l'heure actuelle dans la littérature de développement de ce type de modèle dans le domaine de l'hypertension artérielle et de l'hyperaldostéronisme primaire[28,31,32,88,125]. Cependant, il existe de nombreuses limitations à l'utilisation de ce type de modèle. En effet bien que leur performance puisse dépasser celle des modèles traditionnels, ils sont également soumis à des biais importants et ont par ailleurs, le désavantage de ne pas être explicables, c'est-à-dire que la réponse à une question donnée ne peut pas être examinée à la lumière des facteurs y ayant amené[72,73].

Objectifs

L'objectif de ce travail était donc d'élaborer un algorithme facilitant le dépistage de l'HTA secondaire, ayant une sensibilité et une valeur prédictive négative suffisante, ainsi qu'une interprétabilité permettant son utilisation en médecine de ville. La partie principale du travail présenté, ayant abouti à un article ci-après, concerne le dépistage de l'HAP. Nous présentons ensuite les résultats pour les autres pathologies.

Présentation du manuscrit

Manuscrit soumis à la revue *Hypertension* avec demande de révision

Predicting primary aldosteronism: risk stratification-based and machine learning-based models

Jean-Baptiste de Freminville[a,b], Laurence Amar[a,c,d], Jean Feydy[b,e], Antoine Cremer[f,] Romain Boulestreau[g,h], Philippe Gosse[f], Stéphanie Baron[i,] Michel Azizi[a,d], Bastien Rance[b,j]

a Hypertension Unit, Université Paris-Cité, AP-HP, Hôpital Européen Georges Pompidou, Paris, France

b INSERM, UMRS 1138, CRC, Université Sorbonne-Paris Cité, Paris, France

c Adrenal referral center, Université Paris Cité, AP-HP, France

d Inserm Paris Cardiovascular research center, 75015 Paris, France

e.HeKA team INRIA, INSERM, Université Paris-Cité, Paris, France

f Hypertension Excellence Center, University Hospital of Bordeaux, Hopital Saint André, Bordeaux

f Coronary and vascular disease department, CHU de Bordeaux, Bordeaux
g UMR INSERM U1034, Bordeaux
h Physiology department, AP-HP, Hôpital Européen Georges Pompidou, Université Paris-Cité, Paris, France
i Inserm CIC 1418, F 75015, Paris, France
j Medical Informatics, Biostatistics and Public health department, AP-HP, Hôpital Européen Georges Pompidou, Université Paris-Cité, Paris, France

Short Title: Predicting models for primary aldosteronism

Corresponding author

Jean-Baptiste de Freminville

Hypertension Unit, Hôpital Européen Georges Pompidou, 20 rue Leblanc, 75908 Paris cedex 15, France

email address: jean.de-freminville@polytechnique.org

word count : 4179

ABSTRACT

Background: Primary aldosteronism (PA) screening is difficult in an unselected population of hypertensive patients, as it can be challenging. We report new algorithms based either on risk stratification or machine learning models to improve the screening of PA.

Methods: 15,507 patients with confirmed hypertension aged 18 to 65 years with full work-up for PA screening were included in this study. We developed $PA_{strat,}$ a risk-stratification based algorithm by generating patient groups of interest and assess probability of PA in each group, and two different machine learning model, logistic regression and XGBoost. An external cohort of 768 patients was used for validation.

Results: The AUC of the ROC curves in the derivation cohort were respectively 0.80, 0.82 and 0.83 for the PA_{strat} algorithm, logistic regression, and XGboost. Only 6 out of 768 patients (0.7%) had PA despite incorrect prediction by PA_{strat} (negative predictive value 0.96). Conversely, among the 128 patients who had an indication for PA screening according to guidelines but not according to the algorithm, only 3 (0.23%) patients had PA. Logistic regression and XGBoost had respectively 49 (6.7%) and 84 (10.9%) false negatives (negative predictive value respectively 0.87 and 0.88).

Conclusions: We developed a risk-stratification algorithm for PA screening, which can be simply used in clinical practice. Machine learning models had good performance but their lack of explainability and lower negative value make them less easily usable. Other external validations in a general population must be done to confirm.

Keywords: Hypertension, Hypokalemia, Body Mass Index, Pheochromocytoma, Hyperaldosteronism, Diabetes Mellitus, Renovascular Hypertension

ABBREVIATIONS

ABPM : Ambulatory Blood Pressure Monitoring

AIC : Akaike Information Criteria

ACE : angiotensin converting enzyme

ARB : angiotensin II receptor blockers

ARR : Aldosterone-to-Renin Ratio

BP : Blood Pressure

CHU BDX : Bordeaux University hospital

CKD : Chronic Kidney Disease

CKD-EPI : Chronic Kidney Disease Epidemiology Collaboration

CV : Cardiovascular

eGFR : Estimated glomerular filtration rate

HBPM : Home Blood Pressure Monitoring

HTN : Hypertension

2HTN : Secondary causes of HTN

HEGP : Georges Pompidou Hospital

LC-MS/MS : Liquid Chromatography with tandem Mass Spectrometry

NLP : Natural Language Processing

MR : Mineralocorticoid Receptor

OBP : Office Blood Pressure

OOBP : Out of Office Blood Pressure

OSAS : Obstructive Sleep Apnea Syndrome

PAC : Plasma Aldosterone Concentration

PRC : Plasma Renin Concentration

RIA : Radioimmunoassay

INTRODUCTION

Primary aldosteronism (PA) is one of the main causes of secondary HTN with an estimated prevalence between 6 to 18% of hypertensive adults[1–5], but is probably underdiagnosed. However, its detection is important, since 1) it is a frequent cause of resistant HTN [6–8], 2) aldosterone in excess together with high-sodium environment, is associated with an increased risk of target organ damage, independently of blood pressure (BP) levels [9–12], and 3) PA is accessible to specific treatments including mineralocorticoid receptor antagonists or unilateral adrenalectomy in patients with aldosterone producing adenomas (APA) [13–15].

Screening for PA relies on the determination of the plasma aldosterone:renin ratio which faces various challenges. Indeed it requires 2- to 6- week withdrawal of all medications interfering with the renin-angiotensin-aldosterone system, standardization of dietary sodium and potassium intakes, correction of hypokalemia if needed, blood samplings in standardized conditions and standardized hormonal determinations [13,16],. All these constraints make difficult to diagnose PA in unselected population of hypertensive patients, and consequently result in PA being underdiagnosed worldwide [4,5]. There is thus a need to preselect the population of hypertensive patients in whom probability to have PA is higher.

To facilitate the selection of patients suspect of PA, clinical scoring systems based on traditional statistical methods were developed, albeit with a low diagnostic performance [17–19]. More recently, diagnosis tools, based on machine learning models have been applied to PA diagnosis [20]. However, these models have been subject to major issues. First, these models lack explainability. Second, the risk of overfitting can lead to models with low reproducibility, and external validation, which is necessary to assess generalizability of the models, often show that the performances are not reliable enough to do without interpretability.

We report new algorithms for PA screening based either on risk stratification using simple and easily accessible clinical and biological parameters, or machine learning models. Both were developed on a large cohort of 15,507 patients with HTN consulting a tertiary university center in Paris of whom 1817 (11.7%) had PA. These algorithms underwent an external validation in another cohort of patients referred to another tertiary centre in Bordeaux. We cross-compared the performance of those models and discussed their potential impact in real life.

PATIENTS AND METHODS

We followed the TRIPOD+AI statement guidance for reporting clinical prediction models that use regression or machine learning methods [21].

Derivation and validation cohorts

We included in the derivation cohort patients referred to Hopital Europeen Georges Pompidou (HEGP) with confirmed HTN aged 18 to 65 years between 17/05/2005 to 31/05/2020, after exclusion of patients with kidney replacement therapy (KRT), and patients with too many missing data (more than 25%) (Figure 1: flow chart).

For the validation cohort, we included patients referred to the CHU Bordeaux (CHU BDX) for secondary HTN diagnosis workup 01/01/2019 and 31/01/2023.

The diagnostic criteria for PA followed the current French and European recommendations [16,22]. It is based on a plasma Aldosterone-to-Renin Ratio (ARR) above the threshold on at least 2 separate occasions with an increased basal plasma aldosterone concentration (PAC), post-saline infusion test PAC below cut-off value or 24h-urinary aldosterone overexcretion [16–20][23]. The plasma renin concentration (PRC) was measured by chemiluminescent immunoassay (LIAISON®; DiaSorin®, Sallugia, Italia) and was set to 5 mIU/L for the ARR calculation to avoid overestimation of the ARR in subjects with PRC below 5 mIU/L. At HEGP, PAC and total urine aldosterone excretion after pH1 hydrolysis were measured by radioimmunoassay (RIA, Coat-A-Count®; Siemens Medical Solutions Diagnostics®, Saint-Denis, France) until 2016-03-15 and subsequently by Liquid Chromatography coupled to tandem Mass Spectrometry (LC-MS/MS) since then [24]. The ARR and the PAC cut offs for PA diagnosis were defined in accordance with the use of LC-MS/MS as reported previously [24]. PRC. At CHU BDX, PAC and PRC were both measured by chemiluminescent immunoassay (LIAISON®; Diasorin®, Sallugia, Italia) and the cut-offs were the same as those used at HEGP before 2016.

Plasma hormones were measured after discontinuation of antihypertensive medication interfering with the renin-angiotensin-aldosterone system and correction of hypokalemia. Diuretics, angiotensin converting enzyme (ACE) inhibitors, angiotensin II receptor blockers (ARB), beta blockers were withdrawn for 2 weeks at least and mineralocorticoid receptor (MR) blockers for 6 weeks. If needed, non-interfering antihypertensive drugs were prescribed. PAC and plasma renin concentration (PRC) were measured in standardized conditions with blood collected in the morning after 30 minutes rest in a semi-recumbent position. The final diagnosis of PA was retrieved using multidisciplinary expert round decisions, International Classification of Disease (ICD)-10 diagnoses codes, and text analysis using Natural Language Processing (NLP) on medical reports.

The indication of PA screening according to the guidelines was defined as having hypokalemia, resistant HTN, or less than 40 years old. We did not include other indications for PA screening such as severe hypertension mediated organ damages (HMOD) [25,26], as we wanted indications easily accessible during a consult.

Data extraction

A standardized, computerized medical record has been used in daily practice for both clinical care and research purposes since 1975 at HEGP. It is interfaced with the patients' hospital electronic medical record and completed by senior physicians during standard consultations.

The form is divided into blocks of questions (medical history, biological parameters, etc.) [27–29]. These dedicated medical records are stored on an i2b2 clinical data warehouse together with all other hospital health records.

At CHU BDX, clinical data of interest are collected by Clinical Research Associates from the patients' electronic medical record via a form containing various sections (demographics, medical history, treatment, etc.). The form then implements a password-protected ACCESS database.

We extracted age at diagnosis of HTN, family history of HTN, current smoking status, type I or type II diabetes, dyslipidemia, overweight (BMI ≥ 25 kg/m2), office BP (see below) and diagnosis of apparent resistant HTN (RHTN), as well as the standard biological parameters including among others plasma potassium and creatinine concentrations, from the medical records. Estimated glomerular filtration rate (eGFR) was calculated using the Chronic Kidney Disease (CKD-EPI) formula.

In both centers, office blood pressure (OBP) was recorded as the average of three measurements taken by trained nurses after a 5-minute rest in the lying position with a validated semi-automatic oscillometric device [30]. Out of office blood pressure (OOBP) measurements were not systematically performed and thus could not be included in our analysis because of many missing data in the databases. HTN was defined as OBP greater or equal to 140 and/or 90mmHg or HBPM/daytime ABPM above or equal 135/85 mmHg or by regular prescription of antihypertensive medication(s) [25]. Resistant HTN was defined by inadequate control of BP (≥ 140 and/or 90mmHg) in patients treated with optimal or best-tolerated doses of three or more drugs (at best including a diuretic, typically an ACE inhibitor or an ARB, and a CCB) [25].

The study was approved by the institutional review board (Regional Ethics Committee) no. 00011928), and data collection was approved by the French "National Commission for informatics and liberty (CNIL)", (decision DR-2022-226).

The Risk-stratification-based algorithm: PA_{strat}

Generation of patient groups/cluster of interest

The first step of our approach was to generate patient groups/clusters of interest for the diagnosis of PA. We used the Python 3.9 programing language to generate trees with nodes and branches using randomly selected clinical and biological variables after binarization of continuous variables where necessary. The cut-offs for binarization were mean, median, tertiles, or clinically or biologically relevant values where appropriate (Supplementary table 1). Each branch represented a stratified group that included patients with variables strictly inferior to the defined thresholds.

At each node, odds-ratio (OR) for PA were calculated:

$$OR = \frac{N_{PA\,(group)} \Big/ N_{noPA\,(group)}}{N_{PA\,(out\ of\ group)} \Big/ N_{noPA\,(out\ of\ group)}}$$

We used pruning to simplify the tree structures and prevent overfitting, by removing branches in the tree that were not needed for decision-making. Pruning criteria were as follows:

1. *Support cut-off (size of the sample):* less than 200 patients at the node, or less than 100 patients in children node;
2. OR cut-off:
    a. OR > 3 or OR < 0.5;
    b. OR in the children node less discriminating than OR in the present node;

We initially generated 100,000 trees. We then only kept the groups of interest including less than 4 variables to retain simple explainable groups, leading to identification of 6,848 groups. The process of tree generation and an example are illustrated in Figure 2.

Probability recalculation

We assessed the probability of PA in each of the groups constituted above using bootstrapped logistic regression. We performed multivariate logistic regressions in each group of interest on 1000 bootstrapped samples of our development cohort.

Individual probability assessment

We assessed the individual probability of PA for each patient by finding the groups of interest to which they belonged.

We developed the Risk-stratification-based algorithm PA_{strat} maximizing both its sensitivity and its negative predictive value. Indeed, we considered that the most important feature of the Risk-stratification-based algorithm would be its ability to avoid missing patients with a positive diagnosis of PA. Therefore, we assigned each patient to a group with the highest probability of PA.

We built ROC (receiver operating characteristic) curves of the model predictions in comparison to the predicted probability of PA. We used the ROC curve to establish a threshold for a sensitivity of at least 90% in the development cohort. We then tested the external validity of the algorithm to predict PA in patients with a predicted probability above the threshold in the validation cohort (CHU BDX).

Machine learning models

We developed two supervised learning classifier algorithms (logistic regression, and xgboost) [31].

Missing values were imputed using predictive mean matching, using the MICE package in R.

The cohort was randomized in a training (10,854 patients) and a testing (4653 patients) dataset. We used grid search and 10-fold cross-validation in the training dataset to tune the hyperparameters of the model, and Random Oversampling Examples (ROSE) to deal with class imbalance. The algorithm was then tested on the test cohort. Finally, a prediction of PA was made on the external validation cohort.

The development strategy is shown in Supplementary Figure 1.

Statistical analysis

Results are expressed as percentages or means ± standard deviations. Unpaired Student's t tests or Wilcoxon nonparametric tests for continuous variables and chi-squared tests or Fisher's exact test for qualitative variables were used for comparisons between groups.

We calculated confusion matrix to visualize correct predictions for the Risk-stratification-based algorithm, and for the machine learning models. Rows correspond to the true diagnosis, and columns to the predicted diagnosis.

We also calculated the Net Reclassification Improvement index (NRI) to compare our prediction strategy using the stratification-based algorithm with the standard strategy to diagnose secondary hypertension according to the guidelines (cf above) [25,26] using the following characteristics: age < 40 years old, hypokalemia, and RHTN. NRI is defined as the difference between the proportion of correct reclassification of patients using one strategy rather than another [32]. Positive NRI indicates improvement of the new strategy, whereas 0 indicates no improvement, and negative NRI indicates that the alternative strategy is better.

P-values < 0.05 were considered significant.

Analyses were performed using the statistical software RStudio (RStudio server version 4.0.3).

RESULTS

Population characteristics

Derivation cohort

Characteristics of the population of the derivation cohort are shown in Table 1.

15,507 patients were included from the first outpatient visit at HEGP. Mean age was 47.6 years and 7992 (51.5%) were men. A total of 1006 (8.5%) patients had a history of cardiovascular event, 4313 (28.7%) had dyslipidemia, 1737 (11.5%) had diabetes, and 5420 (36%) were current smokers. Office SBP and DBP were 140.4+/-20.7 and 86.9+/-13.5 mmHg, respectively. Hypokalemia was present in 1309 (9.1%) and RHTN in 1368 (8.9%) patients. After standardized workup, PA was diagnosed in 1817 (11.7%) of whole cohort of patients, and among them 711 (34.9%) had hypokalemia and 194 (11%) had RHTN.

Validation cohort

Characteristics of the population of the validation cohort are shown in Table 2.

A total of 768 patients were included from CHU BDX. Mean age was 38.9 years and 435 (56.6%) were men. 123 (16.0%) of patients had a history of cardiovascular event, 132 (17.2%) had dyslipidemia, 59 (7.7%) had diabetes, and 223 (29%) were current smokers. Office SBP and DBP were 148.9+/-18.3 and 88.7+/-13.6 mmHg respectively.

Hypokalemia was present in 106 (13.8%) and RHTN in 23 (3.5%) patients. PA was diagnosed in 142 (18.5%) patients.

Prediction of PA

The results of the performance of the different models with the confusion matrixes are shown in Table 3

Stratification-based algorithm PA_{strat}

Derivation cohort

The AUC of the ROC curve of the PA_{strat} algorithm was 0.80 (Figure 3). According to the ROC curve, a threshold of 29% for the predicted probability of PA was chosen for a sensitivity of the algorithm around 0.90 (0.89). Patients with a predicted probability above this threshold were therefore classified as having predicted PA.

Among the 15,507 patients at HEGP, 8565 had predicted PA (among which 18,9% had PA), whereas 6784 patients had an apparent indication for PA screening according to the guidelines (hypokalemia, resistant HTN, or less than 40 years old), among which 15.9% had PA. 4575 patients with predicted PA would not have been screened for PA according to the guidelines, and among them 631 (13.8 %) had PA (supplementary Table 2). Conversely, 2794 patients had an indication for PA screening according to the guidelines but had no predicted PA by the algorithm. Among them, only 100 (3.6 %) patients had PA (Figure 4).

Validation cohort

Among the 768 patients of CHU BDX, 590 had predicted PA (among which 23% had PA), whereas 454 patients had an apparent indication for PA screening according to the guidelines

(hypokalemia, resistant HTN, or less than 40 years old), among which 13.7% had PA. 243 patients with predicted PA would not have been screened for PA according to the guidelines, and among them 64 (26.3 %) had PA (supplementary Table 2). Conversely, 127 patients had an indication for PA screening according to the guidelines but had no predicted PA by the algorithm. Among them, only 3 (2.4 %) patients had PA (Figure 4).

Only 6 out of 768 patients (0.7%) had PA despite incorrect prediction by the PA_{strat} algorithm. The sensitivity of the algorithm with the chosen threshold of 29% was 0.96, and the negative predictive value was 0.97 (Table 3).

The net reclassification index of the algorithm in comparison with the guidelines-recommended indication of screening was 0.34 (0.42 for the diagnosis of PA, -0.09 for the absence of PA) (supplementary Table 2).

Machine learning models

The ML-based models showed better global performance than the stratification-based algorithm, with AUC of 0.82 for logistic regression and 0.83 for XGBoost (Table 3). According to the ROC curves (Figure 4), thresholds of 0.36 for logistic regression and 0.09 for XGBoost were chosen for a sensitivity of 0.90.

When applying the models to the validation cohort, the models kept good global performances with accuracy of respectively 0.82 and 0.79, but they had lower negative predictive values with respectively 49 (6.7%) and 84 (10.9%) false negatives (negative predictive value respectively 0.87 and 0.88) (Table 3).

The net reclassification index in comparison with the guidelines-recommended indication of screening were better for diagnosing the absence of PA and were respectively 0.48 (-0.05 for the diagnosis of PA, 0.54 for the absence of PA) and 0.55 (0.06 for the diagnosis of PA, 0.55 for the absence of PA) for logistic regression and XGBoost (supplementary Table 2)

The importance of features used in the models are shown in supplementary Figure 2.

DISCUSSION

In this study, we developed the PA_{strat} algorithm to detect patients with a high probability of PA. The algorithm was based on simple clinical and biological data. Compared with the main indications of screening recommended by the ESH guidelines, the PA_{strat} algorithm increased the proportion of patients to be screened, and among those patients, 26% had PA and would not have been screened. Besides, only 0.8% of patients had a false negative prediction of PA, and among the patients not screened according to the algorithm, but with an indication according to the guidelines, only 2.4% had PA. The machine-learning models had better global performances than the PA_{strat} algorithm but resulted in more false-negative in the validation cohort, with the disadvantage of not being interpretable. Conversely, in the PA_{strat} algorithm, the probability of PA originates from a group of interest, based on clinical and biological characteristics, which can be easily retrieved for each patient, and therefore allow the user to understand the classification.

The underdiagnosis of PA is a major health issue as this pathology is responsible for mortality excess, with target organ damages independent of PA levels [9,12]. The diagnosis can be difficult, therefore the screening is very important and must be simple to be usable at the office of general practioners, which is why the recent ESC guidelines for the management of elevated blood pressure and hypertension suggest the screening of PA by renin and aldosterone measurements in all adults with confirmed hypertension. Our results are in line with this recommendation, as our algorithm increases the number of patients to be screened to avoid missing PA [26].

Our algorithm was designed as a screening tool for PA screening, maximizing sensitivity and negative predictive value. Many parameters intervene in the decision of a clinician to pursue the diagnosis path to the next step, or on the contrary to stop investigations [33]. With limited performances, prediction models must not be accountable for the final decision, but must be a tool for the clinician in a dynamic refinement process of diagnosis [33]. However, it can be difficult when using artificial intelligence models that behave like black-boxes [34].Our algorithm does not aim to give a definite and non-debatable answer. Moreover, the absence of explicability of a model can have a negative influence on the relationship with patients, for example by increasing patients' skepticism [35]. With this in mind, our method of risk stratification is fully explainable, as the probability given by the algorithm can be associated with the clinical group the patient has been linked to. Another constraint is that to be effective, a diagnosis tool must be easy enough to use, and with sufficiently available data, to be used by the greater number of physicians, and especially general practitioner. The algorithm had to be sufficiently explainable, and with sufficiently available data to be used by any physician during a consult. The machine-learning based models, however, had comparable performances, but lack interpretability on specific predictions.

Models to predict secondary hypertension exist, but they are often biased. For example, Extreme Gradient Boosting (XGBoost) was used to develop 5 models to predict 4 etiologies of secondary hypertension using EMR. However, secondary hypertension was only defined using the International Classification of Disease (ICD-10), which we know is very biased in many ways, and moreover, features introduced in the model were sometimes part of the definition of the diagnosis itself (for instance renal artery stenosis on CT for renovascular hypertension) [36]. Also, 2 scores have been published to predict PA. However, they were developed in small datasets, their performance was limited, and more importantly they had no proper validation [37,38]. Wang et al. developed a nomogram to predict PA based on the

records of 806 patients AUC was 0.68 and there was no external validation[17]. Manosroi et al. developed a score based on logistic regression of the data of 305 patients. However AUC was only 0.71 and no external validation was performed [18]. Another recent score was developed in a large database of US veterans, but its performance were not sufficient for clinical use, with more than 5% of patients with a score of 0 having PA [19]. Finally, Buffolo et al. developed a clinical score and a machine-learning based score in a larger population, but the performance of the clinical score dropped significantly when applied to the external cohort of validation regarding the positivity of PA screening., with an AUC of 0.600 and 0.617 respectively [20].

A strength of this study lies in the large sample size of the development cohort, with many data including medical history, clinical exam, and biology, which makes it one of the largest study regarding PA prediction. There also are limitations. First, the algorithm was developed in patients referred to or consulting in a tertiary center, sometimes for suspicion of secondary HTN, including PA. The validation cohort was also extracted from a tertiary center, and therefore the prevalence of PA was high in this cohort. However, as the negative predictive value lowers with prevalence, we can expect our algorithm to be more performant in general population. Nevertheless, this indicates the need of another external validation in the general hypertensive population, which is the target population of this algorithm. Moreover, some patients already had non-interfering treatments such as alpha-blockers when they were referred, which is why non-interfering treatments intervened in the group selection. However, as those patients had a higher probability of PA, this also can be seen as an internal validation of our algorithm, which detected those signals. Of note, our algorithm was developed in patients under 65 years old, because older patients are not screened for PA, which is therefore underdiagnosed. However, in the validation cohort, every patients were included because they had a full work-up for secondary HTN including PA screening. Moreover, the bias induced is against the algorithm because nonspecific rule for older patients was included, and that as PA is less frequent in this population, limiting the validation to patients under 65 years old would overrepresent PA.

In conclusion, we provide a stratification algorithm for PA screening, which can be simply used in clinical practice, is fully explainable and have the potential to improve PA screening. We are currently developing a web interface to make it easily accessible to any general practitioner. Other external validations of this score in a general population must be done in order to confirm its interest.

ACKNOWLEDGEMENTS

None

SOURCES OF FUNDING

This study was not financially supported.

DISCLOSURES

None

REFERENCES

1 Rossi GP, Bernini G, Caliumi C, Desideri G, Fabris B, Ferri C, *et al.* A Prospective Study of the Prevalence of Primary Aldosteronism in 1,125 Hypertensive Patients. *Journal of the American College of Cardiology* 2006; 48:2293–2300.

2 Plouin P-F, Amar L, Chatellier G, on behalf of the COMETE-Conn Study Group. Trends in the prevalence of primary aldosteronism, aldosterone-producing adenomas, and surgically correctable aldosterone-dependent hypertension. *Nephrology Dialysis Transplantation* 2004; 19:774–777.

3 Monticone S, Burrello J, Tizzani D, Bertello C, Viola A, Buffolo F, *et al.* Prevalence and Clinical Manifestations of Primary Aldosteronism Encountered in Primary Care Practice. *J Am Coll Cardiol* 2017; 69:1811–1820.

4 Buffolo F, Monticone S, Burrello J, Tetti M, Veglio F, Williams TA, *et al.* Is Primary Aldosteronism Still Largely Unrecognized? *Horm Metab Res* 2017; 49:908–914.

5 Brown JM, Siddiqui M, Calhoun DA, Carey RM, Hopkins PN, Williams GH, *et al.* The Unrecognized Prevalence of Primary Aldosteronism: A Cross-sectional Study. *Ann Intern Med* 2020; 173:10–20.

6 Douma S, Petidis K, Doumas M, Papaefthimiou P, Triantafyllou A, Kartali N, *et al.* Prevalence of primary hyperaldosteronism in resistant hypertension: a retrospective observational study. *Lancet* 2008; 371:1921–1926.

7 Parasiliti-Caprino M, Lopez C, Prencipe N, Lucatello B, Settanni F, Giraudo G, *et al.* Prevalence of primary aldosteronism and association with cardiovascular complications in patients with resistant and refractory hypertension. *Journal of Hypertension* 2020; 38:1841–1848.

8 Reincke M, Bancos I, Mulatero P, Scholl UI, Stowasser M, Williams TA. Diagnosis and treatment of primary aldosteronism. *Lancet Diabetes Endocrinol* 2021; 9:876–892.

9 Stowasser M, Sharman J, Leano R, Gordon RD, Ward G, Cowley D, *et al.* Evidence for Abnormal Left Ventricular Structure and Function in Normotensive Individuals with Familial Hyperaldosteronism Type I. *The Journal of Clinical Endocrinology & Metabolism* 2005; 90:5070–5076.

10 Rossi GP, Bernini G, Desideri G, Fabris B, Ferri C, Giacchetti G, *et al.* Renal damage in primary aldosteronism: results of the PAPY Study. *Hypertension* 2006; 48:232–238.

11 Monticone S, D'Ascenzo F, Moretti C, Williams TA, Veglio F, Gaita F, *et al.* Cardiovascular events and target organ damage in primary aldosteronism compared with essential hypertension: a systematic review and meta-analysis. *The Lancet Diabetes & Endocrinology* 2018; 6:41–50.

12 Marney AM, Brown NJ. Aldosterone and end-organ damage. *Clin Sci (Lond)* 2007; 113:267–278.

13 Amar L, Baguet JP, Bardet S, Chaffanjon P, Chamontin B, Douillard C, *et al.* SFE/SFHTA/AFCE primary aldosteronism consensus: Introduction and handbook. *Annales d'Endocrinologie* 2016; 77:179–186.

14 de Freminville J-B, Amar L, Azizi M, Mallart-Riancho J. Endocrine causes of hypertension: literature review and practical approach. *Hypertens Res* 2023; 46:2679–2692.

15 Hundemer GL, Curhan GC, Yozamp N, Wang M, Vaidya A. Cardiometabolic outcomes and mortality in medically treated primary aldosteronism: a retrospective cohort study. *The Lancet Diabetes & Endocrinology* 2018; 6:51–59.

16 Mulatero P, Monticone S, Deinum J, Amar L, Prejbisz A, Zennaro M-C, *et al.* Genetics, prevalence, screening and confirmation of primary aldosteronism: a position statement and consensus of the Working Group on Endocrine Hypertension of The European Society of Hypertension ∗. *Journal of Hypertension* 2020; 38:1919–1928.

17 Wang M-H, Li N-F, Luo Q, Wang G-L, Heizhati M, Wang L, *et al.* Development and validation of a novel diagnostic nomogram model to predict primary aldosteronism in patients with hypertension. *Endocrine* 2021; 73:682–692.

18 Manosroi W, Tacharearnmuang N, Atthakomol P. Clinical and biochemical predictors and predictive model of primary aldosteronism. *PLoS One* 2022; 17:e0272049.

19 Ginzberg SP, Kalva S, Wirtalla CJ, Passman JE, Cohen DL, Cohen JB, *et al.* Development of a risk-prediction model for primary aldosteronism in veterans with hypertension. *Surgery* 2024; 175:73–79.

20 Buffolo F, Burrello J, Burrello A, Heinrich D, Adolf C, Müller LM, *et al.* Clinical Score and Machine Learning-Based Model to Predict Diagnosis of Primary Aldosteronism in Arterial Hypertension. *Hypertension* Published Online First: 7 September 2021. doi:10.1161/HYPERTENSIONAHA.121.17444

21 Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; 385:e078378.

22 Funder JW, Carey RM, Mantero F, Murad MH, Reincke M, Shibata H, *et al.* The Management of Primary Aldosteronism: Case Detection, Diagnosis, and Treatment: An Endocrine Society Clinical Practice Guideline. *J Clin Endocrinol Metab* 2016; 101:1889–1916.

23 Cornu E, Steichen O, Nogueira-Silva L, Küpers E, Pagny J-Y, Grataloup C, *et al.* Suppression of Aldosterone Secretion After Recumbent Saline Infusion Does Not Exclude Lateralized Primary Aldosteronism. *Hypertension* 2016; 68:989–994.

24 Baron S, Amar L, Faucon A-L, Blanchard A, Baffalie L, Faucard C, *et al.* Criteria for diagnosing primary aldosteronism on the basis of liquid chromatography-tandem mass spectrometry determinations of plasma aldosterone concentration. *J Hypertens* 2018; 36:1592–1601.

25 Mancia Chairperson G, Kreutz Co-Chair R, Brunström M, Burnier M, Grassi G, Januszewicz A, *et al.* 2023 ESH Guidelines for the management of arterial hypertension The Task Force for the management of arterial hypertension of the European Society of Hypertension Endorsed by the European Renal Association (ERA) and the International Society of Hypertension (ISH). *J Hypertens* Published Online First: 21 June 2023. doi:10.1097/HJH.0000000000003480

26 McEvoy JW, McCarthy CP, Bruno RM, Brouwers S, Canavan MD, Ceconi C, *et al.* 2024 ESC Guidelines for the management of elevated blood pressure and hypertension: Developed by the task force on the management of elevated blood pressure and hypertension of the European Society of Cardiology (ESC) and endorsed by the European Society of Endocrinology (ESE) and the European Stroke Organisation (ESO). *European Heart Journal* 2024; :ehae178.

27 Jeunemaitre X, Degoulet P, Morice V, Chatellier G, Devries C, Plouin PF, *et al.* [Testing an expert system for hypertension]. *Arch Mal Coeur Vaiss* 1986; 79:808–812.

28 Degoulet P, Chatellier G, Devriès C, Lavril M, Ménard J. Computer-assisted techniques for evaluation and treatment of hypertensive patients. *Am J Hypertens* 1990; 3:156–163.

29 Degoulet P, Marin L, Lavril M, Le Bozec C, Delbecke E, Meaux J-J, *et al.* The HEGP component-based clinical information system. *Int J Med Inform* 2003; 69:115–126.

30 La Batide-Alanore A, Chatellier G, Bobrie G, Fofol I, Plouin PF. Comparison of nurse- and physician-determined clinic blood pressure levels in patients referred to a hypertension clinic: implications for subsequent management. *J Hypertens* 2000; 18:391–398.

31 Padmanabhan S, Tran TQB, Dominiczak AF. Artificial Intelligence in Hypertension: Seeing Through a Glass Darkly. *Circ Res* 2021; 128:1100–1118.

32 Cook NR. Quantifying the added value of new biomarkers: how and how not. *Diagnostic and Prognostic Research* 2018; 2:14.

33 Adler-Milstein J, Chen JH, Dhaliwal G. Next-Generation Artificial Intelligence for Diagnosis: From Predicting Diagnostic Labels to "Wayfinding." *JAMA* 2021; 326:2467.

34 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 2021; 3:e745–e750.

35 Funer F. The Deception of Certainty: how Non-Interpretable Machine Learning Outcomes Challenge the Epistemic Authority of Physicians. A deliberative-relational Approach. *Med Health Care Philos* 2022; 25:167–178.

36 Diao X, Huo Y, Yan Z, Wang H, Yuan J, Wang Y, *et al.* An Application of Machine Learning to Etiological Diagnosis of Secondary Hypertension: Retrospective Study Using Electronic Medical Records. *JMIR Med Inform* 2021; 9:e19739.

37 Yamashita T, Shimizu S, Koyama M, Ohno K, Mita T, Tobisawa T, *et al.* Screening of primary aldosteronism by clinical features and daily laboratory tests: combination of urine pH, sex, and serum K. *J Hypertens* 2018; 36:326–334.

38 Kietsiriroje N, Wonghirundecha R, Suntornlohanakul O, Murray RD. Construction of a predictive scoring system as a guide to screening and confirmation of the diagnosis of primary aldosteronism. *Clin Endocrinol (Oxf)* 2020; 92:196–205.

PATHOPHYSIOLOGICAL NOVELTY AND RELEVANCE

What is new?

Primary aldosteronism (PA) is one of the main causes of secondary hypertension and is probably underdiagnosed, notably because its screening is **challenging**. We report here PA_{strat}, a new stratification-based algorithm, and machine-learning models, to help define the population of patients who should be screened for PA.

What is relevant?

When using the PA_{strat} algorithm, only 6 out of 768 patients (1.3%) had PA despite incorrect prediction by the PA_{strat} algorithm. Conversely, among the 128 patients who had an indication for PA screening according to the guidelines but not according to the algorithm, only 3 (0.23%) patients had PA. When using machine-learning models, despite good performances, there were more false negatives.

Clinical/Pathophysiological implications?

Our algorithm for PA screening can be simply used in clinical practice, and is fully explainable. It may have the potential to improve PA screening, which is a major health issue.

TABLES

Table 1- clinical characteristics of patients of the derivation cohort

	Total	Number	No PA	PA	p
	15507		13690	1817	
Male sex, No. (%)	7992 (51.5)	15507	6912 (50.5)	1080 (59.4)	<0.001
Age at diagnosis of HTN, mean (SD), y	47.6 (12.0)	15507	47.6 (12.2)	47.5 (10.1)	0.619
Familial history of HTN, No. (%)	8052 (61.7)	13040	7093 (61.5)	959 (63.4)	0.173
Current smoker, No. (%)	5420 (36.0)	15057	4883 (36.7)	537 (30.5)	<0.001
Diabetes, No. (%)	1737 (11.5)	15083	1474 (11.1)	263 (14.8)	<0.001
Dyslipidemia, No. (%)	4313 (28.7)	15048	3816 (28.7)	497 (28.1)	0.583
BMI, mean (SD), kg/m²	27.79 (5.52)	14854	27.63 (5.55)	28.97 (5.13)	<0.001
Office systolic BP, mean (SD), mmHg	140.38 (20.75)	15181	139.69 (20.74)	145.63 (20.03)	<0.001
Office diastolic BP, mean (SD), mmHg	86.93 (13.50)	15181	86.43 (13.53)	90.71 (12.71)	<0.001
Number of antihypertensive drugs, mean (SD)	1.75 (1.53)	15507	1.68 (1.52)	2.33 (1.49)	<0.001
Resistant HTN, No. (%)	1368 (8.9)	15180	1174 (8.8)	194 (11.0)	0.003
Potassium concentration, mean (SD), mmol/L	4.07 (0.48)	14314	4.12 (0.44)	3.68 (0.53)	<0.001
eGFR, mean (SD), ml/min	80.3 (22.2)	14514	80.3 (22.1)	80.2 (23.3)	0.846
HTN duration (mean (SD))	7.41 (8.71)	14615	7.16 (8.66)	9.21 (8.85)	<0.001
Cardiovascular disease, No (%)	1006 (8.5)	15507	886 (8.6)	120 (8.2)	0.694
LDL, mean(SD) mmol/L	3.1 (1.0)	13238	3.1 (1.0)	3.0 (0.90)	<0.001

Table 2 – characteristics of the validation cohort

	Validation cohort	Derivation cohort	p	Number (validation cohort)
	768	15507		
Male sex, No. (%)	435 (56.6)	7992 (51.5)	0.006	768
Age at diagnosis of HTN, mean (SD), y	39.89 (12.24)	47.62 (11.99)	<0.001	768
Familial history of HTN, No. (%)	294 (38.3)	8052 (61.7)	<0.001	768
Current smoker, No. (%)	223 (29.0)	5420 (36.0)	<0.001	768
Diabetes, No. (%)	59 (7.7)	1737 (11.5)	0.001	768
Dyslipidemia, No. (%)	132 (17.2)	4313 (28.7)	<0.001	768
BMI, mean (SD), kg/m²	28.54 (5.84)	27.79 (5.52)	<0.001	767
Office systolic BP, mean (SD), mmHg	148.9 (18.3)	140.4 (20.8)	<0.001	753
Office diastolic BP, mean (SD), mmHg	88.7 (13.6)	86.9 (13.5)	0.001	753
Number of antihypertensive drugs, mean (SD)	1.38 (1.06)	1.75 (1.53)	<0.001	753
Resistant HTN, No. (%)	23 (3.5)	1368 (9.0)	<0.001	753
Plasma potassium concentration, mean (SD), mmol/L	3.85 (0.38)	4.07 (0.48)	<0.001	764
eGFR, mean (SD), ml/min	102.16 (18.10)	80.30 (22.23)	<0.001	766
HTN duration (mean (SD))	7.16 (8.66)	9.21 (8.85)	<0.001	756
Cardiovascular disease, No (%)	123 (16.0)	1006 (8.5)	0.694	768
LDL, mean(SD) mmol/L	2.4 (2.6)	3.1 (1.0)	<0.001	757

Table 3 – Diagnostic performance of the PA_{strat} algorithm and the ML models

A – PA_{strat} algorithm

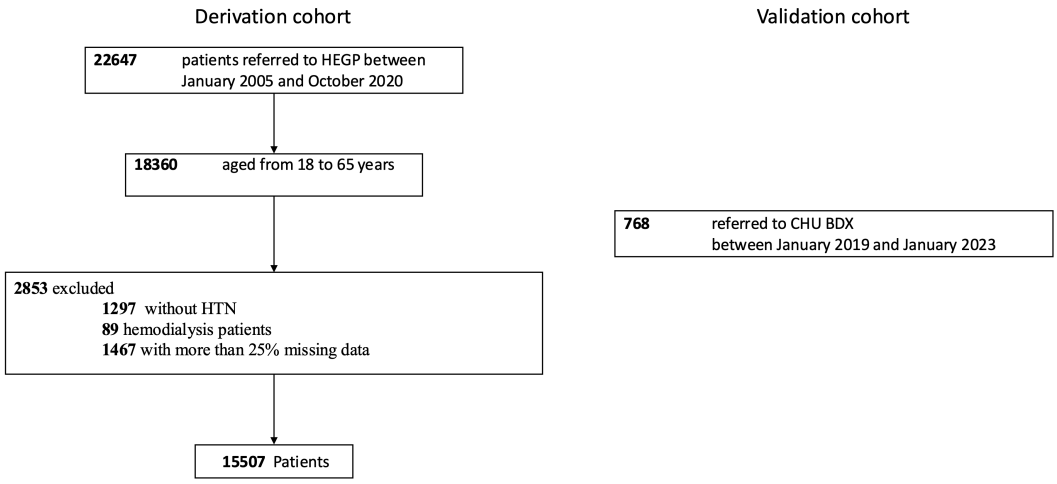
	derivation Cohort (N = 15,507)				Validation cohort (N = 768)		
Stratification algorithm			Predicted PA			Predicted PA	
			No	Yes		No	Yes
	True PA	No	6745	6945		172	454
		Yes	197	1620		6	136
Sensitivity		0.89			0.96		
Specificity		0.49			0.23		
Accuracy		0.54			0.40		
NPV		0.97			0.97		
AUC		0.80					

B – ML models

	Testing cohort (N = 4653)				Validation cohort (N = 768)		
Logistic Regression			Predicted PA			Predicted PA	
			No	Yes		No	Yes
	True PA	No	1841	2266		577	88
		Yes	45	501		49	54
Sensitivity		0.92			0.52		
Specificity		0.44			0.87		
Accuracy		0.50			0.82		
NPV		0.98			0.92		
AUC		0.82			0.82		
			Predicted PA			Predicted PA	
			No	Yes		No	Yes
XGBoost		No	1979	2128		542	71
	Actual PA	Yes	52	494		84	71
Sensitivity		0.90			0.46		
Specificity		0.48			0.88		
Accuracy		0.53			0.79		
NPV		0.97			0.86		
AUC		0.83			0.82		

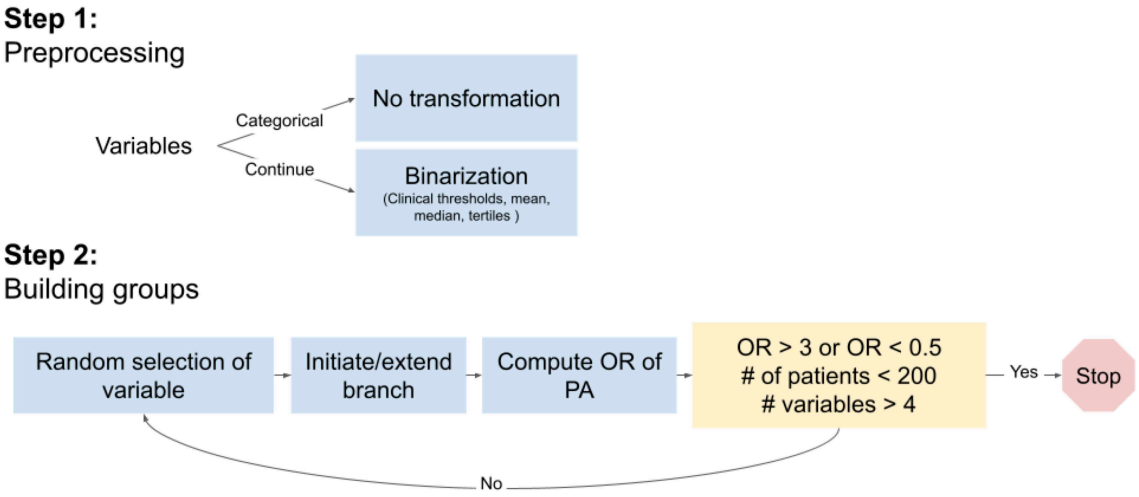
FIGURES

Figure 1: Flow chart - Patients seen in HEGP and CHU BDX

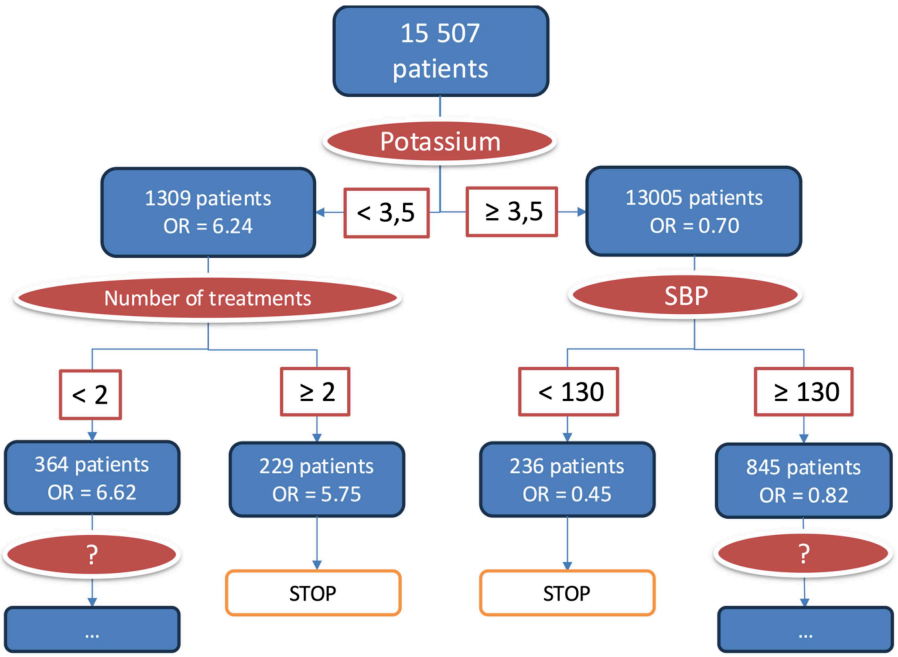


<u>Figure 2: Tree generation</u> – Process and example of tree generation for the PA$_{STRAT}$ algorithm

A – Generation process

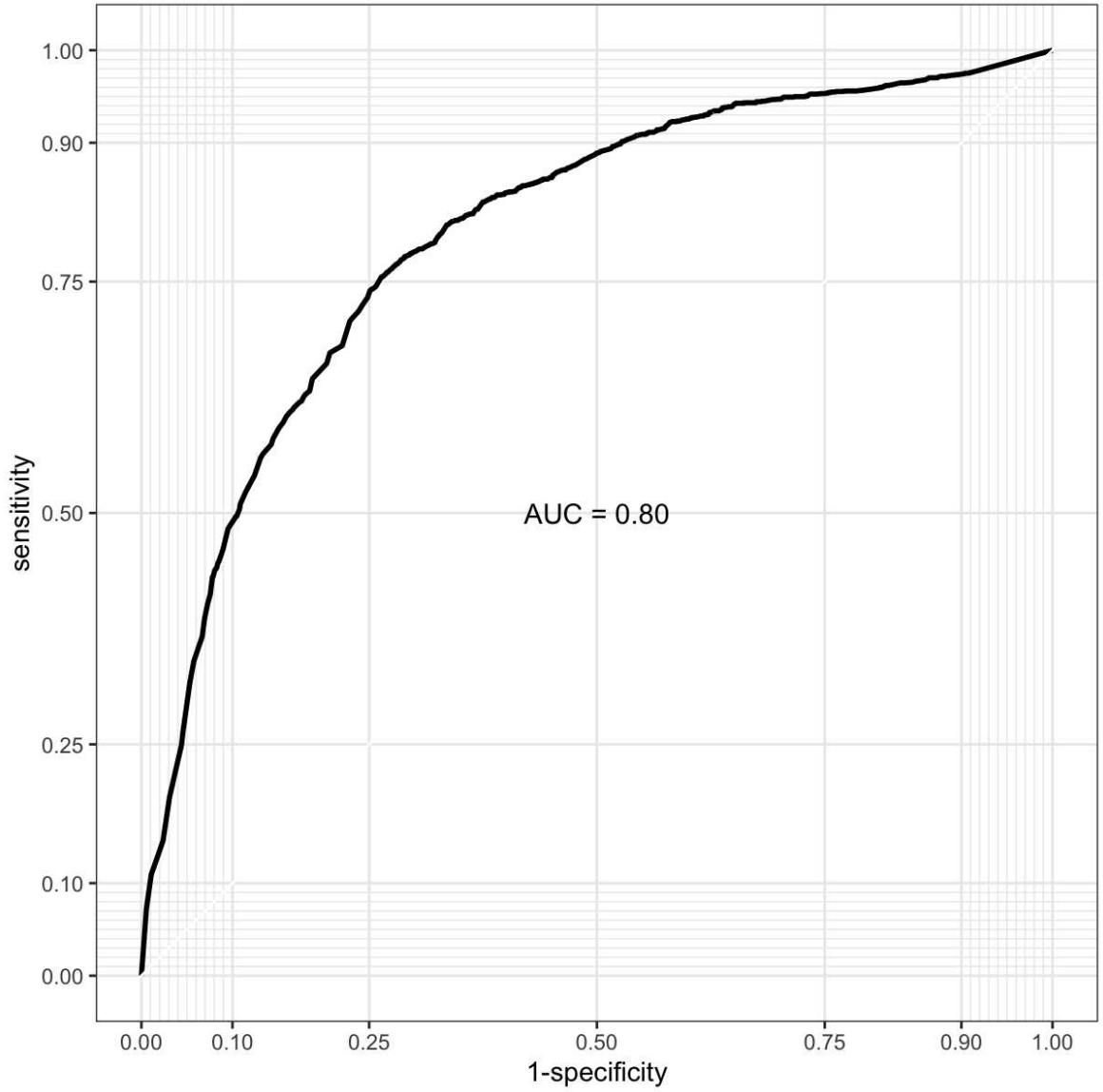


B - Example

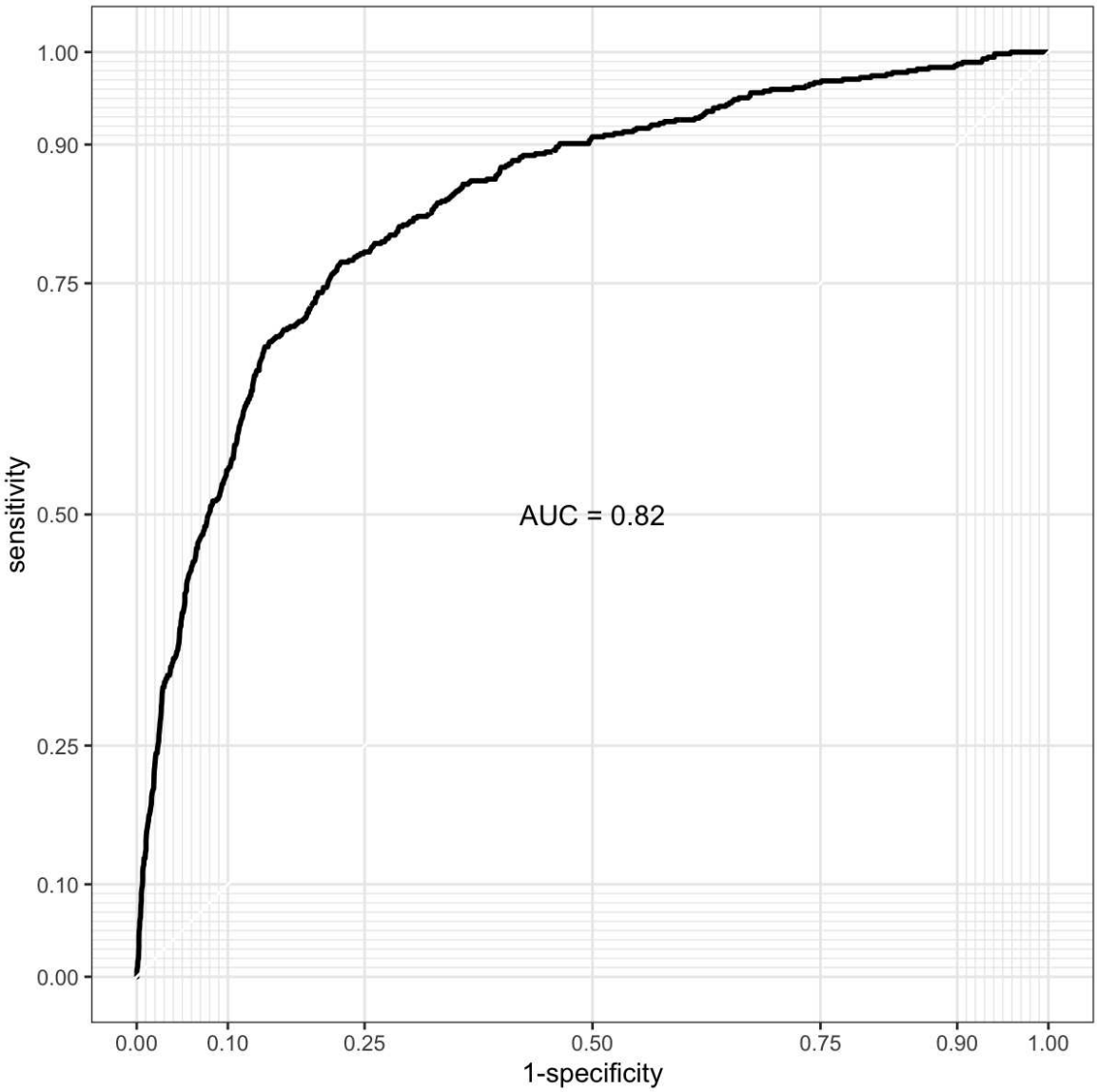


<u>Figure 3: ROC curves –</u> ROC curve of the PA$_{STRAT}$ algorithm and the machine learning models

A - PA$_{STRAT}$ algorithm



B – Logistic Regression


1.00
0.90
0.75
0.50
0.25
0.10
0.00
sensitivity
AUC = 0.82
0.00
0.10
0.25
0.50
0.75
0.90
1.00
1-specificity

C – XGBoost

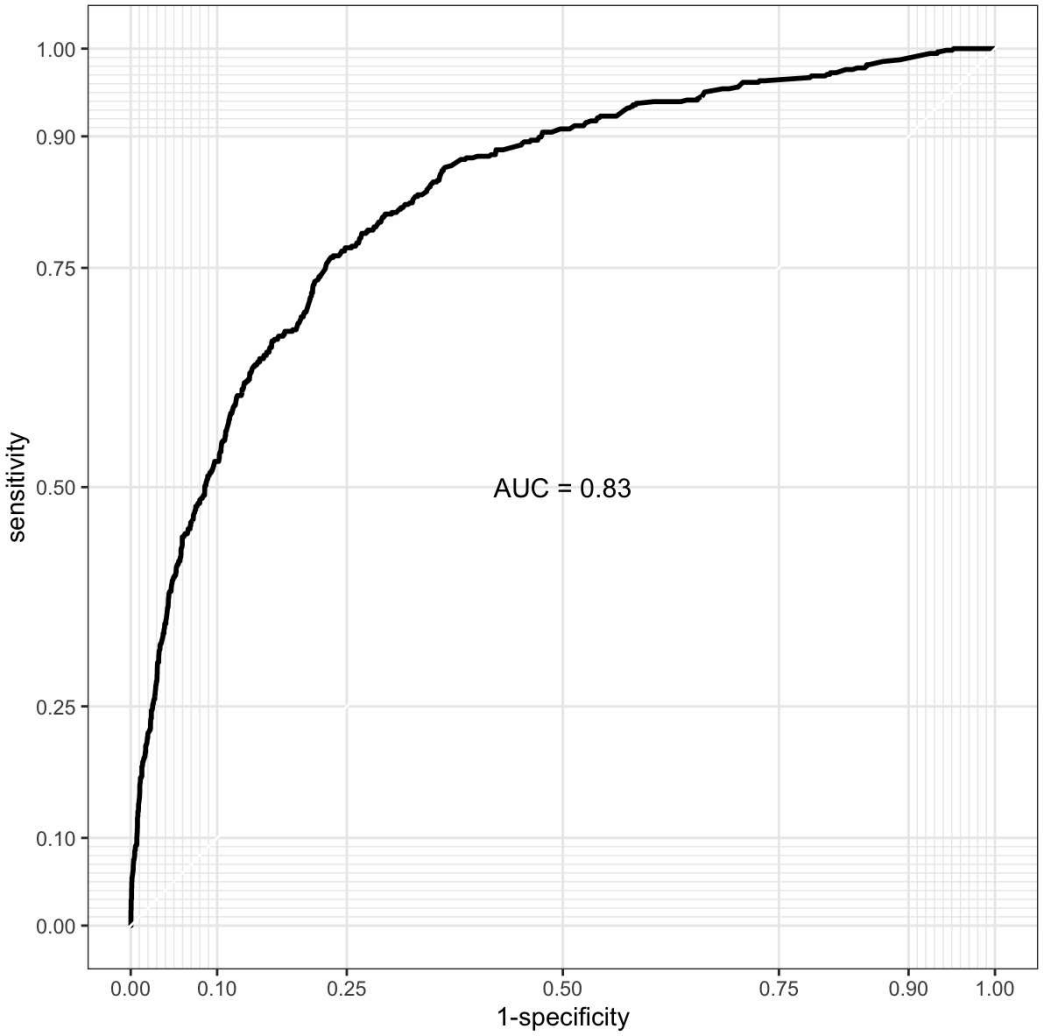
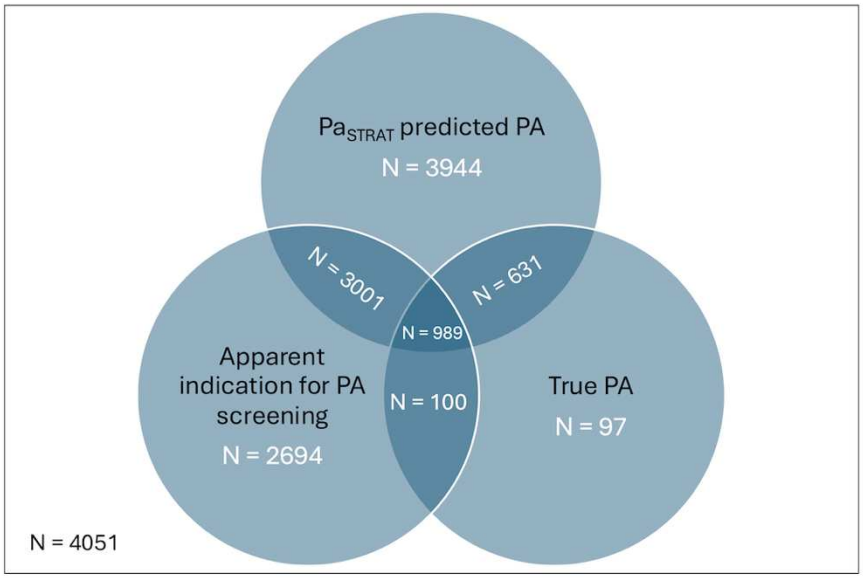
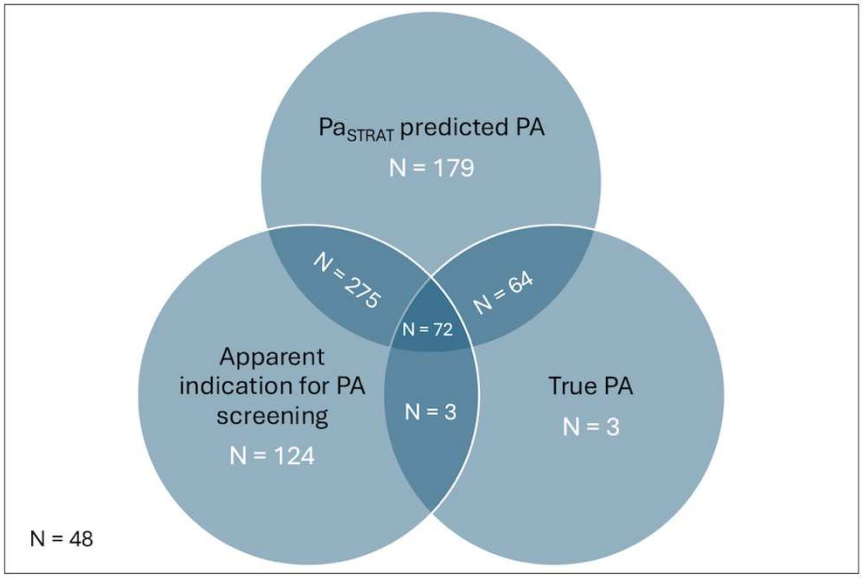

1.00
0.90
0.75
0.50
0.25
0.10
0.00
sensitivity
AUC = 0.83
0.00
0.10
0.25
0.50
0.75
0.90
1.00
1-specificity

Figure 4 : Venn diagram of the PA_{STRAT} algorithm – Venn diagram of the PA_{STRAT} algorithm on the derivation and the validation cohort

Derivation cohort



Validation cohort



SUPPLEMENTAL DATA

Supplementary table 1 – features used in the models and thresholds for variable binarization

Binarized variables/features	threshold 1	threshold 2	threshold 3
abdominal perimeter	88.0	102.0	
age	40.0	50.0	66.0
body mass index (kg/m2)	25.0	30.0	
diastolic blood pressure (mmHg)	80.0	90.0	100.0
glycemia (mmol/l)	6.1		
hdl cholesterol in men (mmol/l)	1.3		
hdl cholesterol in women (mmol/l)	1.0		
heart rate	68.0	79.0	
height (m)	1.65	1.74	
HTN duration	2.0	8.0	
ldl cholesterol (mmol/l)	2.6	4.9	
maximum systolic blood pressure (mmHg)		200.0	
number of treatments	2.0	3.0	
potassium (mmol/l)	3.5	4.0	4.5
systolic blood pressure (mmHg)	130.0	140.0	160.0
triglycerides	1.69		
Binary variables/features			
Familial history of HTN			
Cardiovascular disease			
Current smoking			
Diabetes			
Dyslipidemia			
Depression History			
HTN treatment intolerance			
Sleep Apnea			
Headache			
Palpitations			
Sweating			
Alphablocker treatment			
Antialdosteron treatment			
Betablocker treatment			
RAS blocker treatment			
Centrally active treatment			
Loop diuretic treatment			
Potassium sparing treatment			
Calcium blocker treatment			
Thiazide diuretic treatment			
Statin treatment			

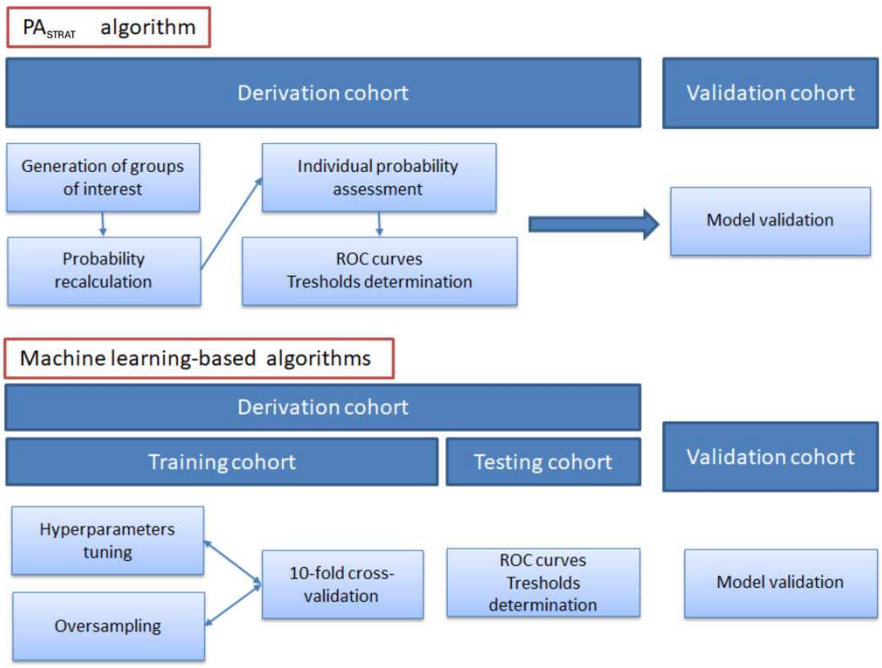
Supplementary table 2 – NRI calculation

PA		**PA_{strat}**					
no PA							
		no	yes	Partial total	Total		
Guidelines	no	3	64	67	293	NRI PA	0,42957746
		47	179	226		NRI no PA	-0,086262
	yes	3	72	75	475	NRI total	0,34331548
		125	275	400			
Partial total		6	136	142			
		172	454	626			
Total		178	590		768		

PA		**Logistic Regression**					
no PA							
		no	yes	Partial total	Total		
Guidelines	no	54	26	80	315	NRI PA	-0,056338
		213	22	235		NRI no PA	0,54632588
	yes	34	28	62	453	NRI total	0,48998785
		364	27	391			
Partial total		88	54	142			
		577	49	626			
Total		665	103		768		

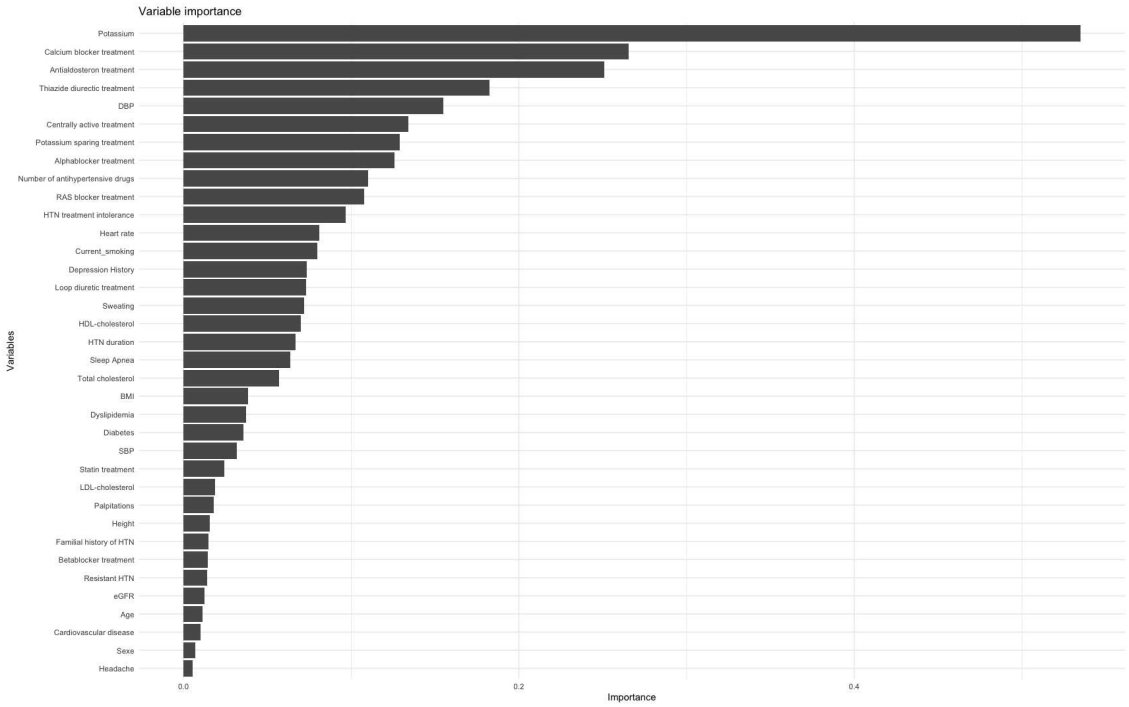
PA		**XGBoost**					
no PA							
		no	yes	Partial total	Total		
Guidelines	no	41	39	80	315	NRI PA	0,06338028
		192	43	235		NRI no PA	0,49041534
	yes	30	32	45	436	NRI total	0,55379562
		350	41	391			
Partial total		71	71	142			
		542	84	626			
Total		572	155		768		

Supplementary Figure 1 – Algorithms' development

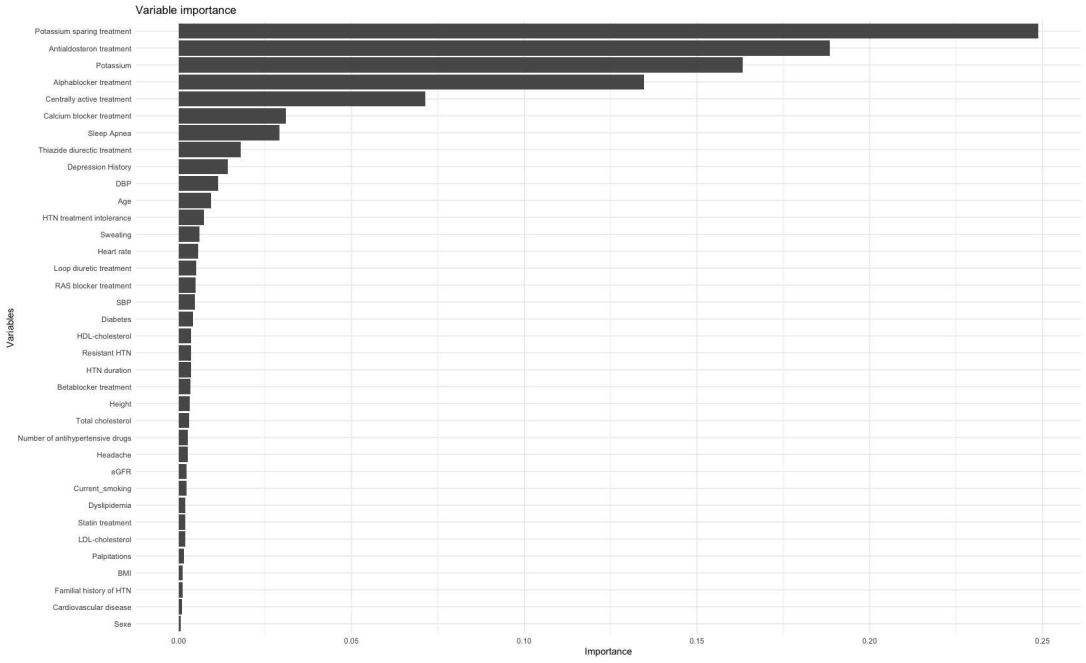


Supplementary Figure 2 – Importance of features used in machine learning predictive algorithms

A – Logistic Regression



B – XGBoost



Complément de méthodes et résultats

Méthodes complémentaires : développement de l'algorithme de stratification

Nous avons développé une méthode de génération de groupes afin de stratifier les patients selon différentes variables, sur le principe de la génération d'arbres de décision.

La génération de ces groupes a été réalisée à partir d'un programme Python.

Nous nous sommes inspirés de la création des arbres de décision C4.5 (issus l'architecture ID3)[126], ainsi que des travaux sur l'identification de motifs fréquents. (*apriori*, *prefixpan*…).[68,69]

pour la structure de l'algorithme de stratification.

Nous avons créé une classe « arbres » comprenant :

- Des nœuds
- Un ensemble d'exemples (patients) associé à chaque nœud

Processus de création d'un arbre :

`Initialisation :`

L'ensemble initial est l'ensemble de la base de données de l'HEGP (15,507 patients, 37 variables, 1)

On choisit aléatoirement une variable et une valeur.

Puis, à chaque étape, pour chaque nœud :

- On compte les diagnostics des patients de l'ensemble associé
- On calcule l'odds-ratio (OR) du diagnostic d'intérêt par rapport à la population globale :

$$OR = \frac{N_{diag\ (noeud)} \Big/ N_{nondiag\ (noeud)}}{N_{diag\ (en\ dehors\ du\ noeud)} \Big/ N_{nondiag\ (en\ dehors\ du\ noeud)}}$$

- On choisit aléatoirement une nouvelle variable qui donnera 2 nœud suivants (nœuds enfants), un pour chaque valeur de la variable

Concernant les choix de variables, certaines règles sont appliquées :

- La variable « sexe » avec la valeur « homme » ne peut être associée dans un arbre aux variables avec des seuils pour les femmes (périmètre abdominal, HDL, LDL)
- La variable « sexe » avec la valeur « femme » ne peut être associée dans un arbre aux variables avec des seuils pour les hommes (périmètre abdominal, HDL, LDL)
- Une variable ne peut intervenir qu'une seule fois dans chaque arbre, même avec un seuil différent (par exemple ka_1, correspondant au seuil de 3,5 mmol/L pour le potassium, exclut les autres seuils de potassium au sein de l'arbre

Les règles suivantes d'appliquent ensuite pour le choix des groupes finaux :

- Si la taille de l'ensemble associé à un nœud est < 100, ou si la taille de l'ensemble associé à un des nœuds enfants est < 50, on arrête le développement de l'arbre
- Si l'OR du nœud enfant est moins intéressant que celui du nœud actuel (inférieur si positif, ou supérieur si négatif), on arrête le développement de l'arbre

Après arrêt du développement :

- Si l'OR est > 3 ou < 0.5, et la taille du groupe est > 10, le groupe final constitué des variables et valeurs associée au noeud et à ses parents est conservé

Ce processus a été répété 100,000 fois.

6848 groupes ont été générés.

Recalcul de probabilité

La probabilité de chaque diagnostic a été à nouveau estimée au sein de chaque groupe, en utilisant une méthode de *bootstrap* (rééchantillonnage avec remise).

Pour chaque groupe d'intérêt identifié avec la méthode présentée ci-dessus, composé de variables avec des valeurs associées :

1. Nous avons estimé la probabilité de ce diagnostic en réalisant 1000 tirages avec remise (*bootstrap*) sur l'ensemble de la population. Il s'agit de la méthode du *bootstrap* décrite en partie 1.
2. La probabilité du diagnostic a été estimée sur chaque échantillon *bootstrap* à partir d'un modèle de régression logistique multivarié prenant en paramètres les variables du groupe (cf partie 4)

3. Nous avons associé à chaque groupe une probabilité moyenne de chaque diagnostic, avec leurs intervalles de confiance

Les 10 groupes avec les probabilités les plus importantes pour nos 3 diagnostics d'intérêt sont montrés dans les tables 5,6 et 7

Tableau 5 - groupes avec les probabilités d'HAP les plus élevées

PROBA HAP	**PROBA RENOVASC**	**PROBA PPGL**	**PROBA HTAE**	**GROUPE**
68.3	8.62	0.955	31.0	ka_1_0 / ttt_hta_alphab_1
56.7	5.16	0.566	46.0	ka_2_0 / ttt_hta_epargneur_k_1
56.4	10.4	1.53	42.1	ttt_hta_alphab_1 / ttt_hta_antialdo_1
55.3	8.66	0.335	48.3	ka_3_0 / ttt_hta_antialdo_1 / ttt_hta_central_1
54.6	6.07	0.286	46.1	atcd_intolerance_ttt_hta_1 / ka_1_0
54.5	7.73	0.394	41.9	ka_1_0 / nb_ttt_2_1
53.2	6.40	1.00	42.8	diabete_1 / ka_1_0
53.1	6.16	0.446	44.6	anciennete_hta_2_1 / ka_1_0
52.7	10.1	0.777	50.5	atcd_intolerance_ttt_hta_1 / ka_2_0 / nb_ttt_2_…
52.6	5.22	0.245	45.9	ka_1_0 / pad_3_1

Tableau 6 - groupes avec les probabilités d'HTA rénovasculaire les plus élevées

PROBA HAP	PROBA RENOVASC	PROBA PPGL	PROBA HTAE	GROUPE
9.96	29.4	2.21	67.0	heredite_0 / imc_1_0 / mdrd_0 / ttt_hta_bb_1
6.38	28.0	2.03	71.8	imc_1_0 / mdrd_0 / pad_1_0 / ttt_hta_bb_1
NA	27.9	NA	73.4	age_3_1 / imc_1_0 / mdrd_0 / ttt_hta_bb_1
9.87	27.7	3.10	68.3	heredite_0 / imc_1_0 / ttt_hta_bb_1 / ttt_stati…
NA	27.5	NA	69.3	age_3_1 / imc_1_0 / mdrd_0 / ttt_hta_ic_1
11.3	27.1	1.30	66.8	imc_1_0 / mdrd_0 / nb_ttt_2_1 / taille_2_0
NA	26.8	NA	71.7	age_3_1 / imc_1_0 / pas_3_1 / ttt_statine_1
9.22	26.2	1.80	69.4	atcd_saos_0 / imc_1_0 / mdrd_0 / ttt_hta_bb_1
17.3	26.1	0.808	59.9	imc_1_0 / mdrd_0 / ttt_hta_central_1
12.1	26.0	1.36	66.1	imc_1_0 / mdrd_0 / nb_ttt_2_1

Tableau 7 - groupes avec les probabilités de PPGL les plus élevées

PROBA HAP	PROBA RENOVASC	PROBA PPGL	PROBA HTAE	GROUPE
8.97	5.20	9.16	75.4	sueurs_1
12.2	8.58	9.09	75.5	age_2_0 / diabete_1 / pad_1_0 / ttt_hta_bsra_0
7.45	14.6	6.90	74.2	diabete_1 / imc_1_0 / pad_1_0 / sexe_1
22.8	6.32	6.78	71.1	age_2_0 / diabete_1 / sexe_1 / ttt_hta_bsra_0
22.7	6.99	6.17	66.8	nb_ttt_2_0 / ttt_hta_alphab_1
5.47	19.3	6.12	75.9	diabete_1 / imc_1_0 / ka_2_1 / ttt_statine_1
10.7	5.50	5.79	77.4	palpitations_1
5.38	6.85	5.70	82.5	atcd_intolerance_ttt_hta_0 / cephalees_1 / imc_…
11.4	7.39	5.66	77.5	age_2_0 / cephalees_1 / heredite_0 / ttt_hta_va…
11.5	7.51	5.63	77.4	age_2_0 / cephalees_1 / heredite_0

Résultats complémentaires

Apprentissage automatique

Non supervisé

Une partie de ce travail a consisté à développer des algorithmes d'apprentissage automatique non supervisés, afin de déterminer s'il était possible de faire émerger des caractéristiques dans certains groupes de patients pouvant faire évoquer un diagnostic d'hypertension artérielle secondaire.

Analyses en composante principale (ACP)

Ci-après les représentations des données selon les deux premières dimensions de l'ACP et les contributions des variables à ces dimensions.

On constate qu'aucun cluster indépendant n'émerge.

Figure 31 - représentation des 2 principales dimensions de l'analyse en composantes principales pour les variables quantitatives, avec visualisation du diagnostic d'HTA primaire

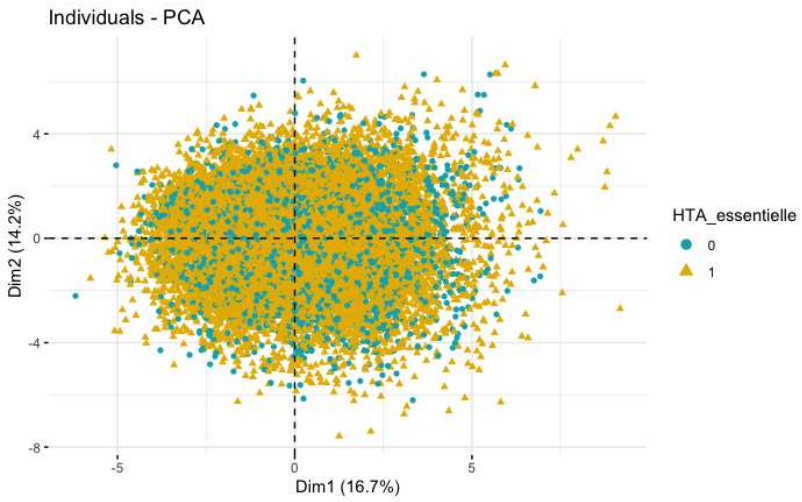


Figure 32 - représentation des 2 principales dimensions de l'analyse en composantes principales pour les variables quantitatives, avec visualisation du diagnostic d'HAP

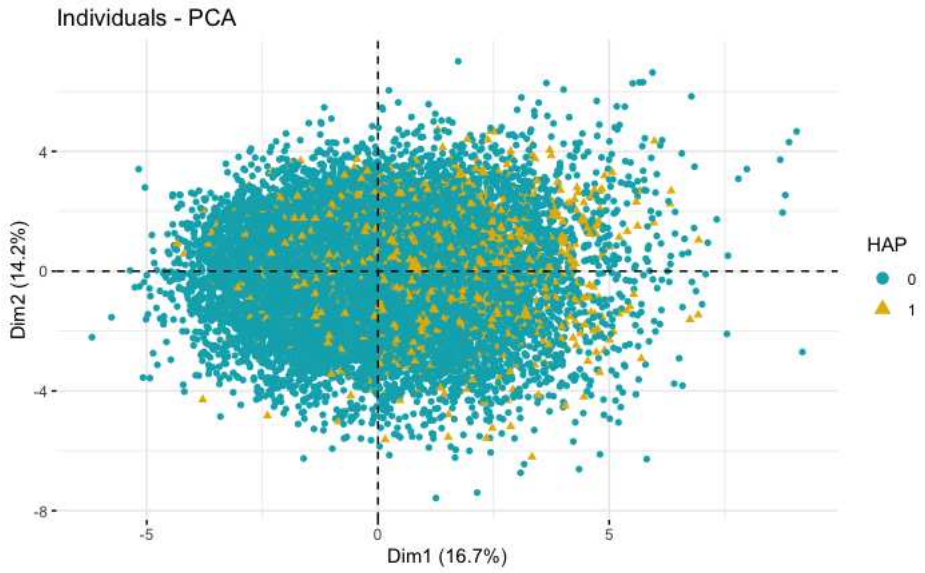
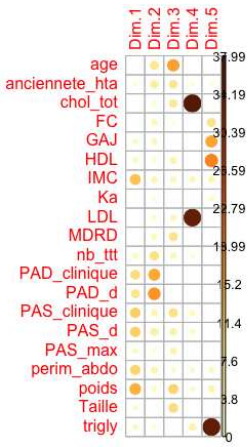


Figure 33 - contribution des variables dans les 5 premières dimensions de l'ACP



Analyse des composantes multiples (ACM)

Ci-après les représentations des données selon les deux premières dimensions de l'ACM et les contributions des variables à ces dimensions.

On constate qu'aucun cluster indépendant n'émerge.

Figure 34 - représentation des 2 principales dimensions de l'analyse des composantes multiples pour les variables catégorielles, avec visualisation du diagnostic d'HTA primaire

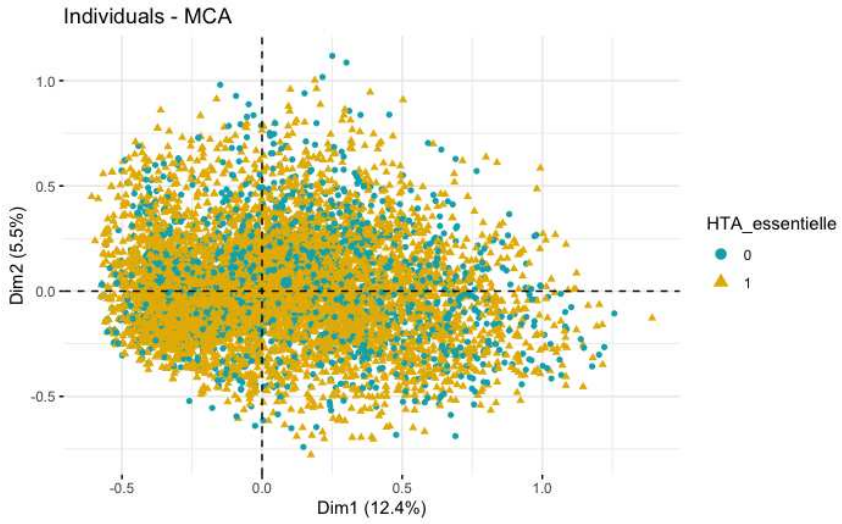


Figure 35 - représentation des 2 principales dimensions de l'analyse des composantes multiples pour les variables catégorielles, avec visualisation du diagnostic d'HAP

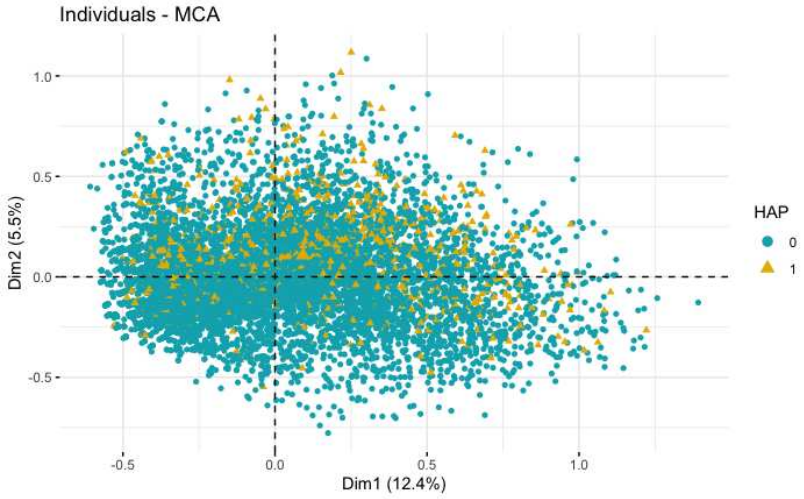
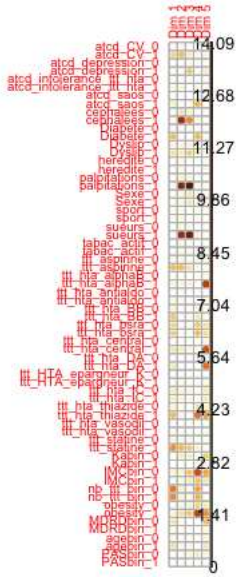


Figure 36 - contribution des variables dans les 5 première dimensions de l'ACM



K-means

Nous avons ensuite cherché à faire émerger des clusters d'intérêt par le modèle des k-means.

Nous avons étudié le modèle pour plusieurs nombres de clusters.

Nous avons pour cela regardé différents nombres de clusters (Figure 37) jusqu'à 8, la distance euclidienne entre les clusters diminuant de manière importante passé ce nombre.

La table 8 montre les valeurs moyennes de chaque variable pour le modèle à 3 clusters.

La figure 38 montre la répartition des clusters selon 3 variables choisies car ayant une valeur moyenne variant significativement dans un modèle à 3 clusters.

A nouveau, on constate l'absence de démarcation significative entre les différents clusters.

Tableau 8 - valeurs moyennes des variables par cluster dans un modèle à 3 clusters

variables	*cluster*	*value*
age	1	59,40308
age	2	47,77827
age	3	55,54702
anciennete_hta	1	13,41426
anciennete_hta	2	5,784124
anciennete_hta	3	11,57836
chol_tot	1	5,130037
chol_tot	2	5,137616
chol_tot	3	5,26
FC	1	69,58388
FC	2	74,96719
FC	3	75,65622
GAJ	1	5,875163
GAJ	2	5,397774
GAJ	3	5,909606
HDL	1	1,404776
HDL	2	1,455138
HDL	3	1,403862
IMC	1	28,45606
IMC	2	26,55748
IMC	3	28,73672
Ka	1	4,193663
Ka	2	4,159695
Ka	3	4,111678
LDL	1	3,078141
LDL	2	3,123204
LDL	3	3,213301
MDRD	1	69,28751
MDRD	2	84,92999
MDRD	3	73,55207

nb_ttt	1	2,650047
nb_ttt	2	1,219098
nb_ttt	3	2,252781
PAD_clinique	1	76,18593
PAD_clinique	2	82,69713
PAD_clinique	3	96,14611
PAD_d	1	78,88071
PAD_d	2	87,68776
PAD_d	3	101,4353
PAS_clinique	1	131,7102
PAS_clinique	2	133,063
PAS_clinique	3	164,632
PAS_d	1	127,5993
PAS_d	2	132,0419
PAS_d	3	164,7144
PAS_max	1	213,7954
PAS_max	2	172,2759
PAS_max	3	205,1416
perim_abdo	1	98,48342
perim_abdo	2	92,3708
perim_abdo	3	99,2815
poids	1	80,38733
poids	2	76,64313
poids	3	82,8761
size	1	2146
size	2	3414
size	3	1978
Taille	1	1,677223
Taille	2	1,696116
Taille	3	1,695364
trigly	1	1,423821
trigly	2	1,230463
trigly	3	1,414505

Figure 37 - représentation à l'échelle des différents clusters selon le nombre choisi, selon la kaliémie et la pression artérielle systolique clinique

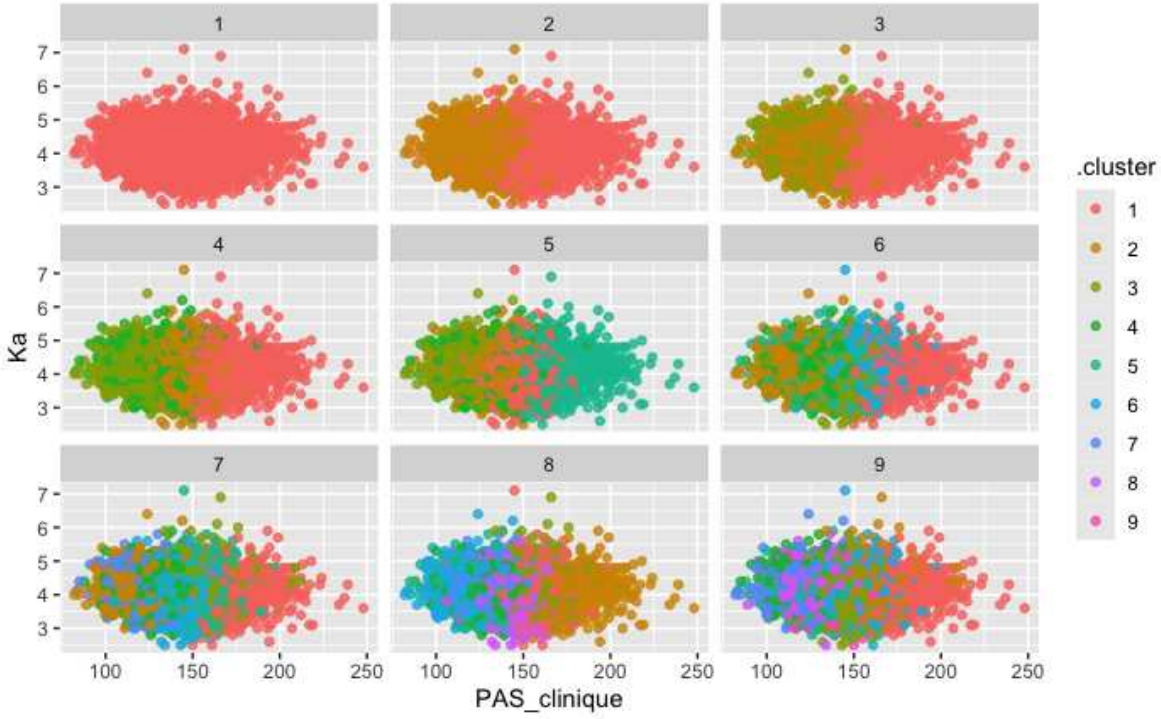
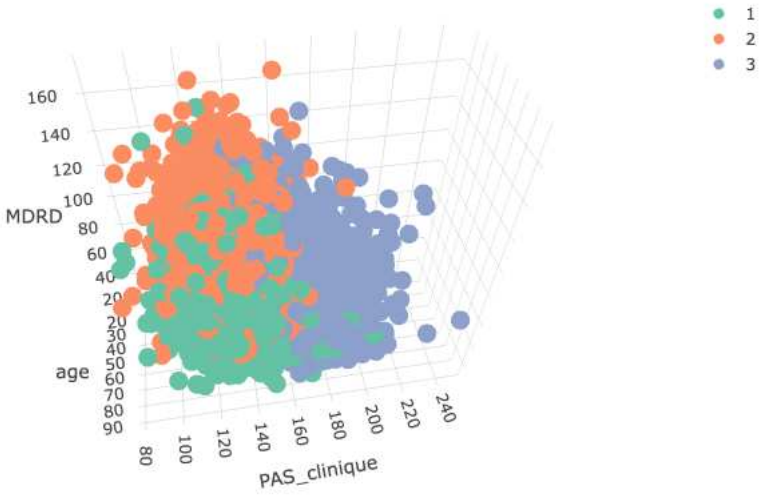


Figure 38 - répartition des patients selon 3 variables dans un modèle à 3 clusters



UMAP

Nous avons enfin utilisé le modèle UMAP pour visualiser en deux dimensions la présence ou non de clusters d'intérêts au sein de notre population.

Comme visualisé sur les figures 39 et 40, aucun cluster ne se forme.

Figure 39 - Représentation d'un clustering UMAP selon le diagnostic d'HTA primaire

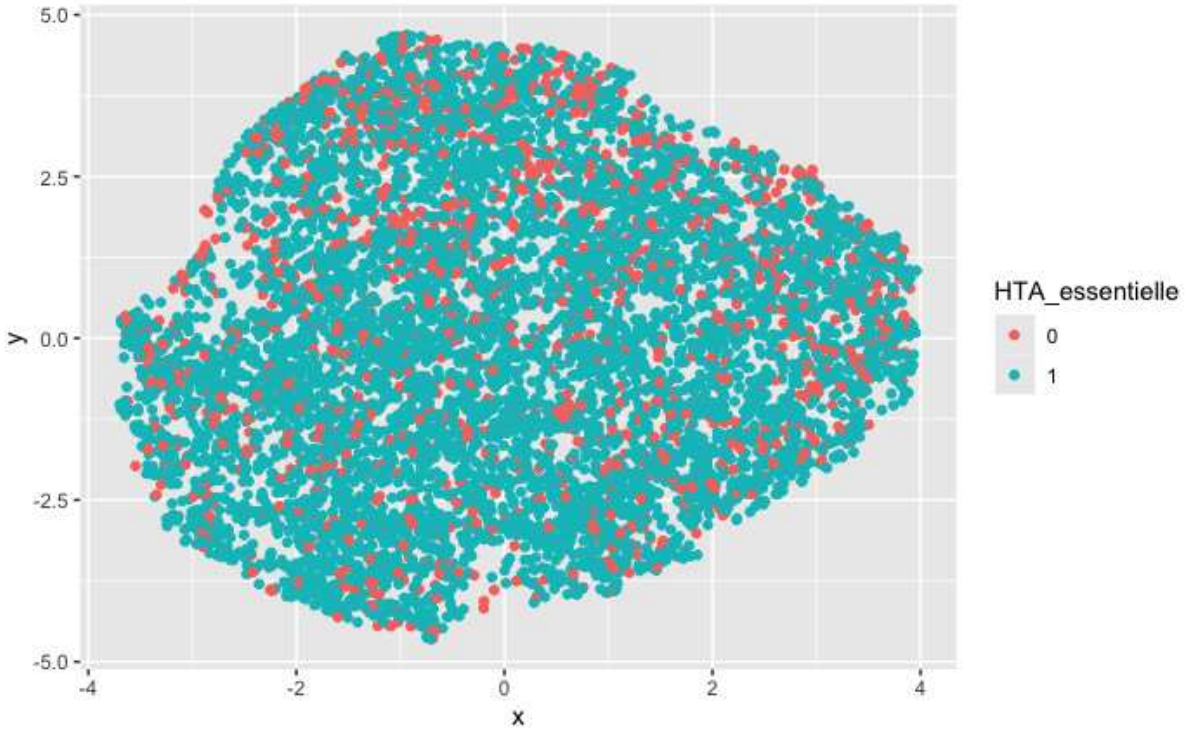
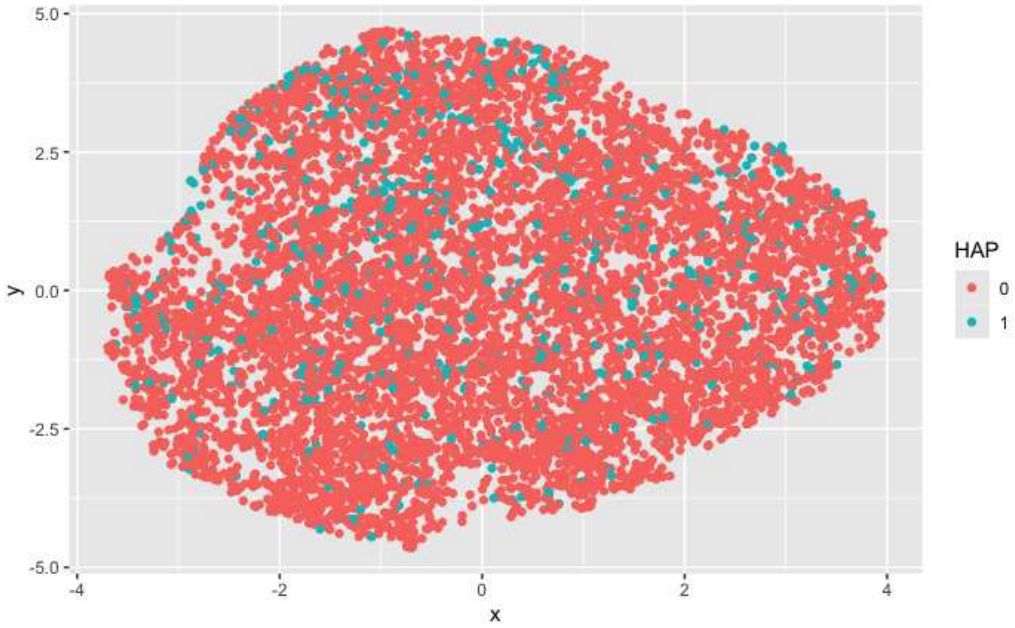


Figure 40 - Représentation d'un clustering UMAP selon le diagnostic d'Hyperaldostéronisme primaire



Supervisé

Nous avons réalisé les analyses supervisées présentées dans l'article (XGBoost et régression logistique) pour les autres diagnostics étudiés : HTA rénovasculaire, phéochromocytomes et paragangliomes.

HTA rénovasculaire

Algorithme de stratification

L'AUC de la courbe ROC était de 0.69 (Figure 41).

Pour une sensibilité d'environ 90%, le seuil choisi était une probabilité égale à 0,09.

Les résultats sont présentés dans la table 9.

Figure 41 - courbe ROC pour l'algorithme de stratification pour l'HTA rénovasculaire

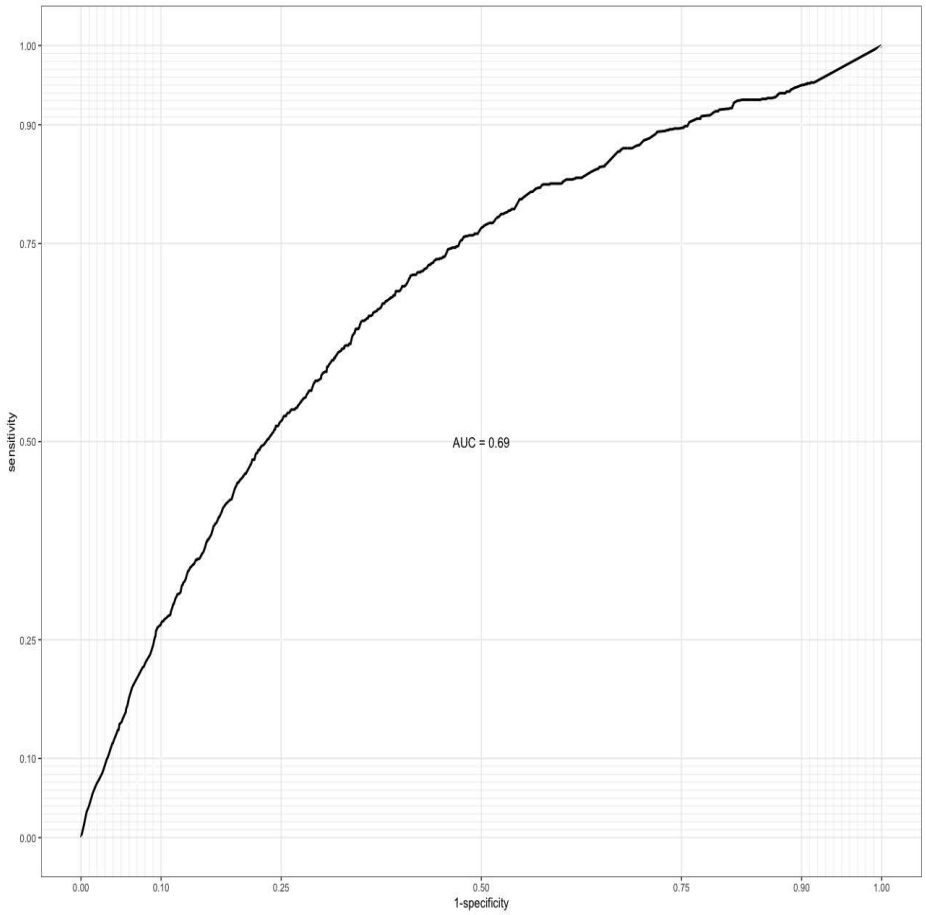


Tableau 9 - matrice de confusion pour l'algorithme de stratification pour l'HTA rénovasculaire

	derivation Cohort (N = 15,507)				Validation cohort (N = 768)		
Stratification algorithm			Predicted Renovascular HTN			Predicted Renovascular HTN	
			No	Yes		No	Yes
	True Renovascular	No	4598	125		168	3
		Yes	9948	839		561	15
Sensitivity		0.87			0.98		
Specificity		0.32			0.26		

Accuracy	0.35	0.25
NPV	0.97	0.83
AUC	0.69	

Régression logistique

Nous avons réalisé une régression de type elastic net.

Nous avons utilisé la fonction R grid search, et une cross-validation à 10 blocs sur le jeu d'entrainement afin de déterminer les meilleurs paramètres.

La regression LASSO pure fournissait les meilleurs résultats (Least Absolute Shrinkage and Selection Operator).

Les paramètres retenus sont détaillés dans la table 10.

Les résultats sont présentés dans la table 11.

Tableau 10 - paramètres des modèles utilisés

		HAP	RENOVASC	PPGL
STRATIFICATION	Seuil proba	0.29	0.09	0.02
REGRESSION	pénalité	0.01	0.01	0.01
XGBOOST	nombre d'arbres	2000	1000	1000
	profondeur des	8	10	8

Figure 42 - courbe ROC pour la régression logistique pour l'HTA rénovasculaire

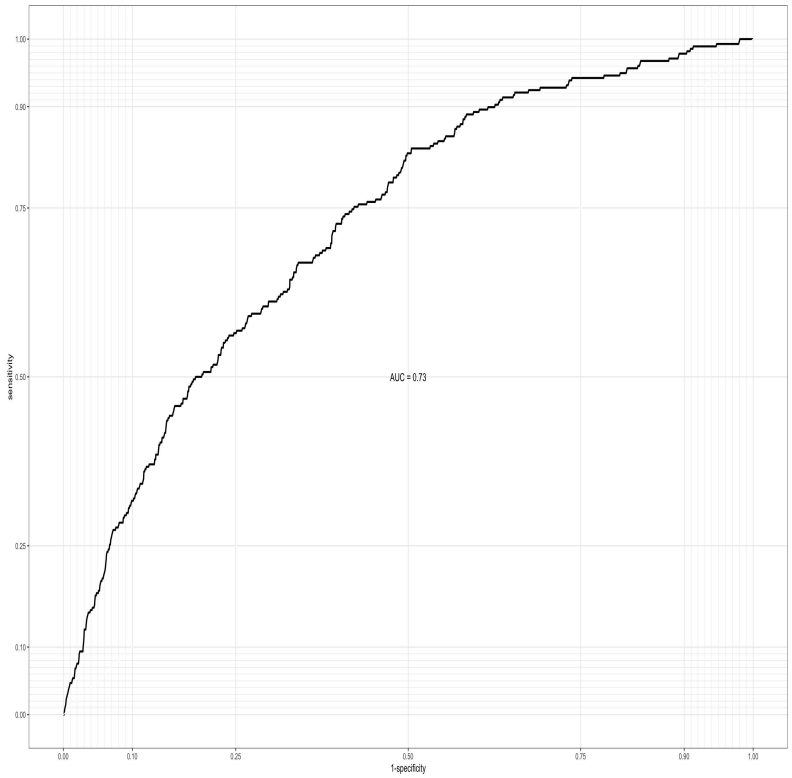
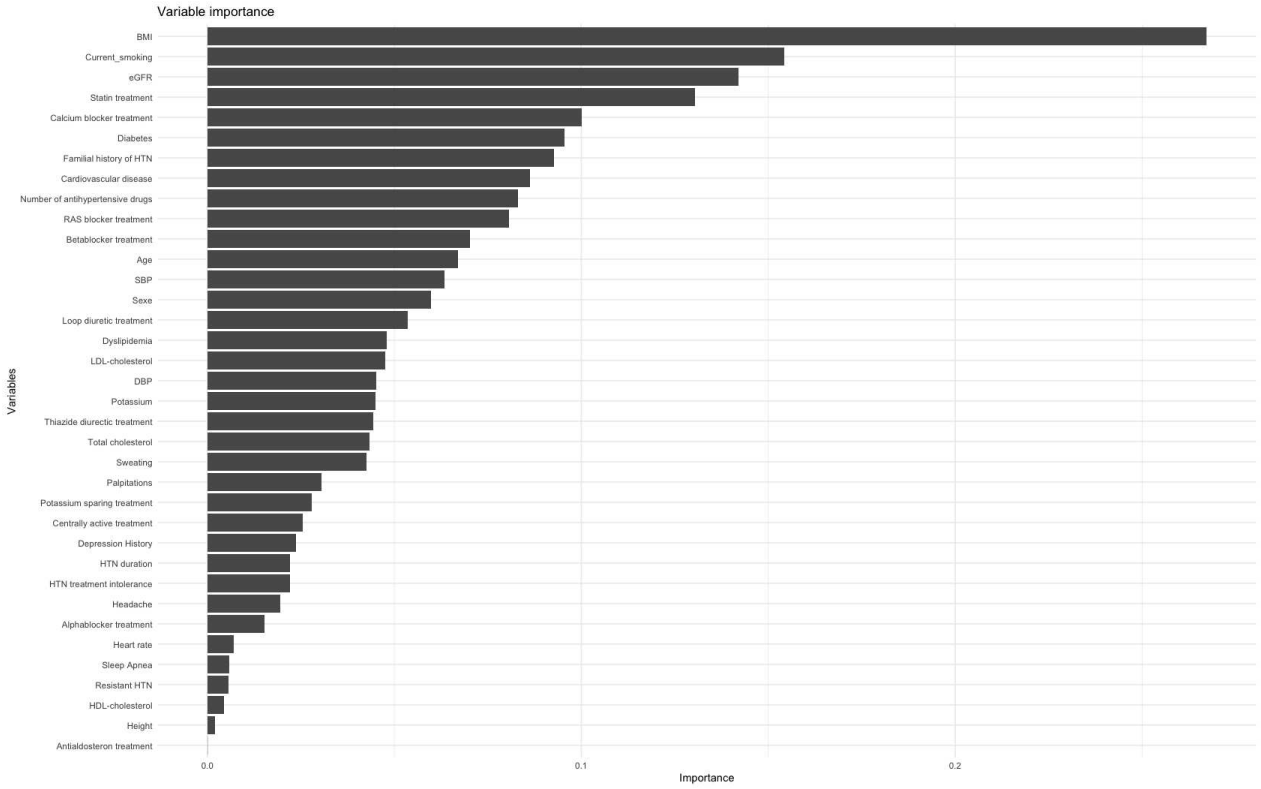


Tableau 11 - matrice de confusion pour la régression logistique pour l'HTA rénovasculaire

	testing Cohort (N = 4653)				Validation cohort (N = 768)		
Logistic regression			Predicted Renovascular HTN			Predicted Renovascular HTN	
			No	Yes		No	Yes
	True	No	1688	29		738	17
	Renovascular	Yes	2687	249		12	1
Sensitivity		0.90			0.56		
Specificity		0.39			0.98		
Accuracy		0.41			0.96		
NPV		0.98			0.98		
AUC		0.73					

Figure 43 - importance des différentes variables dans le modèle de régression logistique



XGboost

Nous avons utilisé la fonction R grid search, et une cross-validation à 10 blocs sur le jeu d'entrainement afin de déterminer les meilleurs paramètres.

Les paramètres étudiés étaient le nombre d'arbres dans l'ensemble et la taille maximale de chaque arbre.

Les paramètres retenus sont détaillés dans la table 10.

Les résultats sont présentés dans la table 12.

Figure 44 - courbe ROC pour XGBoost pour l'HTA rénovasculaire

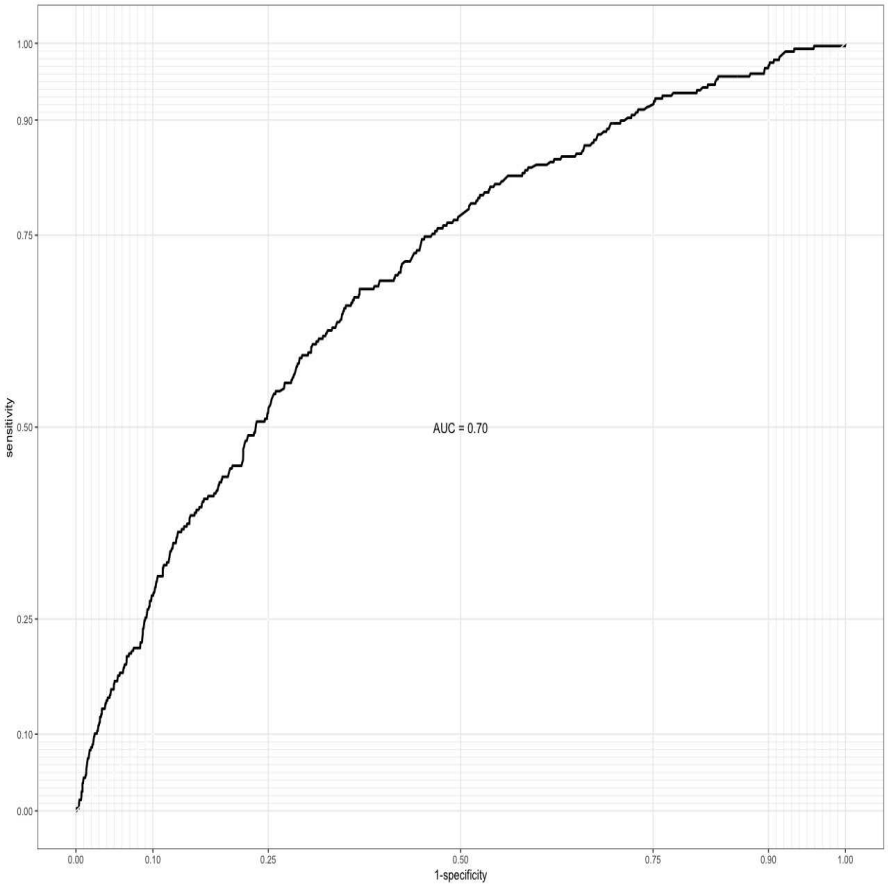
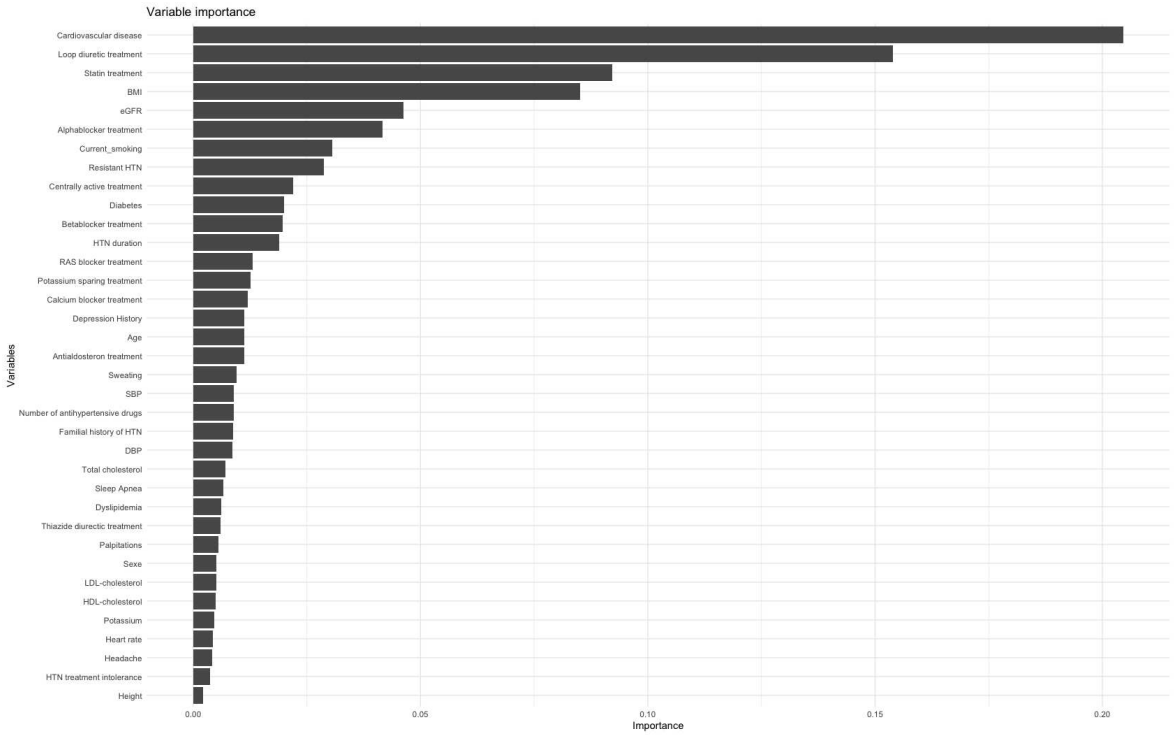


Tableau 12 - matrice de confusion pour XGBoost pour l'HTA rénovasculaire

	testing Cohort (N = 4653)				Validation cohort (N = 768)		
XGBoost			Predicted Renovascular HTN			Predicted Renovascular HTN	
			No	Yes		No	Yes
	True	No	1216	26		138	1
	Renovascular	Yes	3159	252		612	17
Sensitivity		0.91			0.94		
Specificity		0.28			0.18		
Accuracy		0.32			0.20		
NPV		0.98			0.99		
AUC		0.70					

Figure 45 - importance des différentes variables dans le modèle XGBoost



Il n'a donc pas été possible de réaliser une validation externe de ces résultats sur la cohorte bordelaise car elle ne comprenait pas patients atteints de phéochromocytomes ou paragangliomes.

Algorithme de stratification

L'AUC de la courbe ROC était de 0.72 (Figure 45).

Pour une sensibilité de 90%, le seuil choisi était une probabilité égale à 0,02.

Les résultats sont présentés dans la table 14.

Figure 46 - courbe ROC pour l'algorithme de stratification pour les PPGL

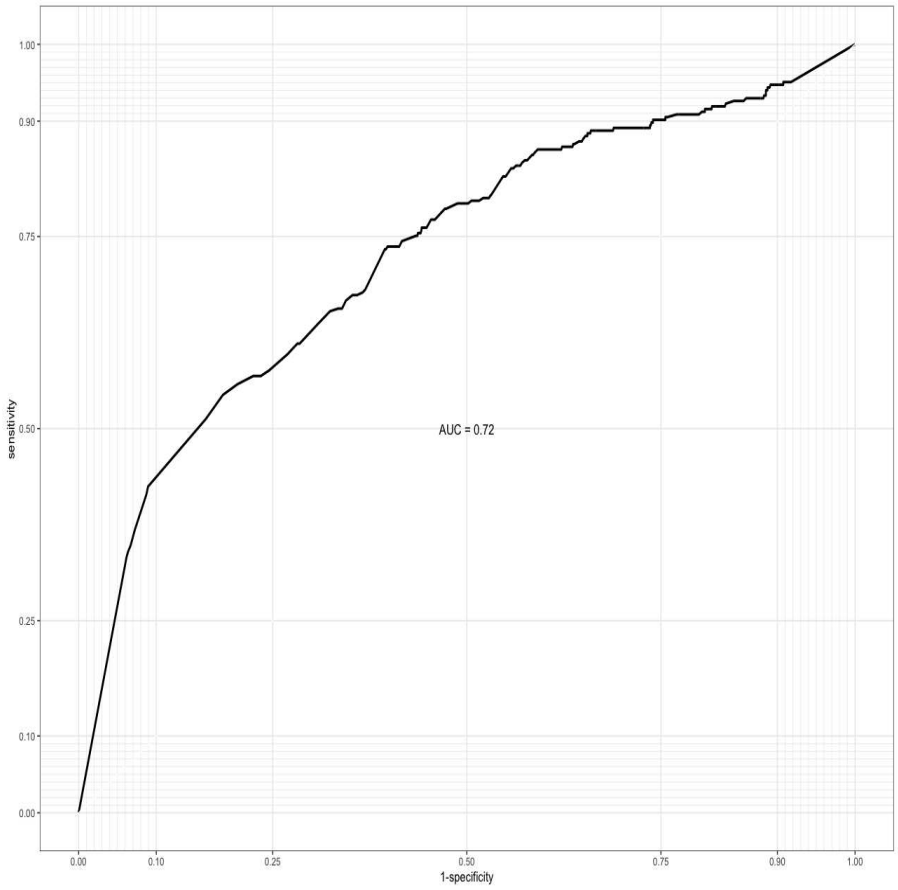


Tableau 13 - matrice de confusion pour l'algorithme de stratification pour les PPGL

	derivation Cohort (N = 15,507)			
Stratification algorithm			Predicted PPGL	
			No	Yes
	True PPGL	No	3692	17
		Yes	11535	258
Sensitivity		0.91		
Specificity		0.24		
Accuracy		0.26		
NPV		0.99		
AUC		0.72		

Régression logistique

Nous avons réalisé une régression de type elastic net.

Nous avons utilisé la fonction R grid search, et une cross-validation à 10 blocs sur le jeu d'entrainement afin de déterminer les meilleurs paramètres.

La regression LASSO pure fournissait les meilleurs résultats (Least Absolute Shrinkage and Selection Operator).

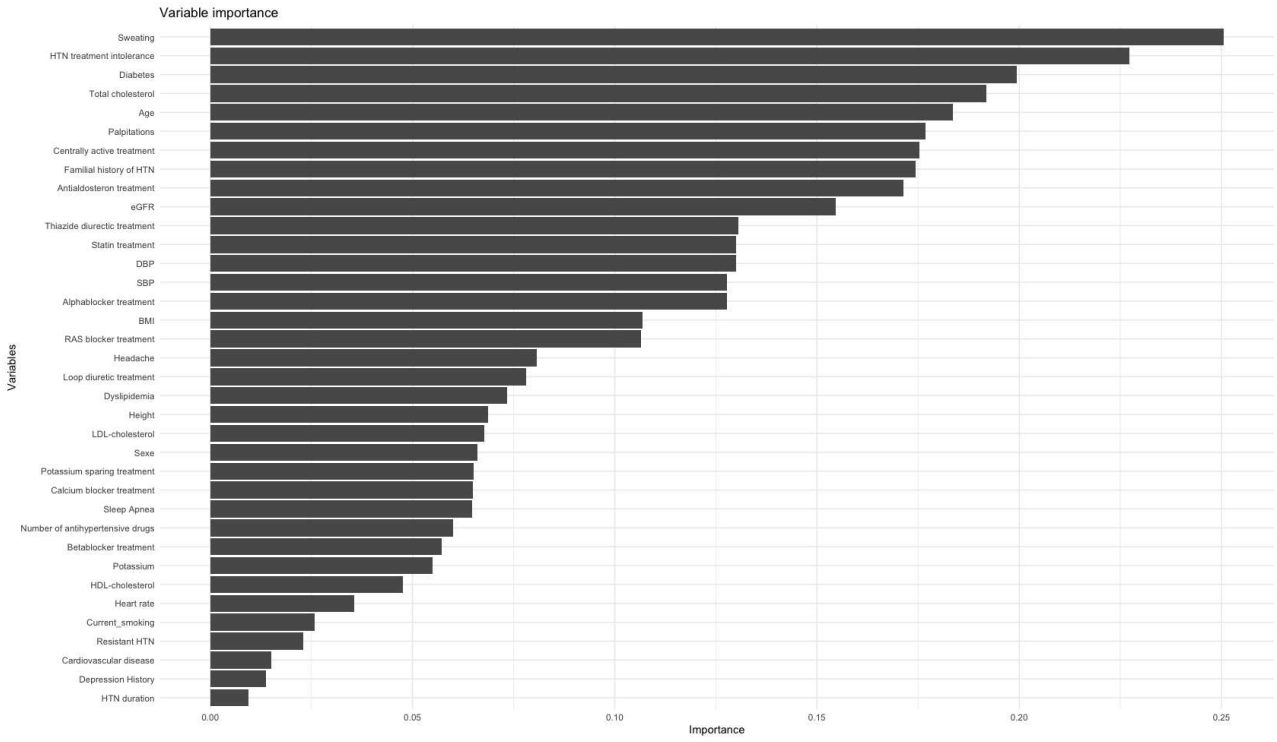
Les paramètres retenus sont détaillés dans la table 10.

Les résultats sont présentés dans la table 16.

Tableau 14 - matrice de confusion pour la régression logistique pour les PPGL

testing Cohort (N = 4653)				
Logistic regression			Predicted Renovascular HTN	
			No	Yes
	True Renovascular HTN	No	1465	8
		Yes	3111	69
Sensitivity		0.90		
Specificity		0.32		
Accuracy		0.33		
NPV		0.99		
AUC		0.80		

Figure 47 - importance des différentes variables dans le modèle de régression logistique



XGboost

Nous avons utilisé la fonction R grid search, et une cross-validation à 10 blocs sur le jeu d'entrainement afin de déterminer les meilleurs paramètres.

Les paramètres étudiés étaient le nombre d'arbres dans l'ensemble et la taille maximale de chaque arbre.

Les paramètres retenus sont détaillés dans la table 10.

Les résultats sont présentés dans la table 15.

Figure 48 - courbe ROC pour XGBoost pour les PPGL

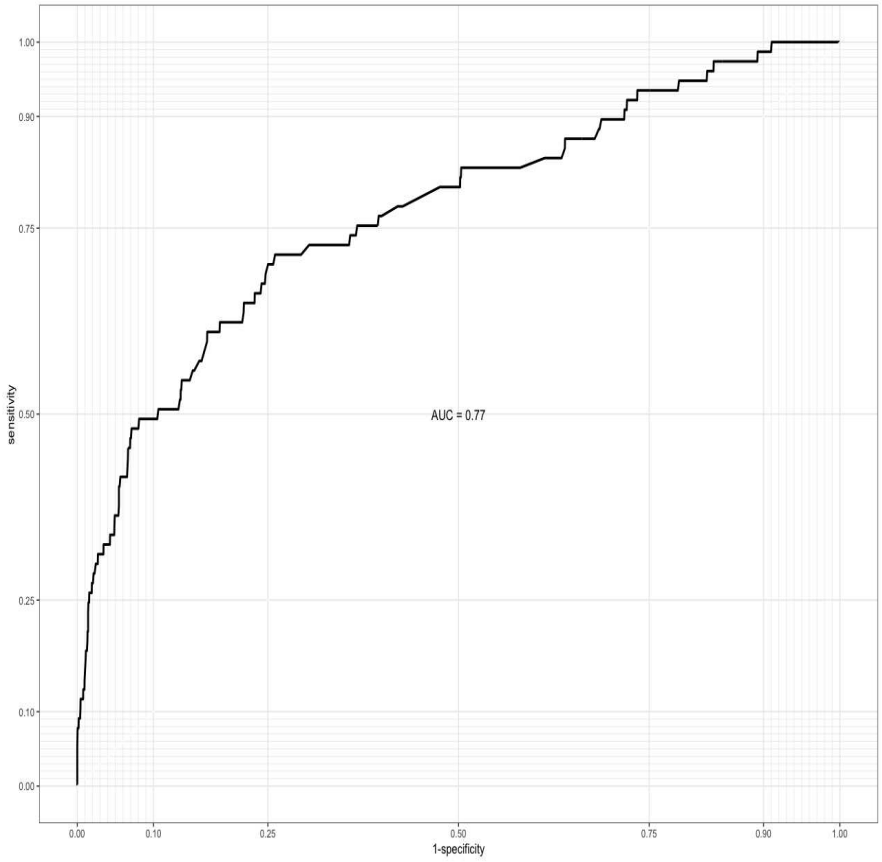
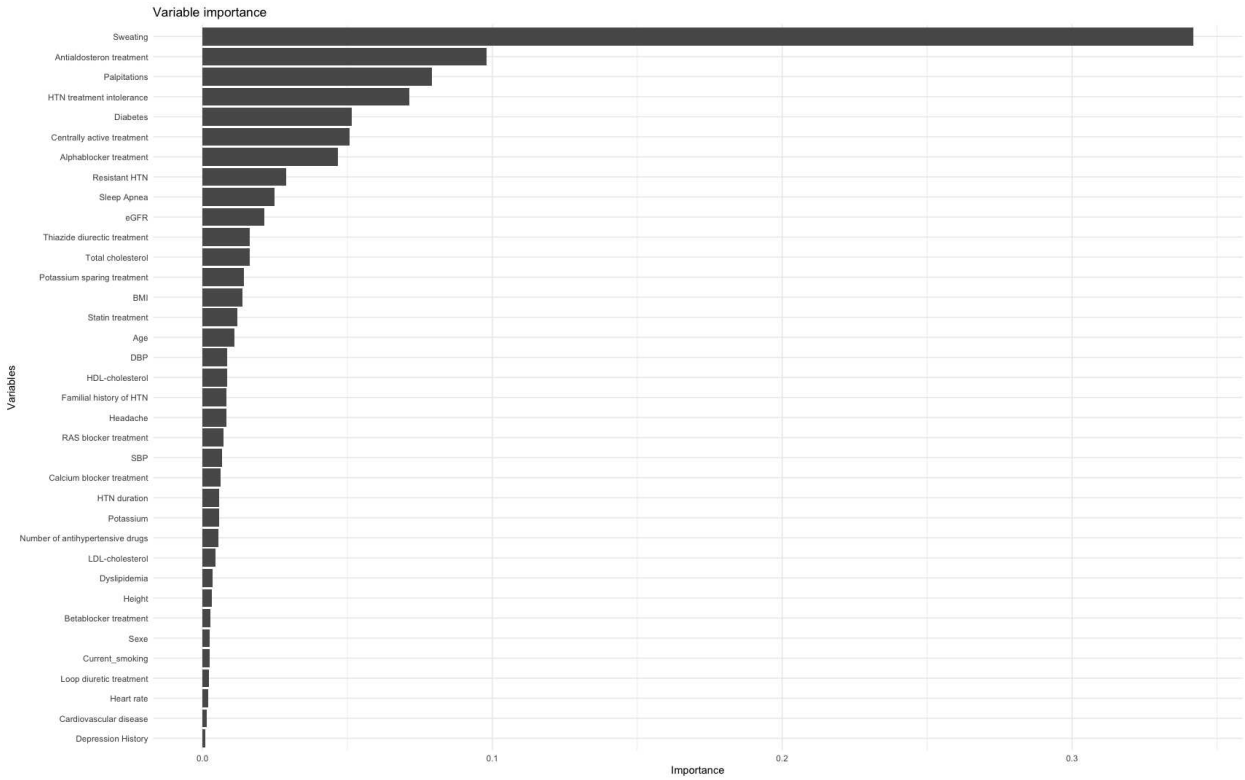


Tableau 15 - matrice de confusion pour XGBoost pour les PPGL

testing Cohort (N = 4653)			
XGBoost		Predicted PPGL	
		No	Yes
	True Renovascular HTN No	1285	7
	Yes	3291	70
Sensitivity	0.91		
Specificity	0.28		
Accuracy	0.29		
NPV	0.99		
AUC	0.77		

Figure 49 - importance des différentes variables dans le modèle XGBoost



Apprentissage profond

Nous avons également développé un réseau de neurones classique, mais les résultats étaient insuffisants pour envisager une application et nous ne les présenterons pas dans ce manuscrit.

Application et interface Web

Nous avons développé une application permettant l'utilisation de l'algorithme de stratification, avec pour projet de la rendre disponible une fois validée à tout praticien via une interface web.

Concrètement, les données du patient peuvent être entrées sur l'interface, codée en R Shiny, et les groupes d'intérêt auquel le patient appartient apparaissent, avec les probabilités calculées de chaque diagnostic.

Dans l'exemple présenté sur la Figure 50, un patient ayant pour caractéristiques :

- Homme
- 35 ans
- Histoire familiale d'HTA
- Sueurs
- 4 médicaments antihypertenseurs
- Kaliémie 3,4 mmol/L

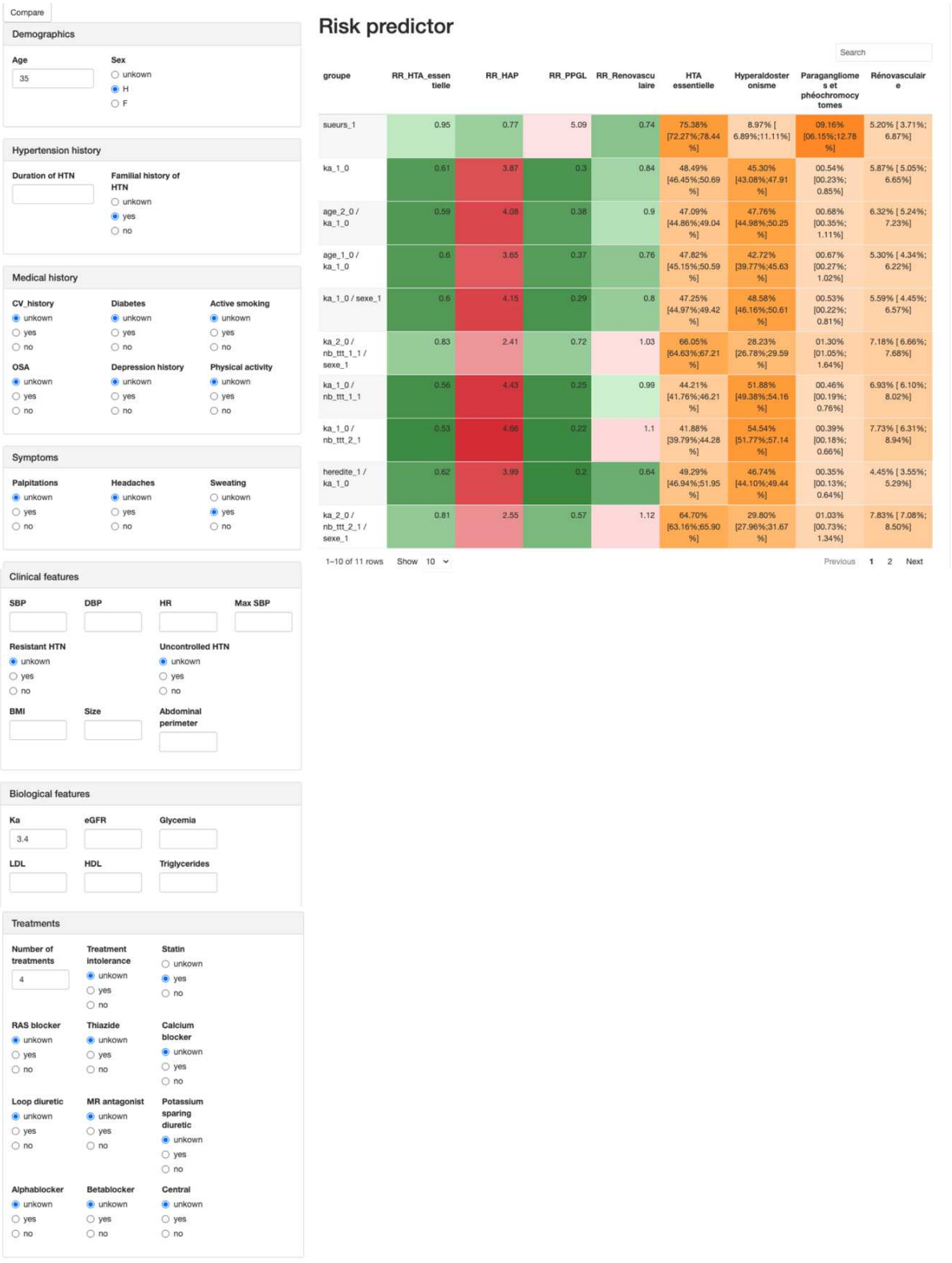
est associé à plusieurs groupes, avec différents niveaux de probabilité (11 groupes dont 10 sont affichés sur la figure).

Les variables sont expliquées dans les tables 3 et 4 du chapitre 3, et le chiffre apparaissant à la suite correspond à la valeur de la variable (1 pour positif ou supérieur au seuil, 0 pour négatif ou inférieur au seuil). Par exemple, *ka_1_0* correspond à un potassium inférieur à 3,5 mmol/L, et heredite_1 correspond à une histoire familiale d'HTA.

On peut noter que la plupart des groupes sont associés à un risque élevé d'HAP, notamment ceux comprenant l'hypokaliémie (ka_1_0), et qu'un groupe est associé à un risque élevé de PPGL, celui comprenant la variable « sueurs ».

Dans un second temps, nous souhaitons proposer une unique recommandation pour chaque patient, avec la possibilité d'accéder aux groupes d'intérêt afin de comprendre la recommandation de l'algorithme.

Figure 50 - exemple d'utilisation de l'application; à gauche les caractéristiques patient, à droite les probabilités diagnostiques (RR et probabilités absolues)



Complément de discussion

Nous avons développé un algorithme de stratification du risque d'HTA secondaire, ayant pour objectif de pouvoir être utilisé en population générale, par des médecins non spécialistes de l'hypertension artérielle.

L'un des points de réflexion majeurs dans cet objectif concerne les potentiels freins à l'adoption de l'algorithme par les médecins, qui dépend fortement de leur confiance en l'algorithme. Les facteurs influençant cette confiance sont la fiabilité, la robustesse, mais aussi la transparence et l'interprétabilité[88]. Les cliniciens peuvent avoir des difficultés à comprendre et à faire confiance à des prédictions issues de « boites noires ».

Différentes techniques d'interprétabilité des modèles existent, comme la « *Shap-value* » (*Shapley Additive exPlanations*), qui évalue à quel point chaque variable contribue à orienter la prédiction vers sa valeur finale. Ou les modèles locaux de substitution, ajustant un modèle interprétable sur des données perturbées dans le voisinage de l'instance à expliquer[72]. Par ailleurs, l'opacité d'un algorithme, c'est-à-dire l'absence de transparence de son processus d'élaboration d'une prédiction, complique la relation entre le médecin et le patient, en remettant en question son autorité épistémique : il ne peut plus justifier ses recommandations, qui sont basées sur des algorithmes dont il ne peut expliquer le fonctionnement[73]. Cependant les méthodes actuelles pour rendre les modèles interprétables ou explicables sont insuffisantes, et peuvent même conduire à des sur-interprétations[71]. Le plus important pour qu'un modèle puisse être utilisé avec confiance est finalement la robustesse du modèle, ce qui nécessite une validation rigoureuse, non seulement sur des bases de données externes, mais également de manière prospective dans des essais cliniques.

Certains modèles de classification ont été développés avec pour objectif d'être interprétables. C'est le cas de le FEAT (*Feature Engineering Automation Tool*), basé sur la génération de régressions symboliques (en pratique des combinaisons linéaires de critères basés sur les variables disponibles), et entrainé de sorte à maximiser la discrimination, et minimiser la complexité du modèle. Les modèles générés sont donc suffisamment simples pour être interprétables[127]. Par exemple, la combinaison linéaire de la pression artérielle moyenne sur différentes visites, de la pression artérielle systolique moyenne, du nombre médian, total de visite comprenant un certain nombre de traitements antihypertenseurs, et de leur écart-type, du nombre de mention de l'HTA dans les comptes-rendus, et de la calcémie,

est utilisée pour prédire l'HTA résistante, si cette combinaison est supérieure à un certain seuil.

Notre algorithme a obtenu des performances intéressantes en termes de valeur prédictive négative pour le dépistage des HAP puisque seulement 6 patients sur 768 dans la cohorte de validation avaient un HAP malgré un dépistage négatif par l'algorithme.

Une des principales limites est le biais de sélection initial de la population, qui est celle d'un centre expert, et donc dans laquelle la prévalence d'HTA secondaire est plus importante que dans la population générale, sur laquelle nous souhaitons appliquer notre algorithme.

En population générale, la prévalence plus faible devrait majorer encore la valeur prédictive négative de l'algorithme, augmentant ainsi la sécurité avec laquelle il peut être utilisé. Un faible taux de faux-négatifs peut avoir un effet bénéfique sur la confiance accordée par le clinicien, pour qui le risque la plus important est de passer à côté d'une pathologie potentiellement grave.

Concernant la prédiction des PPGL et HTA rénovasculaires, les performances étaient nettement moindres, et ne semblent pas à l'heure actuelle apporter de réelle plus-value au dépistage de ces pathologies. On a, en revanche, pu constater que les groupes associés aux plus fortes probabilités de PPGL comprenaient souvent des éléments cliniques (céphalées, sueurs ou palpitations), ce qui est une validation a posteriori du modèle et de son interprétabilité, de la même manière que la présence fréquente de l'hypokaliémie dans le diagnostic d'HAP et que les facteurs de risque cardiovasculaires pour le diagnostic d'HTA rénovasculaire.

Concernant les modèles d'apprentissage automatique, les performances sur la base d'entrainement sont intéressantes et du même niveau que celle de l'algorithme de stratification, mais avec des valeurs prédictives négatives moins fortes, et respectivement 49 et 84 patients sur 768 ayant un diagnostic d'HAP malgré une prédiction négative. Nous avons voulu augmenter l'interprétabilité des modèles en représentant le poids des variables dans chaque algorithme. On retrouvait pour l'HAP, l'hypokaliémie, les traitements épargneurs de potassium et anti-aldostérone, et la prescription d'un traitement neutre. En effet, certains patients avaient déjà été mis sous traitement neutre avant leur première visite à l'HEGP, dans l'optique de réaliser un bilan d'HTA secondaire et notamment de rechercher un HAP, ce qui entraine un biais de confirmation. La présence d'un traitement neutre, que l'on retrouve également dans les groupes de l'algorithmes de stratification pose la question

de l'intégration du traitement dans les variables de l'algorithme. Cependant, nous avons préféré considérer notre algorithme comme pragmatique, et nous l'avons élaboré à partir de données de vie réelle. Cela confirme également la bonne logique de création des groupes.

De la même manière que pour l'algorithme de stratification, les variables associées à la régression logistique et à XGBoost étaient principalement les antécédents vasculaires ou les facteurs de risque cardiovasculaires pour l'HTA rénovasculaire, dont la principale cause est la maladie athéromateuse, et les éléments de la triade de Ménard et le diabète pour les PPGL.

En conclusion, nous avons développé un algorithme dont les performances sont prometteuses pour le dépistage de l'HAP, la cause la plus fréquente d'HTA secondaire mais qui est sous-diagnostiquée. Nous avons également développé de manière exploratoire des algorithmes pour le dépistage de l'HTA rénovasculaire et des PPGL, avec des performances moindres, mais validant le principe de notre algorithme de stratification. L'étape suivante devrait être la validation de cet algorithme de manière prospective, sur la population cible qu'est la population générale de patients hypertendus.

Conclusion et perspectives générales

L'objectif de ce travail de thèse était 1) d'étudier les caractéristiques de l'HTA secondaire, notamment dans une population mal décrite qu'est la population des patients jeunes, chez qui la recherche d'HTA secondaire est une priorité et peut avoir des conséquences à court et long terme, et 2) d'élaborer un modèle prédictif permettant d'améliorer l'efficience des stratégies diagnostiques d'HTA secondaires.

Nous avons dans un premier temps confirmé la prévalence élevée de l'HTA secondaire chez les patients jeunes. Bien que nous ayons déterminé des facteurs prédictifs d'HTA secondaire, notre modèle statistique de régression logistique avec *bootstrap* ne parait pas directement utilisable pour modifier la stratégie diagnostique de l'HTA secondaire dans cette population. En effet, dans cette population jeune, et compte tenu de l'impact majeur de la plupart des causes d'HTA secondaires, il est essentiel de maximiser la valeur prédictive négative de nos modèles prédictifs et de proposer une stratégie minimisant les faux négatifs. Dans ce cas, les sous-groupes de patients avec la probabilité la plus faible affichait malgré tout 10% de risque d'avoir une HTA secondaire, risque inacceptable.

Nous avons ensuite développé un modèle sur la population des patients de moins de 65 ans ayant consulté à l'HEGP (afin de maximiser le nombre de patients ayant eu un bilan d'HTA secondaire complet), avec pour objectif qu'il puisse être appliqué en population générale, chez tout patient hypertendu. Nous avons cette fois développé plusieurs modèles, un par diagnostic d'HTA secondaire, car les enjeux liés à chaque pathologie et leurs risques à court et long terme sont différents.

Le modèle le plus intéressant que nous avons développé concerne le diagnostic de l'HAP. Nous avons particulièrement travaillé ce modèle du fait que c'est la cause la plus fréquente d'HTA secondaire, qu'elle a un impact clinique à long terme important, et qu'elle est très certainement sous-diagnostiquée. Notre modèle devait répondre à la même problématique de maximisation de la valeur prédictive négative et de minimisation des faux positifs.

Un des intérêts principaux de notre modèle est la transparence du processus de classification, c'est-à-dire la possibilité de visualiser et comprendre la raison pour laquelle l'algorithme rend une recommandation de dépistage ou non. En effet, et comme discuté précédemment,

l'interprétabilité d'un modèle a une part importante dans la confiance que vont lui accorder les cliniciens qui vont l'utiliser[73,88].

Nous nous sommes également focalisés sur une étape de prise en charge, qui est la première consultation, avec des données facilement accessibles. Le fait de se concentrer sur une partie de la prise en charge diagnostique et d'intégrer l'algorithme prédictif comme un outil au service du clinicien ajoute à sa fiabilité, et très probablement à son acceptabilité par les utilisateurs. La prédiction en médecine doit se faire dans le cadre de l'aide à la décision, et l'algorithme est un outil au sein du cheminement diagnostic[87].

Cependant, il est probablement encore plus important de produire des modèles prédictifs les plus robustes possibles, et cela passe par la validation. Nous avons réalisé une validation externe, qui confirmé les performances de notre algorithme de stratification du risque d'HAP, mais les données provenaient d'un autre centre expert. Nous avons donc probablement intégré un biais de sélection. De manière générale, cela revient au fait que la qualité des données, qui englobe la population à l'étude, est indispensable pour produire de bons modèles[62].

Les perspectives de ce travail sont la poursuite de l'élaboration d'un outil accessible en ligne, et la validation prospective en population générale. Nous avons développé une application associant chaque patient à ses groupes de probabilité selon ses caractéristiques. Dans notre algorithme, pour maximiser la valeur prédictive négative, nous avons associé au groupe auquel il appartient ayant la probabilité la plus élevée d'HTA secondaire. L'application finale présentera donc plusieurs niveaux d'information : en premier lieu la recommandation (dépistage ou non), puis la possibilité de voir sur quel groupe se base cette recommandation, puis la possibilité de voir l'ensemble des groupes auquel appartient le patient, notamment pour l'exploration de plusieurs diagnostics différents. Cette application intégrera également en parallèle les recommandations actuelles pour le dépistage de l'HTA secondaire[17,18] et si elles s'appliquent ou non au patient décrit.

La seconde perspective est la validation prospective de l'algorithme. Une fois l'application développée, nous envisageons une étude prospective en population générale. Cette étude pour adopter un schéma à 2 bras comparant l'utilisation de l'application à l'application directe des recommandations. La population serait idéalement des patients consultant en médecine de ville, mais une première étape pourrait être à nouveau réalisée en centre expert pour des raisons de puissance et nombre de patients.

L’objectif principal serait de comparer le nombre de patients ayant un diagnostic d’HAP, pou d’une autre cause d’HTA secondaire, avéré, suivant la recommandation de l’application, ou l’application des recommandations. Une étude médico-économique pourrait ensuite être réalisée pour évaluer le bien-fondé de l’utilisation de l’application.

Conclusion

Nous avons développé un modèle prédictif d’HTA secondaire avec des applications prometteuses pour le dépistage de l’HAP. Ce modèle sera implémenté sous la forme d’une application et devra être validé prospectivement dans un essai clinique randomisé.

Bibliographie

1. Forouzanfar MH, Liu P, Roth GA, et al. Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990-2015. *JAMA*. 2017;317(2):165-182. doi:10.1001/jama.2016.19043

2. Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: analysis of worldwide data. *Lancet*. 2005;365(9455):217-223. doi:10.1016/S0140-6736(05)17741-1

3. Zhou B, Carrillo-Larco RM, Danaei G, et al. Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. *The Lancet*. 2021;398(10304):957-980. doi:10.1016/S0140-6736(21)01330-1

4. Masson E. Prévalence de l'hypertension artérielle chez les adultes en France en 2015, étude ESTEBAN. EM-Consulte. Accessed October 30, 2024. https://www.em-consulte.com/article/1202534/prevalence-de-l-hypertension-arterielle-chez-les-a

5. Article - Bulletin épidémiologique hebdomadaire. Accessed October 30, 2024. https://beh.santepubliquefrance.fr/beh/2023/8/2023_8_1.html

6. Mills KT, Bundy JD, Kelly TN, et al. Global Burden of Hypertension: Analysis of Population-based Studies from 89 Countries. *Journal of Hypertension*. 2015;33:e2. doi:10.1097/01.hjh.0000469726.59998.cc

7. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. *Nat Rev Nephrol*. 2020;16(4):223-237. doi:10.1038/s41581-019-0244-2

8. Rimoldi SF, Scherrer U, Messerli FH. Secondary arterial hypertension: when, who, and how to screen? *Eur Heart J*. 2014;35(19):1245-1254. doi:10.1093/eurheartj/eht534

9. Cardoso Claudia R. L., Salles Gil F. Refractory Hypertension and Risks of Adverse Cardiovascular Events and Mortality in Patients With Resistant Hypertension: A Prospective Cohort Study. *Journal of the American Heart Association*. 2020;9(17):e017634. doi:10.1161/JAHA.120.017634

10. Douma S, Petidis K, Doumas M, et al. Prevalence of primary hyperaldosteronism in resistant hypertension: a retrospective observational study. *Lancet*. 2008;371(9628):1921-1926. doi:10.1016/S0140-6736(08)60834-X

11. Courand Pierre-Yves, Dinic Miriana, Lorthioir Aurélien, et al. Resistant Hypertension and Atherosclerotic Renal Artery Stenosis. *Hypertension*. 2019;74(6):1516-1523. doi:10.1161/HYPERTENSIONAHA.119.13393

12. Monticone S, D'Ascenzo F, Moretti C, et al. Cardiovascular events and target organ damage in primary aldosteronism compared with essential hypertension: a systematic

review and meta-analysis. *The Lancet Diabetes & Endocrinology*. 2018;6(1):41-50. doi:10.1016/S2213-8587(17)30319-4

13. Monticone S, Sconfienza E, D'Ascenzo F, et al. Renal damage in primary aldosteronism: a systematic review and meta-analysis. *Journal of Hypertension*. 2020;38(1):3-12. doi:10.1097/HJH.0000000000002216

14. Gupta-Malhotra M, Banker A, Shete S, et al. Essential hypertension vs. secondary hypertension among children. *Am J Hypertens*. 2015;28(1):73-80. doi:10.1093/ajh/hpu083

15. Savard S, Amar L, Plouin PF, Steichen O. Cardiovascular complications associated with primary aldosteronism: a controlled cross-sectional study. *Hypertension*. 2013;62(2):331-336. doi:10.1161/HYPERTENSIONAHA.113.01060

16. Unger Thomas, Borghi Claudio, Charchar Fadi, et al. 2020 International Society of Hypertension Global Hypertension Practice Guidelines. *Hypertension*. 2020;75(6):1334-1357. doi:10.1161/HYPERTENSIONAHA.120.15026

17. Mancia Chairperson G, Kreutz Co-Chair R, Brunström M, et al. 2023 ESH Guidelines for the management of arterial hypertension The Task Force for the management of arterial hypertension of the European Society of Hypertension Endorsed by the International Society of Hypertension (ISH) and the European Renal Association (ERA). *J Hypertens*. Published online June 21, 2023. doi:10.1097/HJH.0000000000003480

18. McEvoy JW, McCarthy CP, Bruno RM, et al. 2024 ESC Guidelines for the management of elevated blood pressure and hypertension: Developed by the task force on the management of elevated blood pressure and hypertension of the European Society of Cardiology (ESC) and endorsed by the European Society of Endocrinology (ESE) and the European Stroke Organisation (ESO). *European Heart Journal*. Published online August 30, 2024:ehae178. doi:10.1093/eurheartj/ehae178

19. Rossi GP, Bisogni V, Rossitto G, et al. Practice Recommendations for Diagnosis and Treatment of the Most Common Forms of Secondary Hypertension. *High Blood Press Cardiovasc Prev*. 2020;27(6):547-560. doi:10.1007/s40292-020-00415-9

20. Siddiqui MA, Mittal PK, Little BP, et al. Secondary Hypertension and Complications: Diagnosis and Role of Imaging. *RadioGraphics*. 2019;39(4):1036-1055. doi:10.1148/rg.2019180184

21. Rossi GP, Bernini G, Caliumi C, et al. A Prospective Study of the Prevalence of Primary Aldosteronism in 1,125 Hypertensive Patients. *Journal of the American College of Cardiology*. 2006;48(11):2293-2300. doi:10.1016/j.jacc.2006.07.059

22. Brown JM, Siddiqui M, Calhoun DA, et al. The Unrecognized Prevalence of Primary Aldosteronism: A Cross-sectional Study. *Ann Intern Med*. 2020;173(1):10-20. doi:10.7326/M20-0065

23. Plouin PF, Amar L, Chatellier G, on behalf of the COMETE-Conn Study Group. Trends in the prevalence of primary aldosteronism, aldosterone-producing adenomas, and surgically correctable aldosterone-dependent hypertension. *Nephrology Dialysis Transplantation*. 2004;19(4):774-777. doi:10.1093/ndt/gfh112

24. Monticone S, Burrello J, Tizzani D, et al. Prevalence and Clinical Manifestations of Primary Aldosteronism Encountered in Primary Care Practice. *J Am Coll Cardiol*. 2017;69(14):1811-1820. doi:10.1016/j.jacc.2017.01.052

25. Buffolo F, Monticone S, Burrello J, et al. Is Primary Aldosteronism Still Largely Unrecognized? *Horm Metab Res*. 2017;49(12):908-914. doi:10.1055/s-0043-119755

26. Krittanawong C, Johnson KW, Rosenson RS, et al. Deep learning for cardiovascular medicine: a practical primer. *Eur Heart J*. 2019;40(25):2058-2073. doi:10.1093/eurheartj/ehz056

27. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Wilson Tang WH. Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension. *Curr Hypertens Rep*. 2018;20(9):75. doi:10.1007/s11906-018-0875-x

28. Buffolo F, Burrello J, Burrello A, et al. Clinical Score and Machine Learning-Based Model to Predict Diagnosis of Primary Aldosteronism in Arterial Hypertension. *Hypertension*. Published online September 7, 2021. doi:10.1161/HYPERTENSIONAHA.121.17444

29. D'Ascenzo F, De Filippo O, Gallone G, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *The Lancet*. 2021;397(10270):199-207. doi:10.1016/S0140-6736(20)32519-8

30. Koren G, Nordon G, Radinsky K, Shalev V. Machine learning of big data in gaining insight into successful treatment of hypertension. *Pharmacol Res Perspect*. 2018;6(3). doi:10.1002/prp2.396

31. Deo RC. Artificial Intelligence and Machine Learning in Cardiology. *Circulation*. 2024;149(16):1235-1237. doi:10.1161/CIRCULATIONAHA.123.065469

32. Padmanabhan S, Tran TQB, Dominiczak AF. Artificial Intelligence in Hypertension: Seeing Through a Glass Darkly. *Circ Res*. 2021;128(7):1100-1118. doi:10.1161/CIRCRESAHA.121.318106

33. Hall JE, Guyton AC. *Guyton and Hall Textbook of Medical Physiology*. 12th ed. Saunders/Elsevier; 2011.

34. Shevchenko YL, Tsitlik JE. 90th Anniversary of the Development by Nikolai S. Korotkoff of the Auscultatory Method of Measuring Blood Pressure. *Circulation*. 1996;94(2):116-118. doi:10.1161/01.CIR.94.2.116

35. Vander AJ, Sherman JH, Luciano DS, Brière R. *Physiologie humaine*. 3e éd. Chenelière/McGraw-Hill; 1995.

36. Effects of Treatment on Morbidity in Hypertension: Results in Patients With Diastolic Blood Pressures Averaging 115 Through 129 mm Hg. *JAMA*. 1967;202(11):1028-1034. doi:10.1001/jama.1967.03130240070013

37. Dahlöf B, Devereux RB, Kjeldsen SE, et al. Cardiovascular morbidity and mortality in the Losartan Intervention For Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. *Lancet*. 2002;359(9311):995-1003. doi:10.1016/S0140-6736(02)08089-3

38. Parving HH, Lehnert H, Bröchner-Mortensen J, Gomis R, Andersen S, Arner P. The Effect of Irbesartan on the Development of Diabetic Nephropathy in Patients with Type 2 Diabetes. *New England Journal of Medicine*. 2001;345(12):870-878. doi:10.1056/NEJMoa011489

39. Williams B, MacDonald TM, Morant S, et al. Spironolactone versus placebo, bisoprolol, and doxazosin to determine the optimal treatment for drug-resistant hypertension (PATHWAY-2): a randomised, double-blind, crossover trial. *Lancet*. 2015;386(10008):2059-2068. doi:10.1016/S0140-6736(15)00257-3

40. Jamerson K, Weber MA, Bakris GL, et al. Benazepril plus Amlodipine or Hydrochlorothiazide for Hypertension in High-Risk Patients. *New England Journal of Medicine*. 2008;359(23):2417-2428. doi:10.1056/NEJMoa0806182

41. Omura M, Saito J, Yamaguchi K, Kakuta Y, Nishikawa T. Prospective study on the prevalence of secondary hypertension among hypertensive patients visiting a general outpatient clinic in Japan. *Hypertens Res*. 2004;27(3):193-202. doi:10.1291/hypres.27.193

42. Anderson GH, Blakeman N, Streeten DH. The effect of age on prevalence of secondary forms of hypertension in 4429 consecutively referred patients. *J Hypertens*. 1994;12(5):609-615. doi:10.1097/00004872-199405000-00015

43. Singh M, Singh AK, Pandey P, Chandra S, Singh KA, Gambhir IS. Molecular genetics of essential hypertension. *Clin Exp Hypertens*. 2016;38(3):268-277. doi:10.3109/10641963.2015.1116543

44. de Freminville JB, Gardini M, Cremer A, et al. Prevalence and Risk Factors for Secondary Hypertension in Young Adults. *Hypertension*. 2024;81(11):2340-2349. doi:10.1161/HYPERTENSIONAHA.124.22753

45. de Freminville JB, Amar L, Azizi M, Mallart-Riancho J. Endocrine causes of hypertension: literature review and practical approach. *Hypertens Res*. 2023;46(12):2679-2692. doi:10.1038/s41440-023-01461-1

46. Amar L, Baguet JP, Bardet S, et al. SFE/SFHTA/AFCE primary aldosteronism consensus: Introduction and handbook. *Annales d'Endocrinologie*. 2016;77(3):179-186. doi:10.1016/j.ando.2016.05.001

47. Ananda RA, Gwini S, Beilin LJ, et al. Relationship Between Renin, Aldosterone, Aldosterone-to-Renin Ratio and Arterial Stiffness and Left Ventricular Mass Index in Young Adults. *Circulation*. 0(0). doi:10.1161/CIRCULATIONAHA.124.070039

48. Douillard C, Houillier P, Nussberger J, Girerd X. SFE/SFHTA/AFCE Consensus on Primary Aldosteronism, part 2: First diagnostic steps. *Ann Endocrinol (Paris)*. 2016;77(3):192-201. doi:10.1016/j.ando.2016.02.003

49. Rossi GP, Rossi FB, Guarnieri C, Rossitto G, Seccia TM. Clinical Management of Primary Aldosteronism: An Update. *Hypertension*. 2024;81(9):1845-1856. doi:10.1161/HYPERTENSIONAHA.124.22642

50. Rossi GP, Bernini G, Caliumi C, et al. A prospective study of the prevalence of primary aldosteronism in 1,125 hypertensive patients. *J Am Coll Cardiol*. 2006;48(11):2293-2300. doi:10.1016/j.jacc.2006.07.059

51. Rossi GP, Bagordo D, Rossi FB, Pintus G, Rossitto G, Seccia TM. “Essential” arterial hypertension: time for a paradigm change. *J Hypertens*. 2024;42(8):1298-1304. doi:10.1097/HJH.0000000000003767

52. Rossi GP, Rossi FB, Guarnieri C, Rossitto G, Seccia TM. Clinical Management of Primary Aldosteronism: An Update. *Hypertension*. 2024;81(9):1845-1856. doi:10.1161/HYPERTENSIONAHA.124.22642

53. Gornik HL, Persu A, Adlam D, et al. First international consensus on the diagnosis and management of fibromuscular dysplasia. *J Hypertens*. 2019;37(2):229-252. doi:10.1097/HJH.0000000000002019

54. Raman G, Adam GP, Halladay CW, Langberg VN, Azodo IA, Balk EM. Comparative Effectiveness of Management Strategies for Renal Artery Stenosis: An Updated Systematic Review. *Ann Intern Med*. 2016;165(9):635-649. doi:10.7326/M16-1053

55. Lenders JWM, Kerstens MN, Amar L, et al. Genetics, diagnosis, management and future directions of research of phaeochromocytoma and paraganglioma: a position statement and consensus of the Working Group on Endocrine Hypertension of the European Society of Hypertension. *Journal of Hypertension*. 2020;38(8):1443-1456. doi:10.1097/HJH.0000000000002438

56. Lenders JWM, Duh QY, Eisenhofer G, et al. Pheochromocytoma and paraganglioma: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab*. 2014;99(6):1915-1942. doi:10.1210/jc.2014-1498

57. Zelinka T, Petrák O, Turková H, et al. High incidence of cardiovascular complications in pheochromocytoma. *Horm Metab Res*. 2012;44(5):379-384. doi:10.1055/s-0032-1306294

58. Castinetti F, De Freminville JB, Guerin C, Cornu E, Sarlon G, Amar L. Controversies about the systematic preoperative pharmacological treatment before

pheochromocytoma or paraganglioma surgery. *Eur J Endocrinol*. 2022;186(5):D17-D24. doi:10.1530/EJE-21-0692

59. Taïeb D, Hicks RJ, Hindié E, et al. European Association of Nuclear Medicine Practice Guideline/Society of Nuclear Medicine and Molecular Imaging Procedure Standard 2019 for radionuclide imaging of phaeochromocytoma and paraganglioma. *Eur J Nucl Med Mol Imaging*. 2019;46(10):2112-2137. doi:10.1007/s00259-019-04398-1

60. Fleseriu M, Auchus R, Bancos I, et al. Consensus on diagnosis and management of Cushing's disease: a guideline update. *Lancet Diabetes Endocrinol*. 2021;9(12):847-875. doi:10.1016/S2213-8587(21)00235-7

61. Navarro CLA, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;375:n2281. doi:10.1136/bmj.n2281

62. Shah Rashmee U. We Don't Need More Data, We Need the Right Data. *Circulation*. 2020;142(3):197-198. doi:10.1161/CIRCULATIONAHA.120.045968

63. Siontis KC, Noseworthy PA, Arghami A, et al. Use of Artificial Intelligence Tools Across Different Clinical Settings: A Cautionary Tale. *Circ: Cardiovascular Quality and Outcomes*. 2021;14(9). doi:10.1161/CIRCOUTCOMES.121.008153

64. Serra-Burriel M, Locher L, Vokinger KN. Development Pipeline and Geographic Representation of Trials for Artificial Intelligence/Machine Learning–Enabled Medical Devices (2010 to 2023). *NEJM AI*. 2023;1(1):AIp2300038. doi:10.1056/AIpc2300038

65. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593

66. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer; 2009. doi:10.1007/978-0-387-84858-7

67. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv.org. February 9, 2018. Accessed October 31, 2024. https://arxiv.org/abs/1802.03426v3

68. Hegland M. The apriori algorithm ? a tutorial. In: *Mathematics and Computation in Imaging Science and Information Processing*. Vol Volume 11. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore. WORLD SCIENTIFIC; 2007:209-262. doi:10.1142/9789812709066_0006

69. Pei J, Han J, Mortazavi-Asl B, et al. Mining sequential patterns by pattern-growth: the PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*. 2004;16(11):1424-1440. doi:10.1109/TKDE.2004.77

70. Mihan A, Pandey A, Spall HGV. Mitigating the risk of artificial intelligence bias in cardiovascular care. *The Lancet Digital Health*. 2024;0(0). doi:10.1016/S2589-7500(24)00155-9

71. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*. 2021;3(11):e745-e750. doi:10.1016/S2589-7500(21)00208-9

72. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak*. 2019;19(1):146. doi:10.1186/s12911-019-0874-0

73. Funer F. The Deception of Certainty: how Non-Interpretable Machine Learning Outcomes Challenge the Epistemic Authority of Physicians. A deliberative-relational Approach. *Med Health Care Philos*. 2022;25(2):167-178. doi:10.1007/s11019-022-10076-1

74. Layton AT. AI, Machine Learning, and ChatGPT in Hypertension. *Hypertension*. 0(0). doi:10.1161/HYPERTENSIONAHA.124.19468

75. Chang W, Liu Y, Xiao Y, et al. A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics (Basel)*. 2019;9(4):178. doi:10.3390/diagnostics9040178

76. Wu X, Yuan X, Wang W, et al. Value of a Machine Learning Approach for Predicting Clinical Outcomes in Young Patients With Hypertension. *Hypertension*. 2020;75(5):1271-1278. doi:10.1161/HYPERTENSIONAHA.119.13404

77. Wang MH, Li NF, Luo Q, et al. Development and validation of a novel diagnostic nomogram model to predict primary aldosteronism in patients with hypertension. *Endocrine*. 2021;73(3):682-692. doi:10.1007/s12020-021-02745-7

78. Liu Y, Wang M, Qiu X, et al. A novel clinical-imaging nomogram for predicting primary aldosteronism in patients with hypertension. *Hypertens Res*. 2023;46(12):2603-2612. doi:10.1038/s41440-023-01374-z

79. Rossi GP, Rossi E, Pavan E, et al. Screening for primary aldosteronism with a logistic multivariate discriminant analysis. *Clin Endocrinol (Oxf)*. 1998;49(6):713-723. doi:10.1046/j.1365-2265.1998.00608.x

80. Yamashita T, Shimizu S, Koyama M, et al. Screening of primary aldosteronism by clinical features and daily laboratory tests: combination of urine pH, sex, and serum K. *J Hypertens*. 2018;36(2):326-334. doi:10.1097/HJH.0000000000001511

81. Kietsiriroje N, Wonghirundecha R, Suntornlohanakul O, Murray RD. Construction of a predictive scoring system as a guide to screening and confirmation of the diagnosis of primary aldosteronism. *Clin Endocrinol (Oxf)*. 2020;92(3):196-205. doi:10.1111/cen.14142

82. Diao X, Huo Y, Yan Z, et al. An Application of Machine Learning to Etiological Diagnosis of Secondary Hypertension: Retrospective Study Using Electronic Medical Records. *JMIR Med Inform*. 2021;9(1):e19739. doi:10.2196/19739

83. Reel PS, Reel S, Kralingen JC van, et al. Machine learning for classification of hypertension subtypes using multi-omics: A multi-centre, retrospective, data-driven study. *eBioMedicine*. 2022;84. doi:10.1016/j.ebiom.2022.104276

84. Venot A, Burgun A, Quantin C. *Informatique médicale, e-Santé*. Springer; 2013. doi:10.1007/978-2-8178-0338-8

85. Wang Q, Dong H, Li HW, et al. Development of a diagnostic model for pre-washout screening of primary aldosteronism. *J Endocrinol Invest*. Published online March 27, 2024. doi:10.1007/s40618-024-02337-y

86. Ginzberg SP, Kalva S, Wirtalla CJ, et al. Development of a risk-prediction model for primary aldosteronism in veterans with hypertension. *Surgery*. 2024;175(1):73-79. doi:10.1016/j.surg.2023.04.063

87. Adler-Milstein J, Chen JH, Dhaliwal G. Next-Generation Artificial Intelligence for Diagnosis: From Predicting Diagnostic Labels to "Wayfinding." *JAMA*. 2021;326(24):2467. doi:10.1001/jama.2021.22396

88. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*. 2020;22(6):e15154. doi:10.2196/15154

89. Reyna MA, Nsoesie EO, Clifford GD. Rethinking Algorithm Performance Metrics for Artificial Intelligence in Diagnostic Medicine. *JAMA*. 2022;328(4):329. doi:10.1001/jama.2022.10561

90. Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. *Int J Med Inform*. 2017;102:21-28. doi:10.1016/j.ijmedinf.2017.02.006

91. Boussadi A, Zapletal E. A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2. *BMC Medical Informatics and Decision Making*. 2017;17(1):120. doi:10.1186/s12911-017-0513-6

92. Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. *Stud Health Technol Inform*. 2010;160(Pt 1):193-197.

93. Zapletal E, Bibault JE, Giraud P, Burgun A. Integrating Multimodal Radiation Therapy Data into i2b2. *Applied Clinical Informatics*. 2018;09:377-390. doi:10.1055/s-0038-1651497

94. Kimball R, Ross M. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley; 2002.

95. Degoulet P, Chatellier G, Devriès C, Lavril M, Ménard J. Computer-assisted techniques for evaluation and treatment of hypertensive patients. *Am J Hypertens*. 1990;3(2):156-163. doi:10.1093/ajh/3.2.156

96. Degoulet P, Marin L, Lavril M, et al. The HEGP component-based clinical information system. *Int J Med Inform*. 2003;69(2-3):115-126. doi:10.1016/s1386-5056(02)00101-6

97. Neuraz A. Compréhension du langage naturel pour le dossier patient informatisé: accès à l'information et extraction d'information.

98. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:181004805 [cs]*. Published online May 24, 2019. Accessed September 17, 2021. http://arxiv.org/abs/1810.04805

99. Halimi JM, de Fréminville JB, Gatault P, et al. Long-term impact of cardiorenal syndromes on major outcomes based on their chronology: a comprehensive French nationwide cohort study. *Nephrol Dial Transplant*. Published online April 19, 2022:gfac153. doi:10.1093/ndt/gfac153

100. de Freminville JB, Halimi JM, Maisons V, et al. Unsupervised Cluster Analysis in Patients with Cardiorenal Syndromes: Identifying Vascular Aspects. *Journal of Clinical Medicine*. 2024;13(11):3159. doi:10.3390/jcm13113159

101. Halimi JM, de Fréminville JB, Gatault P, et al. Characteristics and Prognosis of Patients With Hypertensive Encephalopathy: A French Nationwide Cohort Study. *Hypertension*. 2023;80(8):1716-1727. doi:10.1161/HYPERTENSIONAHA.123.21226

102. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ*. 2010;341:c4226. doi:10.1136/bmj.c4226

103. ENS@T - home. Accessed November 1, 2024. https://ensat.wildapricot.org/

104. Plouin PF, Gimenez-Roqueplo AP, Bertagna X. Le réseau national COMETE sur les tumeurs de la surrénale. *Bulletin de l'Académie Nationale de Médecine*. 2008;192(1):73-85. doi:10.1016/S0001-4079(19)32869-9

105. Délibération n° 2018-155 du 3 mai 2018 portant homologation de la méthodologie de référence relative aux traitements de données à caractère personnel mis en œuvre dans le cadre des recherches n'impliquant pas la personne humaine, des études et évaluations dans le domaine de la santé (MR-004) - Légifrance. Accessed November 2, 2024. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037187498

106. La loi Informatique et Libertés. Accessed November 2, 2024. https://www.cnil.fr/fr/la-loi-informatique-et-libertes

107. Viera AJ, Neutze DM. Diagnosis of secondary hypertension: an age-based approach. *Am Fam Physician*. 2010;82(12):1471-1478.

108. Trifanescu R, Carsote M, Caragheorgheopol A, et al. Screening for secondary endocrine hypertension in young patients. *Maedica (Bucur)*. 2013;8(2):108-115.

109. Noilhan C, Barigou M, Bieler L, Amar J, Chamontin B, Bouhanick B. Causes of secondary hypertension in the young population: A monocentric study. *Ann Cardiol Angeiol (Paris)*. 2016;65(3):159-164. doi:10.1016/j.ancard.2016.04.016

110. Marney AM, Brown NJ. Aldosterone and end-organ damage. *Clin Sci (Lond)*. 2007;113(6):267-278. doi:10.1042/CS20070123

111. Burrello Jacopo, Monticone Silvia, Losano Isabel, et al. Prevalence of Hypokalemia and Primary Aldosteronism in 5100 Patients Referred to a Tertiary Hypertension Unit. *Hypertension*. 2020;75(4):1025-1033. doi:10.1161/HYPERTENSIONAHA.119.14063

112. Thompson Philip, Logan Ian, Tomson Charles, Sheerin Neil, Ellam Timothy. Obesity, Sex, Race, and Early Onset Hypertension. *Hypertension*. 2020;76(3):859-865. doi:10.1161/HYPERTENSIONAHA.120.15557

113. Hall JE, Mouton AJ, da Silva AA, et al. Obesity, kidney dysfunction, and inflammation: interactions in hypertension. *Cardiovasc Res*. 2021;117(8):1859-1876. doi:10.1093/cvr/cvaa336

114. Spyroglou A, Handgriff L, Müller L, et al. The metabolic phenotype of patients with primary aldosteronism: impact of subtype and sex - a multicenter-study of 3566 Caucasian and Asian subjects. *Eur J Endocrinol*. 2022;187(3):361-372. doi:10.1530/EJE-22-0040

115. La Batide-Alanore A, Chatellier G, Plouin PF. Diabetes as a marker of pheochromocytoma in hypertensive patients. *J Hypertens*. 2003;21(9):1703-1707. doi:10.1097/00004872-200309000-00020

116. Li R, Zhou J, Wang L. ESTIMATION OF THE BINARY LOGISTIC REGRESSION MODEL PARAMETER USING BOOTSTRAP RE-SAMPLING. *Latin American Applied Research*. Published online 2018.

117. Bootstrapping the coefficients of multiple logistic regression model in medicine data | Asian-European Journal of Mathematics. Accessed January 23, 2023. https://www.worldscientific.com/doi/10.1142/S1793557122502485

118. Cevenini G, Barbini P. A bootstrap approach for assessing the uncertainty of outcome probabilities when using a scoring system. *BMC Med Inform Decis Mak*. 2010;10:45. doi:10.1186/1472-6947-10-45

119. S B, L A, Al F, et al. Criteria for diagnosing primary aldosteronism on the basis of liquid chromatography-tandem mass spectrometry determinations of plasma aldosterone concentration. *Journal of hypertension*. 2018;36(7). doi:10.1097/HJH.0000000000001735

120. Mulatero P, Tizzani D, Viola A, et al. Prevalence and Characteristics of Familial Hyperaldosteronism. *Hypertension*. 2011;58(5):797-803. doi:10.1161/HYPERTENSIONAHA.111.175083

121. Elliott WJ. Renovascular hypertension: an update. *J Clin Hypertens (Greenwich)*. 2008;10(7):522-533. doi:10.1111/j.1751-7176.2008.07788.x

122. Yoshihara F, Nishikimi T, Yoshitomi Y, et al. Left Ventricular Structural and Functional Characteristics in Patients with Renovascular Hypertension, Primary Aldosteronism and Essential Hypertension. *American Journal of Hypertension*. 1996;9(6):523-528. doi:10.1016/0895-7061(96)00031-3

123. Bhalla V, Textor SC, Beckman JA, et al. Revascularization for Renovascular Disease: A Scientific Statement From the American Heart Association. *Hypertension*. 2022;79(8). doi:10.1161/HYP.0000000000000217

124. Castinetti F, De Freminville JB, Guerin C, Cornu E, Sarlon G, Amar L. Controversies about the systematic preoperative pharmacological treatment before pheochromocytoma or paraganglioma surgery. *Eur J Endocrinol*. 2022;186(5):D17-D24. doi:10.1530/EJE-21-0692

125. Eisenhofer G, Durán C, Cannistraci CV, et al. Use of Steroid Profiling Combined With Machine Learning for Identification and Subtype Classification in Primary Aldosteronism. *JAMA Netw Open*. 2020;3(9):e2016209. doi:10.1001/jamanetworkopen.2020.16209

126. Salzberg SL. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach Learn*. 1994;16(3):235-240. doi:10.1007/BF00993309

127. La Cava WG, Lee PC, Ajmal I, et al. A flexible symbolic regression method for constructing interpretable clinical prediction models. *npj Digit Med*. 2023;6(1):1-14. doi:10.1038/s41746-023-00833-8

Table des matières

INTRODUCTION 9

CHAPITRE 1 - CONTEXTE MEDICAL 11

La pression artérielle 12
La mesure de la pression artérielle 13
L'hypertension artérielle 15

Hypertension artérielle secondaire 17
Hyperaldostéronisme primaire 18
Hypertension artérielle rénovasculaire 21
Phéochromocytomes et paragangliomes (PPGL) 22
Les autres formes d'HTA secondaire 23

CHAPITRE 2 – L'INTELLIGENCE ARTIFICIELLE ET LES MODELES PREDICTIFS EN MEDECINE ET DANS L'HYPERTENSION ARTERIELLE SECONDAIRE 24

Modèles prédictifs en médecine 25
Étapes du développement d'un modèle prédictif. 25
Métriques et évaluation des modèles 26
Utilisation de l'intelligence artificielle 27

L'apprentissage automatique : grands principes et principaux algorithmes 28
Le principe de l'apprentissage automatique 28
Analyses non supervisées 28
Analyses supervisées 30

Traitement Automatique des Langues Naturelles (*Natural Language Processing*) 37
Word-Embedding 37

Réseaux de neurones et apprentissage profond (*Deep learning*) 38
LSTM/Bi-LSTM 39

Les biais liés aux algorithmes d'intelligence artificielle dans le domaine de la médecine 40

Etat de l'art des algorithmes prédictifs dans le domaine de l'HTA secondaire 40

Aide à la décision – aide au diagnostic 45

CHAPITRE 3 : CONSTRUCTION D'UNE BASE DE DONNEES DE PATIENTS HYPERTENDUS 47

Description de l'entrepôt de données de l'HEGP 48
Données médicales à l'Hôpital Européen Georges Pompidou 48
Questionnaires d'hypertension artérielle 51

Méthodes d'identification des diagnostics d'hypertension artérielle secondaire 53
Identification à partir de données textuelles 53
Identification à partir du PMSI 60
Identification à partir des questionnaires 62

Bases de données existantes 63
Algorithme retenu pour l'identification du diagnostic final 63

Mise en forme des bases de données finales 64
Formatage des données 64
Sélection des patients 65
Qualité des données 66
Sélection des variables 68
Binarisation des variables 70

Base de données Bordeaux 71

Aspects technico-règlementaires 71

CHAPITRE 4 : HYPERTENSION ARTERIELLE SECONDAIRE CHEZ LE SUJET JEUNE 73

Introduction 74

Objectifs 75

Présentation du manuscrit 76

Complément de méthodes et résultats 104
Méthodes : estimation du risque d'HTA secondaire par méthode régression logistique et bootstrap 104

Complément de discussion : apport de l'algorithme dans la stratégie de dépistage des patients jeunes 106

CHAPITRE 5 : MODELE PREDICTIF D'HYPERTENSION ARTERIELLE SECONDAIRE 108

Introduction 109

Objectifs 110

Présentation du manuscrit 111

Complément de méthodes et résultats 142
Méthodes complémentaires: développement de l'algorithme de stratification 142
Résultats complémentaires 147

Complément de discussion 170

CONCLUSION ET PERSPECTIVES GENERALES 173

BIBLIOGRAPHIE 176

TABLE DES MATIERES 187

TABLE DES ILLUSTRATIONS 189

Table des illustrations

Figure 1 - pression artérielle à différents niveaux du système circulatoire chez une personne en position allongée 12
Figure 2 - Mesure non-invasive de la pression artérielle, mesure auscultatoire vs. mesure oscillométrique 13
Figure 3 - tensiomètre oscillométrique 14
Figure 4 - mesure de la pression artérielle 14
Figure 5 – facteurs impliqués dans la physiopathologie de l'Hypertension artérielle 17
Figure 6 - algorithme diagnostic de l'hyperaldostéronisme primaire 20
Figure 7 - exemple de surapprentissage 26
Figure 8 - nombre de publications (incidence) en recherchant le terme « intelligence artificielle » sur pubmed depuis 1950 27
Figure 9 - exemple d'apprentissage non supervisé. L'algorithme a identifié trois types de formes différentes. Il saura ensuite les reconnaitre sur un nouveau jeu de données 29
Figure 10 - Principe de développement d'un algorithme d'apprentissage automatique supervisé, avec séparation du jeu de données initial en un jeu d'entrainement et un jeu de test 32
Figure 11 - Relations linéaires entre les mots vectorisés par embedding. La relation entre homme et femme correspond à un certain déplacement dans l'espace, tout comme le lien entre participe présent et passé, ou entre un pays et sa capitale 37
Figure 12 – architecture d'un réseau de neurones 39
Figure 13 – algorithmes prédictifs d'HTA secondaire 43
Figure 14 - raisonnement hypothético-déductif 45
Figure 15 - données stockées dans l'entrepôt de l'HEGP 49
Figure 16 - quantité de données disponible selon le type 50
Figure 17 - questionnaire d'hypertension artérielle 52
Figure 18 - Dictionnaire 54
Figure 19 – outil BRAT. 56
Figure 20 - poster présentant les résultats de la reconnaissance diagnostique par NLP 58
Figure 21 - performance par catégorie de diagnostic 59
Figure 22 - matrice de confusion des résultats du NER spécialisé 60
Figure 23 – diagnostic renseigné sur le questionnaire d'hypertension artérielle 62
Figure 24 - algorithme de choix du diagnostic selon les sources disponibles et leur concordance 64
Figure 25 - exemple de présentation des données extraites 65
Figure 26 - Flow Chart patients final des patients inclus dans la base HEGP 66
Figure 27 - représentation de l'ancienneté de l'HTA en fonction de l'âge après suppression des valeurs aberrantes. Tous les points au-dessus de la droite identité (ligne rouge) et les valeurs aberrantes ont été exclus 66
Figure 28 - violin plot de la taille après suppression des valeurs aberrantes 67
Figure 29 - Représentation de la PA systolique en fonction de la PA diastolique. Tous les points au-dessous de la droite identité (ligne rouge) et les valeurs aberrantes ont été exclus 67
Figure 30 - représentation des données manquantes par variable après binarisation avant sélection de la base finale 69

Figure 31 - représentation des 2 principales dimensions de l'analyse en composantes principales pour les variables quantitatives, avec visualisation du diagnostic d'HTA primaire 147
Figure 32 - représentation des 2 principales dimensions de l'analyse en composantes principales pour les variables quantitatives, avec visualisation du diagnostic d'HAP 148
Figure 33 - contribution des variables dans les 5 premières dimensions de l'ACP 148
Figure 34 - représentation des 2 principales dimensions de l'analyse des composantes multiples pour les variables catégorielles, avec visualisation du diagnostic d'HTA primaire 149
Figure 35 - représentation des 2 principales dimensions de l'analyse des composantes multiples pour les variables catégorielles, avec visualisation du diagnostic d'HAP 149
Figure 36 - contribution des variables dans les 5 première dimensions de l'ACM 150
Figure 37 - représentation à l'échelle des différents clusters selon le nombre choisi, selon la kaliémie et la pression artérielle systolique clinique 153
Figure 38 - répartition des patients selon 3 variables dans un modèle à 3 clusters 153
Figure 39 - Représentation d'un clustering UMAP selon le diagnostic d'HTA primaire 154
Figure 40 - Représentation d'un clustering UMAP selon le diagnostic d'Hyperaldostéronisme primaire 155
Figure 41 - courbe ROC pour l'algorithme de stratification pour l'HTA rénovasculaire 156
Figure 42 - courbe ROC pour la régression logistique pour l'HTA rénovasculaire 158
Figure 43 - importance des différentes variables dans le modèle de régression logistique 160
Figure 44 - courbe ROC pour XGBoost pour l'HTA rénovasculaire 161
Figure 45 - importance des différentes variables dans le modèle XGBoost 162
Figure 46 - courbe ROC pour l'algorithme de stratification pour les PPGL 163
Figure 47 - importance des différentes variables dans le modèle de régression logistique 165
Figure 48 - courbe ROC pour XGBoost pour les PPGL 166
Figure 49 - importance des différentes variables dans le modèle XGBoost 167
Figure 50 - exemple d'utilisation de l'application; à gauche les caractéristiques patient, à droite les probabilités diagnostiques (RR et probabilités absolues) 169