Résultats projet PREDIVASC-IOC

1. Rappel des objectifs

1.1. Contexte

Le syndrome d'apnées-hypopnées obstructives du sommeil (SAHOS) est une pathologie chronique fréquente (environ 10% de la population adulte) très hétérogène dans sa présentation clinique, la nature et la fréquence de ses complications. Évaluer lors du diagnostic initial le risque de survenue d'une complication cardiovasculaire chez un patient exploré pour un SAHOS, constitue pour le clinicien un enjeu majeur de prise en charge. De plus, le risque cardiovasculaire associé au SAHOS étant multifactoriel, l'évaluation du pronostic cardiovasculaire impose de prendre en compte de nombreux paramètres liés à la sévérité des troubles respiratoires du sommeil, aux autres facteurs de risque cardiovasculaire (hypertension, diabète, dyslipidémie, etc.) et à leur prise en charge pharmacologique.

Le SAHOS constitue le principal thème de recherche en pneumologie au centre hospitalier universitaire d'Angers. L'équipe de pneumologie Angevine a fait partie des centres précurseurs en France pour le développement du diagnostic du SAHOS et de son traitement par la pression positive continue. Dès 1985, Jean-Louis Racineux a initié une collaboration avec le laboratoire de recherche de l'ESEO, dirigé par Jean Pinguet, aboutissant sur un dépôt de brevet et la création de l'entreprise CIDELEC en 1990. Au cours de ces 15 dernières années, la recherche pneumologique angevine s'est focalisée sur l'étude de la morbidité associée aux apnées du sommeil, en particulier dans le domaine des pathologies cardiométaboliques grâce, en particulier, à la mise en place par l'institut de recherche en santé respiratoire des Pays de la Loire (IRSR), en 2007, de la Cohorte Sommeil des Pays de la Loire. Elle comporte aujourd'hui plus de 15 000 adultes investigués pour suspicion clinique de syndrome d'apnées du sommeil avec les outils CIDELEC, dans un des 9 centres participants, et parfaitement caractérisés à l'inclusion en termes de données anthropomorphiques, socio-économiques, de sévérité des troubles respiratoires du sommeil et de comorbidités. Elle a donné lieu à de nombreuses études publiées dans des revues anglo-saxonnes. En 2017, l'obtention, par l'IRSR, d'un appariement avec les données du système national d'information inter-régimes de l'Assurance maladie (SNIIRAM) nous offre une opportunité unique pour évaluer de façon exhaustive et dynamique le parcours de soin et la morbidité cardiovasculaire incidente pour plus de 12 000 patients, constituant l'étude prospective morbidité et sommeil (ERMES) qui tient compte des caractéristiques à l'inclusion, de l'adhérence au traitement du SAHOS, des comorbidités et médicaments connus pour impacter le pronostic cardiovasculaire.

L'objectif du projet de thèse, PREDIVASC, est de développer un outil permettant au clinicien d'évaluer lors du diagnostic du SAHOS le pronostic évolutif de la maladie dans le domaine cardiovasculaire. L'outil devra aussi bien tenir compte des contraintes temps, matérielles et financières. Ainsi, il pourra être facilement intégré au dispositif de diagnostic du SAHOS, sans nécessiter de matériel ou d'examens supplémentaires.

PREDIVASC a un premier versant associé au domaine de la recherche fondamentale puis un second avec un objectif d'industrialisation avec l'intégration du produit dans les dispositifs de diagnostic des troubles du sommeil proposés par la société CIDELEC (PREDIVASC-IOC). CIDELEC étant le financeur de ce projet, à travers la chaire PEPITES mise en place avec l'ESEO. La thèse est rattachée au laboratoire d'acoustique de l'université du Mans (LAUM), une unité mixte de recherche de l'université du Mans et du centre national de la recherche scientifique (unité mixte de recherche

6613). Ce projet a été autorisé par un accord de la commission nationale de l'informatique et des libertés (CNIL) pour l'exploitation des données de la cohorte sommeil appariées au SNIIRAM (demande d'autorisation numéro 921411).

1.2. Introduction

Un sommeil de mauvaise qualité peut avoir de multiples conséquences sur la santé, à court terme comme à long terme. Une personne présentant une somnolence diurne importante peut se rendre chez son médecin pour passer un examen du sommeil. Ce dernier consiste à dormir avec plusieurs capteurs enregistrant les différentes activités physiologiques du corps : le système cardiovasculaire, le système respiratoire, etc. Le lendemain, le médecin peut lire l'enregistrement du sommeil, donner son verdict sur la présence d'un trouble du sommeil et mettre en place un traitement. Le SAHOS se manifeste par des pauses respiratoires au cours du sommeil, d'une durée de plus de 10 secondes et allant jusqu'à plus d'une minute parfois. Les différents systèmes physiologiques du corps humain réagissent à ce manque d'oxygène et génèrent automatiquement une réaction qui permet une reprise ventilatoire. Ces réactions à répétitions vont avoir de graves conséquences sur la santé, à court terme comme à long terme.

En effet, de nombreuses études ont montré un lien associatif entre l'augmentation du risque cardiovasculaire et des indicateurs de ces réactions nocturnes, à répétition, des différents systèmes physiologiques. Cependant, les informations sur le sommeil ne sont pas considérées pour estimer un risque cardiovasculaire, alors que cela permettrait une meilleure stratification du risque des patients et ainsi une meilleure prise en charge.

Aujourd'hui, les maladies cardiovasculaires sont la première cause de décès dans le monde et la deuxième en France. Le Forum économique mondial estime qu'elles vont coûter jusqu'à 1 044 milliards de dollars en 2030. Pour des raisons de santé et économiques, leur prévention est un enjeu majeur.

De plus, la littérature a déjà montré que les patients atteints de SAHOS ont un risque cardiovasculaire plus élevé que les patients sains. Cependant, les estimateurs de risque cardiovasculaire actuels sont développés sur des populations générales et non pas des patients SAHOS. Ainsi, de nouveaux estimateurs pourraient être développés, plus spécifiques, à la population SAHOS.

Par ailleurs, les estimateurs de risque existant sont basés sur des modèles statistiques reposant sur des hypothèses contraignantes. Le risque cardiovasculaire étant multifactoriel, les méthodes plus modernes d'intelligence artificielle, s'affranchissant des hypothèses des statistiques classiques, permettent de combiner tous les facteurs de risque, cliniques comme sommeils.

Aujourd'hui, il existe quelques études qui proposent un estimateur de risque cardiovasculaire, basé sur des modèles d'intelligence artificielle et construit avec des variables cliniques et des indicateurs sur le sommeil des patients SAHOS. Mais ces estimateurs ne permettent pas au clinicien d'évaluer le risque cardiovasculaire tout en diagnostiquant les troubles du sommeil. Ils nécessitent des informations parfois complexes, couteuses et pas disponibles rapidement.

Enfin, des méthodes encore plus modernes permettent de s'affranchir de l'utilisation d'indicateurs sur le sommeil, en utilisant directement les signaux enregistrés pendant le sommeil. D'après notre connaissance, aucune étude n'a encore proposé une telle application, c'est l'objet de notre travail de recherche.

2. Résumé des données

Come évoqué dans la partie contexte, la Cohorte Sommeil des Pays de la Loire comporte plus de 15000 adultes investigués pour suspicion clinique de syndrome d'apnées du sommeil. De plus, les données ont été appariées avec les données du SNIIRAM offrant la possibilité de suivre le parcours de soin et la morbidité cardiovasculaire incidente des patients (étude ERMES, demande d'autorisation numéro 908157). Les variables pertinentes pour l'étude PREDIVASC-IOC se divisent en deux catégories : les variables de la cohorte et les variables issues du SNIIRAM.

2.1. Données cohorte : clinique + sommeil

Parmi les variables de la cohorte il y a des variables dites « clinique » et les variables dites « sommeil » extraites des enregistrements de sommeil : index d'apnée hypopnée, index de désaturation en oxygène, temps en dessous de 90% de saturation en oxygène, charge hypoxique, position, paramètres de variabilité de la fréquence cardiaque... Ces données seront précisées tout au long de l'étude. Les variables sont présentées dans le tableau ci – dessous.

Variables cohorte	**Type**
Enregistrement sommeil	Signaux continus
Date de l'examen	Jour/mois/année
Age	Variable continue : Années
Genre	Variable binaire : Homme / Femme
Indice de masse corporel	Variable continue : kg/m²
Consommation d'alcool	Variable continue : nombre de verre par jour
Consommation de tabac	Variable catégorielle : fume/arrêt/jamais
Pression artérielle systolique	Variable continue : mmol/L
Traitement hypertension	Variable binaire : oui/non
Traitement cholestérol	Variable binaire : oui/non
Traitement diabète	Variable binaire : oui/non
Antécédents d'insuffisance cardiaque	Variable binaire : oui/non
Antécédents de maladies cérébrovasculaires	Variable binaire : oui/non
Antécédents d'arythmies	Variable binaire : oui/non
Antécédents de coronaropathies	Variable binaire : oui/non
Variables sommeil	Variables continues

2.2. Données SNDS

Les événements cardiovasculaires indésirables majeurs (MACE), principal résultat composite de l'étude, ont été définis comme la première hospitalisation due à un infarctus du myocarde, un AVC, une insuffisance cardiaque congestive, une procédure de revascularisation (intervention coronarienne percutanée, pontage aortocoronarien) ou un décès toutes causes confondues. La première incidence est définie à tout moment entre l'enregistrement du sommeil et la date de suivi final du 31 décembre 2019. Les patients ayant déjà rencontré ces évènements cardiovasculaires, avant l'enregistrement du sommeil, ont été exclus de l'étude.

3. Méthodes

3.1. Modèle d'apprentissage automatique : AdaBoost

Un premier modèle a été développé, basé sur des arbres de classification d'une seule séparation : AdaBoost. Un poids initial est attribué à chaque patient. L'erreur de classification de l'arbre est calculée par l'indice de Gini. Un nouveau poids est calculé pour chaque patient en fonction de son erreur de classification. Comme le modèle est basé sur le principe du boosting, chaque arbre est

construit en fonction de l'erreur de classification des patients de l'arbre précédent. La probabilité finale d'appartenance à une classe est établie en fonction de la décision de chaque arbre. Les codes sources d'AdaBoost sont disponibles sous forme de package open source Python : https://github.com/scikit-learn.

Pour former et évaluer le modèle, l'ensemble de données a été séparé de manière aléatoire et homogène en ensembles de données d'entrainement (deux tiers) et de test (un tiers). Pour évaluer le modèle, il a été entraîné trois fois pour obtenir un score moyen de l'AUC. Il y avait trois ensembles de données d'entraînement différents : deux tiers de l'ensemble de données d'entraînement ont été extraits de manière aléatoire à chaque itération pour entraîner le modèle. L'ensemble de données de test est resté le même pour évaluer le modèle trois fois.

Le modèle était composé de 1000 arbres avec un taux d'apprentissage de 0,01. La combinaison de neuf caractéristiques a été utilisée : le sexe, l'âge, la présence d'un traitement contre l'hypertension et le diabète, la pression artérielle systolique, l'index de désaturation en oxygène, la fréquence cardiaque moyenne, l'aire moyenne de la fréquence cardiaque et le rapport LF/HF. Au final, cinq caractéristiques cliniques et quatre caractéristiques d'oxymétrie nocturne ont été sélectionnées.

L'échantillon final de l'étude comprenait 5234 patients sans MACE au moment de l'étude diagnostique du sommeil. Après un suivi médian de 6,0 [3,5-8,5] ans, 426 patients ont reçu un diagnostic de MACE et 259 sont décédés.

3.2. Modèle d'apprentissage automatique profond : CNN

Pour utiliser directement les signaux issus de l'enregistrement du sommeil, un modèle basé sur un réseau de neurones à convolution (CNN) d'une dimension a été choisi. Une fonction de coût basée sur l'optimisation de l'indice de concordance (CI) a été utilisée pour obtenir le meilleur classement des patients en fonction de leur risque cardiovasculaire et de leur durée de suivi, dans la population.

Pour former et évaluer le modèle, l'ensemble de données a été séparé de manière aléatoire et homogène en ensembles de données d'entrainement (80%) et de test (20%). Une validation croisée, en 5 groupes, a été effectuée avec l'ensemble de données d'entraînement. Le meilleur modèle a été sélectionné au regard du meilleur résultat obtenu avec l'ensemble de données de validation, issu de l'ensemble d'entrainement lors de la validation croisée. Ensuite, l'ensemble des données de test a été utilisé pour calculer le CI et l'AUC. L'AUC a été calculée avec une sortie binaire (survenue ou non de MACE). Les deux mesures ont été calculées afin de pouvoir comparer facilement nos résultats avec ceux des autres études. Le modèle clinique et les modèles mixtes avec un signal en entrée ont utilisé 100 itérations. Le nombre d'itérations a été défini en fonction des courbes d'apprentissage. Dans cette étude, les modèles ont été écrits en Python avec la bibliothèque Tensorflow (https://www.tensorflow.org/) et la bibliothèque Scikit Survival a été utilisée pour construire la fonction de perte (https://github.com/sebp/scikit-survival).

La population finale comprenait 5 506 patients. Les patients ayant des données de sommeil non valides (temps d'enregistrement consécutif < 4 heures) et les patients ayant des antécédents d'AVC, d'insuffisance cardiaque et/ou des coronaropathies ont été retirés de l'étude. L'objectif était, encore, de prédire le risque de MACE. Après un suivi médian de 6,0 [3,9-8,7] ans, 383 patients ont reçu un diagnostic de maladies cardiovasculaires et 230 sont décédés, soit 613 cas de MACE.

Les signaux utilisés dans cette étude sont présentés dans le tableau ci-dessous. Ils sont issus de l'enregistrement de sommeil, et les signaux d'une durée de 90 minutes consécutives ont été utilisés dans les modèles.

Signaux	Fréquence	Normalisation	Signification
SpO2	1 Hz	Entre 0 et 1 en conservant la ligne de base	Taux d'oxygène dans le sang, dans le doigt enregistré par un oxymètre
Photopléthysmogramme (PPG)	2 Hz	Entre 0 et 1	Variations du volume sanguin dans le doigt enregistré par un oxymètre
Pression nasale	2 Hz	Entre 0 et 1	Flux d'air circulant par le nez, enregistré par une lunette nasale
Sons respiratoires	2 Hz	Entre 0 et 1	Bruits de la respiration entre 200 Hz et 2000 Hz, enregistrés par le capteur de CIDELEC (PneaVoX®)
Manifestation autonomique	0,5 Hz	Entre 0 et 1 en conservant la ligne de base	Activité autonomique comme une accélération du pouls, une vasoconstriction ou un mouvement

4. Résultats

4.1. Modèle d'apprentissage automatique : AdaBoost

Les performances du modèle AdaBoost ont été évaluées, une AUC moyenne de 0,78 a été atteinte. Pour obtenir des scores de sensibilité et de spécificité, un seuil de décision permettant de créer deux classes a été défini. L'indice de ROC, un compromis entre sensibilité et spécificité a été calculé. Les résultats sont présentés dans le tableau ci-dessous.

	Modèle AdaBoost
AUC	0,78
Sensibilité	73,5%
Spécificité	70,0%

4.2. Modèle d'apprentissage automatique profond : CNN

Pour les variables cliniques, l'architecture était composée de deux couches consécutives de 64 neurones. Pour les modèles basés sur des signaux, les modèles étaient composés de couches de neurones convolutifs consécutifs suivis d'une couche de neurones classiques. Par exemple, pour le signal de Manifestation Autonomique (MA), le modèle était composé de trois couches de convolution de 64, 32 et 16 neurones, suivies d'une couche de 64 neurones, permettant le traitement des caractéristiques extraites par convolution. Le modèle combinant le signal MA et les variables cliniques a atteint le meilleur CI de 0,779 et une AUC de 0,815. Les résultats sont présentés dans le tableau ci-dessous.

Entrées	**Architecture**	**CI**	**AUC**
Variables cliniques Signal SpO2	NN : 64 64 CNN : 16 16 16 16 NN : 64	0,769	0,805
Variables cliniques Signal Sons	NN : 64 64 CNN : 16 16 16 16 NN : 128	0,776	0,813
Variables cliniques Signal PPG	NN : 64 64 CNN : 8 16 32 64 NN : 64	0,767	0,802

Variables cliniques Signal Flux nasal	NN : 64 64 CNN : 8 8 8 8 NN : 64	0,772	0,803
Variables cliniques Signal de manifestation autonomique	NN : 64 64 CNN : 64 32 16 NN : 64	0,779	0,815

Des analyses en sous-groupe de genre ont également été réalisées, les résultats sont présentés dans le tableau ci-dessous. Les analyses par sous-groupes ont été réalisées avec les modèles ayant les mêmes architectures que ci-dessus. Le signal PPG semble plus pertinent chez les hommes, et le signal MA semble plus pertinent chez les femmes.

		Femmes	**Hommes**
Total		2166	3329
Variables cliniques Signal PPG	CI	0,781	**0,797**
	AUC	0,808	**0,825**
Variables cliniques Signal MA	CI	**0,790**	0,744
	AUC	**0,815**	0,773

5. Publications

Les publications (articles et congrès) ont été réalisées dans le cadre de la thèse, dans un objectif de recherche, répondant à la finalité de l'étude ERMES et non pas celle de PREDIVASC-IOC. Les méthodes et les populations décrites dans ces publications sont celles utilisées dans l'étude PREDIVASC-IOC. Mais les résultats peuvent être légèrement différents.

5.1. Article

- M. Blanchard et al., "Cardiovascular risk and mortality prediction in patients suspected of sleep apnea: a model based on an artificial intelligence system," Physiol. Meas., vol. 42, no. 10, p. 105010, Oct. 2021, doi : 10.1088/1361-6579/ac2a8f

5.2. Congrès

- M. Blanchard *et al.*, "Estimation du risque cardiovasculaire et de la mortalité chez les patients suspectés du syndrome d'apnées du sommeil : un modèle de machine learning," Médecine du Sommeil, vol. 19, no. 1, pp. 53–54, 2022, doi : https://doi.org/10.1016/j.msom.2022.01.006.
- EMBC 2022 (à venir) : "Deep Learning Based on Nocturnal Oximetry to Predict Cardiovascular Diseases and Mortality in Patients Suspected of Sleep Apnea", the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).

5.3. Manuscrit de thèse

Le manuscrit de thèse mentionne ce nouvel objectif PREDIVASC-IOC, ainsi que certains résultats. Mais le manuscrit est en grande partie basé sur les résultats obtenus dans le cadre de la finalité ERMES.