Physiological
Measurement

PAPER

Cardiovascular risk and mortality prediction in patients suspected of sleep apnea: a model based on an artificial intelligence system

To cite this article: Margaux Blanchard *et al* 2021 *Physiol. Meas.* **42** 105010

IOP Publishing Physiol. Meas. 42 (2021) 105010 https://doi.org/10.1088/1361-6579/ac2a8f

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
2 July 2021

REVISED
14 September 2021

ACCEPTED FOR PUBLICATION
27 September 2021

PUBLISHED
29 October 2021

PAPER

Cardiovascular risk and mortality prediction in patients suspected of sleep apnea: a model based on an artificial intelligence system

Margaux Blanchard[1,2], Mathieu Feuilloy[1,2], Chloé Gervès-Pinquié[3], Wojciech Trzepizur[4,5], Nicole Meslier[4,5], François Goupil[6], Thierry Pigeanne[7], Jean-Louis Racineux[3], Frédéric Balusson[8], Emmanuel Oger[8], Frédéric Gagnadoux[4,5], Jean-Marc Girault[1,2] on behalf of the ERMES study group

[1] ESEO, Angers, France
[2] LAUM, UMR CNRS 6613, Le Mans, France
[3] Pays de la Loire Respiratory Health Research Institute, Beaucouzé, France
[4] Department of Respiratory and Sleep Medicine, Angers University hospital, Angers, France
[5] INSERM Unit 1063, Angers, France
[6] Department of Respiratory Diseases, Le Mans General Hospital, Le Mans, France
[7] Respiratory Unit, Pôle santé des Olonnes, Olonne sur Mer, France
[8] Rennes University, Rennes University Hospital, EA 7449 [Pharmacoepidemiology and Health Services Research] REPERES, F5043 Rennes, France

E-mail: margaux.blanchard@eseo.fr

Keywords: sleep apnea, cardiovascular risk, machine learning, adaptive boosting

Abstract

Objective. Cardiovascular disease (CVD) is one of the leading causes of death worldwide. There are many CVD risk estimators but very few take into account sleep features. Moreover, they are rarely tested on patients investigated for obstructive sleep apnea (OSA). However, numerous studies have demonstrated that OSA index or sleep features are associated with CVD and mortality. The aim of this study is to propose a new simple CVD and mortality risk estimator for use in routine sleep testing. *Approach*. Data from a large multicenter cohort of CVD-free patients investigated for OSA were linked to the French Health System to identify new-onset CVD. Clinical features were collected and sleep features were extracted from sleep recordings. A machine-learning model based on trees, AdaBoost, was applied to estimate the CVD and mortality risk score. *Main results*. After a median [inter-quartile range] follow-up of 6.0 [3.5–8.5] years, 685 of 5234 patients had received a diagnosis of CVD or had died. Following a selection of features, from the original 30 features, 9 were selected, including five clinical and four sleep oximetry features. The final model included age, gender, hypertension, diabetes, systolic blood pressure, oxygen saturation and pulse rate variability (PRV) features. An area under the receiver operating characteristic curve (AUC) of 0.78 was reached. *Significance*. AdaBoost, an interpretable machine-learning model, was applied to predict 6 year CVD and mortality in patients investigated for clinical suspicion of OSA. A mixed set of simple clinical features, nocturnal hypoxemia and PRV features derived from single channel pulse oximetry were used.

1. Introduction

Cardiovascular disease (CVD) is one of the leading causes of death worldwide and it is estimated that it will cost $1044 billion by 2030 (Bloom *et al* 2011). Hence, the issue of prevention is critical. With nearly 1 billion adults affected worldwide, obstructive sleep apnea (OSA) is a highly prevalent disease characterized by recurrent episodes of complete or partial upper airway obstruction during sleep, leading to intermittent hypoxia (Benjafield *et al* 2019). Evidence from population- and clinic-based cohort studies supports a causal association between OSA and the incidence of CVD, including hypertension, coronary heart disease, arrhythmia, heart failure, and stroke (Drager *et al* 2017, Javaheri *et al* 2017).

OSA is very heterogeneous in its clinical presentation and the nature and frequency of its complications (Osman *et al* 2018). Assessing the risk of occurrence of a cardiovascular (CV) complication, in a patient

investigated for OSA, during the initial diagnosis, constitutes a major management issue for the clinician. Given that the CV risk associated with OSA is multifactorial, the evaluation of the CV prognosis requires taking into account many parameters related to the severity of the sleep-disordered breathing, the other CV risk factors (hypertension, diabetes, dyslipidaemia, etc) and their pharmacological management. Earlier research has extracted some sleep features from sleep recordings, and showed that they were associated with CVD and all-causes mortality (Azarbarzin *et al* 2019, Baumert *et al* 2019, Heinzer *et al* 2019). Most of these studies used univariate and multivariate Cox proportional-hazards models (Cox 1972) and features were studied in quantiles. For example, Blanchard *et al* (2021a) demonstrated that in patients investigated for clinical suspicion of OSA, indices of sleep apnea-related hypoxic burden and heart rate variability during sleep were associated with the incidence of a stroke. A recent study (Wallace *et al* 2021) presented a tree-based machine learning model to investigate the predictive value of each clinical and polysomnographic (PSG) features for predicting all-cause mortality. However, not all of these papers propose a risk estimator that combines all the features studied.

There are many CVD risk estimators, such as the Framingham risk score, QRISK2 and the European Society of Cardiology and the European Society of Hypertension (ESC/ESH) risk score, which predict the occurrence of a patient's fatal or non-fatal CVD. The risk factors generally studied included clinical and biological features such as age, gender, smoking status, cholesterol, systolic blood pressure (SBP), diabetes and hypertension. However, getting biological results can be time consuming and expensive. Moreover, integration of sleep parameters could be interesting to get a more adapted CVD risk. These risk estimators often use standard statistical methods such as the Cox proportional-hazards models (Cox 1972).

Today, the trend is more towards using new machine learning methods to overcome assumptions of classical statistical models. Machine learning is the part of artificial intelligence where computers use data to learn to correctly predict the endpoint of each sample. Several studies show that the accuracy to predict CVD was improved with a machine learning model (Weng *et al* 2017). Few studies have been performed to integrate clinical and sleep features in a machine-learning model to estimate a CVD risk score. Mazzotti *et al* (2019) have compared several supervised machine learning models on 3,674 individuals from the sleep heart health study (SHHS). The model with the highest performance to predict CVD risk was a logistic regression model. They combined 14 clinical and polygraphic features, but also used three biological features. Segura *et al* (2020) have tried many machine learning algorithms too, and combinations of features to predict CVD risk on the SHHS cohort: 3696 patients were followed up over 5.5 years, 866 patients developed CVD. The final model, based on a neural network, included four clinical features and one PSG feature (age, diabetes, reported coronary artery bypass graft, smoking status, percentage of time in stage one). However, a neural network, also called 'black box', is not well accepted in clinical practice due to an important limitation related to its interpretability (Watson *et al* 2019). For example, models based on trees are easier to explain and understand. Zhang *et al* (2020) used the eXtrem Gradient Boosting (XGB) model, an interpretable classification and regression trees (CARTs) based machine learning algorithm, to predict CVD outcomes in the SHHS database, with 11 clinical and 18 PSG sleep features. Specifically, features of heart rate variability during each sleep stage. PSG from 2111 patients were analyzed, participants were followed up for 11.8 years and 1252 patients developed a CVD. However, reading sleep stages to extract parameters can be time consuming.

The aim of this study is to develop an artificial intelligence model, for patients suspected of OSA, which allows the clinician to evaluate the risk of CV morbidity and all-causes mortality (MMCV) while diagnosing sleep disorders. The model should be interpretable with the smallest combination of easily accessible clinical and sleep features derived from sleep recordings. The model will be tree-based to make it easy to interpret. In order for it to be simple, fast and inexpensive to use in clinical routine, the model will not use biological and sleep stage features.

2. Methods

2.1. Study population

The study was conducted on 9876 patients of the *Pays de la Loire Sleep Cohort*, linked with data from the French administrative health care database (SNDS), which contains individualized, anonymous, and comprehensive data on health spending reimbursements (Fagot *et al* 2013, Justeau *et al* 2020, Blanchard *et al* 2021b).

All patients from *Pays de la Loire Sleep Cohort* investigated by PSG or respiratory polygraphy (RP) between 15 May 2007 and 31 December 2018, with available SNDS data were eligible for the present study. There were no available SNDS data for 3708 excluded patients. In this research, 916 patients had been diagnosed with a stroke, myocardial infraction or heart failure before the sleep study and were excluded. All patients had given their written, informed consent.

According to French guidelines, patients with a high clinical probability of OSA were investigated by RP (CID102LTM; CIDELEC, Sainte-Gemmes-sur-Loire, France) (Meurice and Gagnadoux 2010). Patients with a

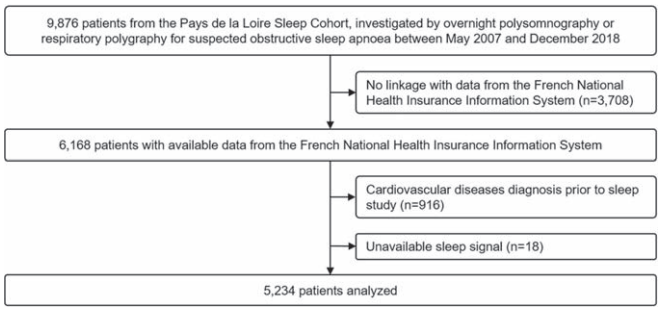

Figure 1. Flow diagram of the study subjects.

Table 1. Baseline characteristics of the study population.

	Total	No-MMCV	MMCV	*p*-values[a]
n	5234	4549	685	
Age, years	61 [50–70]	59 [49–68]	71 [63–79]	<0.0001
Gender, % female	37.43	39.24	25.40	<0.0001
Body mass index, kg m^{-2}	29 [25–34]	29 [25–33]	30 [27–34]	0.0009
Alcohol intake, g day^{-1}	0 [0–20]	0 [0–10]	10 [0–20]	<0.0001
Smoker, %	58.23	57.53	62.90	0.0115
Prevalent disease, %				
Diabetes	12.99	11.15	24.27	<0.0001
Hypertension	30.21	27.05	49.62	<0.0001
Cholesterol	21.05	18.48	37.05	<0.0001
AHI, events/h	19 [8–34]	18 [7–32]	28 [13–43]	<0.0001
SBP, mmHg	130 [120–140]	130 [120–140]	133 [121–143]	<0.0001
PAP adherent, %	35.79	34.54	44.09	<0.0001

Note. Data are expressed as median [interquartile range] or percentages.

[a] Patients No-MMCV versus MMCV; statistical significance was assessed using the nonparametric chi square for qualitative features, Mann–Whitney U test for quantitative features. Abbreviations: MMCV, cardiovascular morbidity and all-causes mortality.

low likelihood of OSA and/or coexisting sleep disorders received a PSG (CID102L8DTM, CIDELEC, Sainte-Gemmes-sur-Loire, France). Respiratory events were scored manually using recommended criteria (Berry *et al* 2012). Eighteen patients had unavailable data and were excluded.

According to the criteria defined by the French national health insurance, positive airway pressure (PAP) therapy was prescribed in patients with severe OSA and in those with moderate OSA and CV comorbidities or severe daytime sleepiness (Meurice and Gagnadoux 2010). The average PAP adherence during the follow-up period was calculated to determine treatment status. Patients with objectively measured PAP adherence $\geqslant$4 h per night during follow-up were considered to be adherent to the treatment. Patients with no therapy or objectively measured PAP adherence <4 h per night during follow-up were considered to be untreated.

The final study sample comprised 5234 CVD-free patients at the time of the diagnostic sleep study (figure 1). As shown in table 1, the study population consisted of typical OSA patients, predominantly male, obese or overweight, frequently presenting CV and metabolic comorbidities. Among the 5234 patients included, 16.8% had no OSA, 24.8% had mild OSA, 26.9% had moderate OSA, and 31.4% had severe OSA. During follow-up, 35.79% of patients were PAP adherent with objective PAP adherence $\geqslant$4 h per day.

2.2. Endpoint

The study endpoint is the first incidence of MMCV (including cerebrovascular disease, coronaropathy, and heart failure). The first incidence is defined at any time between the sleep study and the final follow-up date of 31 December 2019 in patients free of CVD at baseline. The algorithm to identify the first incidence of MMCV is based on data from the SNDS, which anonymously links information for all health care insurance reimbursement claims to the national hospital discharge database (PMSI).

The PMSI provides detailed medical information about all public and private hospitalizations, including main and associated discharge diagnosis ICD-10 codes. In French hospitals, all physicians routinely contribute to PMSI data collection, with annual quality control of coding resulting in a decreased risk of misclassification

bias (Fagot *et al* 2013). Patients who did not develop CVD, or did not die, were censored at the final follow-up date. The first incidence of a stroke in an inpatient was defined as the date of entry of the first hospitalization with a primary discharge diagnosis of G45, G46, I60-I64, or I69; or the date of entry of the first hospitalization with the same associated diagnosis. For patients identified in both streams, the earlier of the two dates was selected as the first incidence of a stroke. The first incidence of HF in an inpatient was defined as the date of entry of the first hospitalization with a primary discharge diagnosis of I50, I110 or I13; or the date of entry of the first hospitalization with the same associated diagnosis. For patients identified in both streams, the earlier of the two dates was selected as the first incidence of HF. The first incidence of coronaropathy in an inpatient was defined as the date of entry of the first hospitalization with a primary discharge diagnosis of I20, I21, I22 or I23; or the date of entry of the first hospitalization with the same associated diagnosis: or the first coronary revascularization intervention: angioplasty or bypass surgery. For patients identified in all three streams, the earliest of the three dates was used as the first incidence of coronaropathy.

After a median follow-up of 6.0 [3.5–8.5] years, 426 patients had received a diagnosis of CVD, and 259 had died (incidence density rate $=$ 21.8 per 1000 person-years). The comparison of patients with and without incident MMCV showed significant differences (p-value $<$ 0.05) for age, gender, body mass index (BMI), alcohol intake, smoking status, medical history of diabetes, hypertension and cardiac disease, indices of OSA severity, SBP, and PAP adherence (table 1).

2.3. Extraction of features

Thirty features have been used according to related works (Redline *et al* 2010, Gami *et al* 2013, Javaheri *et al* 2016, Azarbarzin *et al* 2019, Baumert *et al* 2019, Heinzer *et al* 2019, Blanchard *et al* 2021b); nine clinical features and twenty-one features derived from sleep recordings. Clinical features included age, sex, BMI, alcohol consumption, smoking status, medical history of diabetes, hypertension and heart disease, and SBP. Sleep features were selected based on their predictive value for MMCV risk in previous studies. They include apnea severity, nocturnal desaturations, pulse rate variability (PRV), vasoconstriction and respiratory rate features. All features from PSG were calculated as RP features. Apnea features were computed from scored respiratory events and based on total recording time for both, PSG and RP.

Apnea features included the obstructive apnea-hypopnea index, central apnea index and mean event duration and mortality (Redline *et al* 2010, Javaheri *et al* 2016). Saturation features were extracted from the oximeter saturation signal such as the 3% oxygen desaturation indices (ODI), percentage of recording time with oxygen saturation (SaO_2) <90% (T90), lowest SaO_2, mean SaO_2, hypoxic burden (SASHB) and lung to finger oxygen circulation time (Gami *et al* 2013, Azarbarzin *et al* 2019). A recent review concluded that PRV derived from photoplethysmography (PPG) technology provided an accurate estimation of ECG-derived heart rate variability in subjects at rest (Schäfer and Vagedes 2013). PRV signal processing analyses were performed from the PPG signal sampled at 64 Hz. The pulse-to-pulse intervals (PPI) were then calculated by detecting local peaks. Time series measures analyzing normal-to-normal (NN) beat intervals included the mean of the NN intervals, the root mean square of the successive NN differences, the proportion of adjacent NN intervals differing by more than 50 ms (pNN50) which reflects short-term PRV, and the standard deviation of NN intervals which reflects global PRV. Frequency domain features were extracted from the spectral analysis of the NN intervals. The normalized low frequency power (LF, from 0.04 to 0.15 Hz) and the normalized high frequency power (HF, from 0.15 to 0.4Hz) were calculated. The LF to HF (LF/HF) ratio was used as a measure of sympathetic/parasympathetic tone (Blanchard *et al* 2021b). The heart rate area was calculated from PPI signal envelopes linking peaks. Vasoconstriction and respiratory rate features included pulse wave amplitude drop index (Heinzer *et al* 2019) and a mean nocturnal respiratory rate greater than 16 breaths per minute (Baumert *et al* 2019) respectively.

All features were developed on MATLAB® (MathWorks, Inc., Natick, MA, USA). Missing values were imputed with a Python package using a multivariate imputation method for continuous features and by the most frequent value for categorical features (Buuren and Groothuis-Oudshoorn 2011). As machine-learning algorithms are sensitive to outliers, they were treated and filled with the nearest non-outlier value.

2.4. Machine learning algorithms

AdaBoost, also known as adaptive boosting, is a supervised machine learning algorithm (Freund and Schapire 1997). In a supervised model, an output is associated with each patient's feature vector. To create a new model, the computer learns to correctly classify each patient, based on its output and feature vector. To do this, it uses various mathematical and statistical operations. As described in figure 2, the AdaBoost model is based on trees, composed of a single division and called stumps. An initial weight is assigned to each sample. The classification error of the stump is calculated by the Gini index (Gini 1921). A new weight is calculated for each sample based on its classification error. As the model is based on the boosting principle, each stump is built

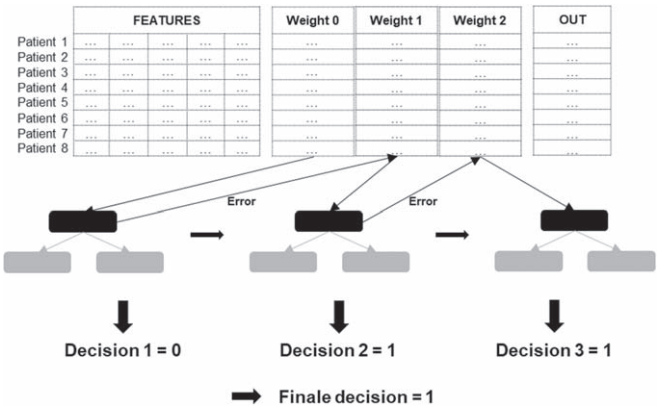

Figure 2. Schematic diagram of AdaBoost model. Patients and output were used in each stump to train model. At the first iteration, an initial weight is calculated (generally, 1 divided by the number of patients). First stump is constructed, its decision rules are saved and first classification error is calculated. A second weight is set for all patients and a new stump is constructed. And so on. The final decision is the majority decision.

according to the samples classification error of the previous stump. The final probability of belonging to a class is established according to the decision of each stump. The source codes of AdaBoost are available as a Python open source package: https://github.com/scikit-learn.

A selection of features was performed in order to simplify the model and avoid the risk of overfitting. The sequential floating forward selection (SFFS) (Somol *et al* 2010) method was executed to perform features selection, using an AdaBoost model with 5-fold cross validation. SFFS is a step-by-step features selection method. If the added feature improves the performance score, the feature is included in the set of the best features. If it results in less information than an alternative added feature, it is removed.

To train and evaluate the model, the dataset was randomly and homogeneously separated into training (two-thirds) and testing (one-third) datasets. To select the model hyperparameters, 3-fold cross validation was applied with the training dataset. To evaluate the model, it was trained three times to get a mean area under the receiver operating characteristic curve (AUC) score (Fawcett 2006). There were three different training datasets: two thirds of the training dataset were randomly extracted at each iteration to train the model. The test dataset remained the same to evaluate the model three times. Accuracy was not used because of its measurement bias on unbalanced data.

The AdaBoost model is a tree-based model, the importance of features could easily be extracted. The score is based on the total number of times a feature was used and split to build a stump. The more a feature has been used, the more important it is considered to be.

On another note, decision rules of each tree of the model can be extracted. Indeed, a decision rule defines a feature, a threshold and an output. Each decision rule was extracted from each stump. A median threshold was calculated for each continuous feature. Weights of the threshold were the number of stumps used divided by the total number of stumps; weights can be interpreted as a confidence index.

Moreover, age and gender subgroup analyses were performed to show which features were more or less informative for each group. The population was divided into four groups: women over 60, women under 60, men over 60 and men under 60 years of age. Clinical and sleep features were used: gender, age, hypertension, diabetes, SBP, ODI, mean heart rate, heart rate area, and LF/HF ratio. Four new AdaBoost models were trained.

Performance of the AdaBoost model was compared to the Framingham risk score on the cohort. The Framingham risk score is usually used with biological features, but a version exists with biology replaced by BMI. So, seven features were selected: gender, age, BMI, SBP, hypertension, smoking status and diabetes. To apply the Framingham risk score on the cohort, score sheets were used (D'Agostino *et al* 2008).

Performance of the AdaBoost model was also compared with the two other models that use clinical features and sleep features. (i) A feedforward neural network consists of several layers, composed of neurons and connected to each other with weights: an input layer, hidden layers and an output layer. The input layer is used to read the features, and the output layer to compute the final prediction. At the output of the model, an error is calculated between the final prediction and the real output, then the weights previously used are readjusted. This is called the gradient descent. And so on, until the gradient is minimal, and thus the prediction is closest to reality (Bishop 1995). (ii) The XGB model is very similar to AdaBoost, it is a model of a set of trees that are created one after the other and trained to correct the errors of the previous ones. But unlike AdaBoost, trees can be deeper and are not all the same depth. The first tree is built from the average of the observations, and the objective of the

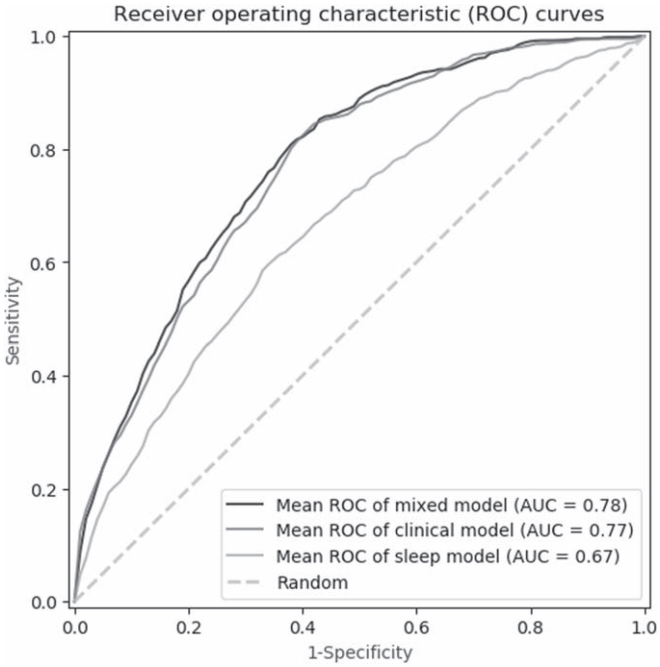

Figure 3. Mean ROC curves of the mixed model, clinical model and sleep model to predict cardiovascular morbidity and all-causes mortality risk. The mixed model included 9 features: gender, age, hypertension, diabetes, SBP, ODI, mean HR, HR area, and LF/HF ratio. The clinical model included 5 features: gender, age, hypertension, diabetes and SBP. The sleep model included 4 features: ODI, mean HR, HR area, and LF/HF ratio. Abbreviations: HR, heart rate; LF/HF, low frequency/high frequency; SBP, systolic blood pressure; ODI, oxygen desaturation index.

Table 2. Results of mixed model trained with clinical and sleep features, clinical model trained with clinical features and sleep model trained with sleep features.

	Mixed model	Clinical model	Sleep model
AUC	0.78	0.77	0.67
Sensitivity	73.5%	84.8%	63.2%
Specificity	70.0%	60.3%	65.2%

following trees is to get further and further away from this average in order to get closer to reality. This deviation is called residuals; it is the particularity of XGB (Chen and He 2021).

3. Results

3.1. AdaBoost scores

Hyperparameters of the AdaBoost model were calculated, the model was composed of 1000 stumps with a learning rate of 0.01. The features selection method was executed; the highest train score of AUC was 0.78, with the best features combination of nine features: gender, age, hypertension, diabetes, SBP, ODI, mean heart rate, heart rate area, and LF/HF ratio. In the end, these five clinical features and four nocturnal oximetry features were selected.

Performances of the model were evaluated, a mean AUC of 0.78 was reached. A clinical model was also trained and tested; it included only clinical features: gender, age, hypertension, diabetes and SBP. A mean AUC of 0.77 was reached. A sleep model was additionally trained and tested; it included ODI, mean heart rate, heart rate area and LF/HF ratio. A mean AUC of 0.67 was reached. In figure 3, mean ROC curves of the mixed model, clinical model and sleep model are presented. To get sensitivity and specificity scores, a decision threshold to create two classes was defined. The ROC index (Desquilbet 2020), a compromise between sensitivity and specificity was calculated. Results are presented in table 2.

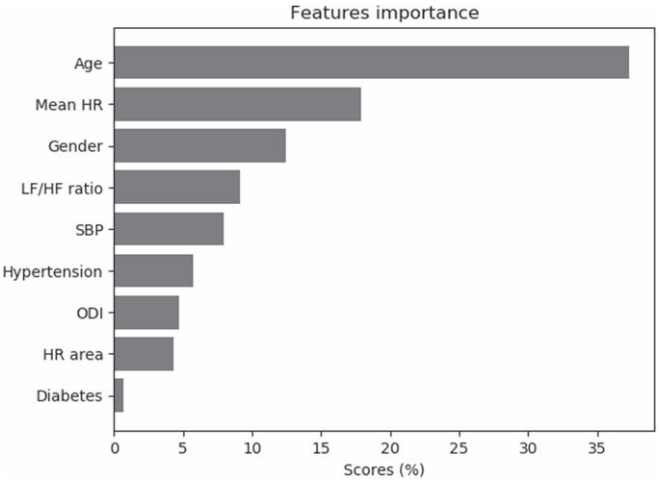

Figure 4. The importance of features used in the final mixed AdaBoost model. Scores are calculated as the number of trees using this feature. Abbreviations: HR, heart rate; LF/HF, low frequency/high frequency; SBP, systolic blood pressure; ODI, oxygen desaturation index.

Table 3. Details of model-derived decision thresholds for cardiovascular morbidity and all-causes mortality risk estimation. Each tree splits one feature into two parts according to a threshold. Threshold of trees can be different; results are presented as median and standard deviation. Patients with features in the category or under/over thresholds are classified as being at risk. Abbreviations: HR, heart rate; LF/HF, low frequency/high frequency; SBP, systolic blood pressure; ODI, oxygen desaturation index.

Features	Thresholds	Standard deviations	Weights
Gender	Men	—	0.124
Hypertension	Yes	—	0.057
Diabetics	Yes	—	0.007
LF/HF ratio	$\leqslant$0.93	$\pm$0.24	0.091
HR area	$\leqslant$ 36.1	0	0.043
ODI	$\geqslant$64.2 desat./h.	$\pm$11.4 desat./h.	0.047
Age	$\leqslant$56.5	$\pm$8.8	0.156
	$\geqslant$74.5	$\pm$18.5	0.217
Mean HR	$\leqslant$869 ms	$\pm$119 ms	0.163
	$\geqslant$1349 ms	0 ms	0.016
SBP	$\leqslant$90.5 mmHg	0 mmHg	0.021
	$\geqslant$ 136.5 mmHg	$\pm$6.2 mmHg	0.058

3.2. Features of importance

The importance of the features was extracted from the model. As shown in figure 4, age was a predominant feature to estimate MMCV risk. Mean heart rate, a sleep feature, was in second place before gender. A second PRV feature, LF/HF ratio, was also at the top of the ranking. Next, SBP was more important than hypertension, ODI, and heart rate area. Finally, the least important feature was diabetes.

3.3. Decision rules

The decision rules were retrieved and summarized in table 3. Men, patients with hypertension and diabetics were more at risk, with weight of 0.124, 0.057 and 0.007 respectively. Patients with a ratio under 0.93, a heart rate area under 36.1 or more than 64.2 desaturations/hour were more at risk, with a weight of 0.091, 0.043 and 0.047 respectively. Patients under 56.5 or over 74.5 years of age were at risk, with a weight of 0.156 and 0.217. Patients with a mean heart rate under 869 ms or over 1349 ms were at risk, with a weight of 0.163 and 0.016. Finally, patients with SBP under 90.5 mmHg or over 136.5 mmHg were at risk, with a weight of 0.021 and 0.058 for the threshold, respectively.

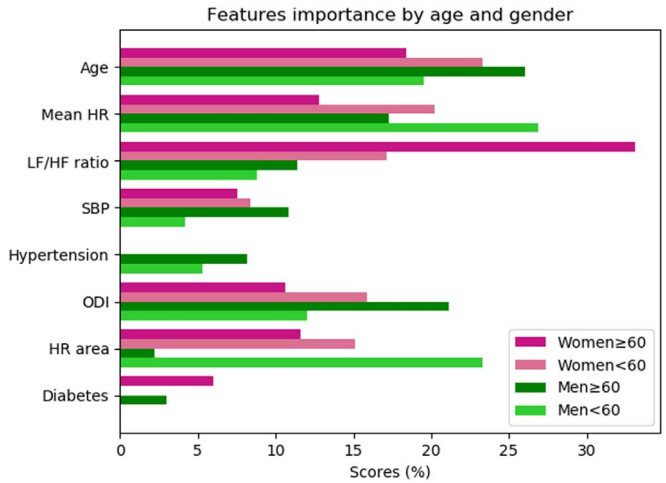

Figure 5. The importance of features used in subgroup mixed AdaBoost models. Four models were trained for each subgroup: women over 60 (n = 1039, incidence = 14%, AUC = 0.63), women under 60 (n = 920, incidence = 3%, AUC = 0.53), men over 60 (n = 1737, incidence = 25%, AUC = 0.65) and men under 60 years of age (n = 1538, incidence = 5%, AUC = 0.57). Scores are calculated as the number of trees using this feature. Abbreviations: HR, heart rate; LF/HF, low frequency/high frequency; SBP, systolic blood pressure; ODI, oxygen desaturation index.

3.4. Subgroup analyses

Age and gender subgroup analyses were performed. Results are presented in figure 5. Diabetes feature was not used to estimate MMCV risk in patients under 60 years of age. Hypertension feature was not used to estimate MMCV risk in women. In women over 60 years of age, the 4 sleep features were part of the 5 most important features, they represented 68% of the score, with in addition LF/HF ratio which was more important than age. In men under 60 years of age: the 4 sleep features represented 71% of the score, with in addition mean heart rate, heart rate area which were more important than age.

3.5. Other models

The Framingham risk score, without biology features, was calculated with the score sheets on the test dataset. An AUC score of 0.73 was reached.

XGB and neural network models were also tested on the cohort. Hyperparameters were selected with a 3-fold cross validation. The neural network model had four layers (1 input layer, 2 hidden layers, 1 output layer) and the XGB model had 100 trees. Results of both reached a mean AUC of 0.76.

4. Discussion

An interpretable machine learning model based on classification trees was presented in this study to predict 6 year MMCV risk. The model was trained with clinical and sleep features. In the final mixed model only 9 features were used, making the model robust. Features were simple, interpretable, easily accessible and none were biological. Indeed, sleep selected features can be collected with a nocturnal pulse oximeter only. Automated generation of hypoxemia and PRV features only required oximetry data, which are readily available from RP. An instant result after the night is therefore possible. In the clinical context, prevention is easier with a minimum of features. This means that less information is needed to make a decision, resulting in reduced costs.

Patients with OSA have a high MMCV risk. Nevertheless, OSA is very heterogeneous in its clinical phenotype, its comorbidities and its severity. Therefore, the level of MMCV risk is different within this population. There is a need for risk stratification in order to propose an adapted management of each patient. During the diagnosis of OSA, all the parameters that can contribute to the evaluation of the MMCV risk are taken into account. Sleep recordings also provide access to physiological measures of potential interest to participate in risk stratification. This study could allow estimation of MMCV risk at the end of the sleep test, which would allow better management of the patient's sleep and cardiac disorders. This risk estimator is specialized for patients with OSA. In line with previous reports, our study indicates that sleep medicine is moving toward applying precision medicine tools to patients with OSA in order to understand their MMCV risk. Such approaches are bound to enable the design of more rigorous clinical trials and more personalized treatment approaches for our patients (Zinchuk and Yaggi 2019).

The presented model reached a performance comparable to the other studies with fewer, and more accessible features. Mazzotti *et al* (2019) reached an AUC of 0.75 with 14 clinical, PSG and biological features; Zhang *et al* (2020) reached a sensitivity of 87.9 and a specificity of 57.0 with 29 clinical and PSG features; whereas our study reached an AUC of 0.78 with only 9 clinical and sleep oximetry features. A tree-based model was presented in this work, all decision rules were extracted and resumed (table 3). Segura *et al* (2020) reached good results (AUC $=$ 0.76) but their model was not as interpretable.

The chosen clinical features (section 2.3) are usually presented as potential MMCV risk factors in other publications. For example, age, which stands out most in the model, has already been shown to be a predominant factor in the estimation of MMCV risk (Zhang *et al* 2020). Usually, MMCV risk estimators are sex-specific, because the incidence of CVD in women seems to be lower than in men (Gao *et al* 2019). Some factors have a greater effect in men like age or hypertension. For women, other factors have a greater influence like smoking, diabetes and reproductive or pregnancy disorders. In the presented model, genders were not separated because of the low number of women and their low MMCV incidence in the data used.

SBP is almost systematically used in well-known estimators and it is directly associated with hypertension and CVD above specific thresholds (Flint *et al* 2019). In this study, SBP was the fifth most important feature. Comorbidities such as hypertension and diabetes did not appear to be of great importance in the model. However, they were selected among the 30 initial features. In Zhang's study (Zhang *et al* 2020) results are similar, especially for diabetes.

ODI is generally used as a measure of the diagnosis and severity of SAOS, but its clinical role and relationship to MMCV are not yet well established. One study noted a role in the measurement of intermittent hypoxia and autonomous system deregulation, and demonstrated that ODI was correlated with hypertension (Frangopoulos *et al* 2020).

In this study, we were interested in the impact of respiratory events on the autonomic nervous system. PRV is a standard non-invasive method for evaluating autonomic nervous system function. Moreover, sleep is a quiet period without too many artefacts to measure PRV. Night-time PRV analyses have revealed altered sympathetic/parasympathetic tone in subjects with moderate or severe OSA compared to healthy checks. The pronounced sympathetic activation that occurs towards the end of sleep-related obstructive events is accompanied by vagal mediated bradycardia due to activation of the diving reflex (Linz *et al* 2018). A high mean heart rate and lower PRV seem to be predictors of MMCV (Tsuji *et al* 1996).

Strengths of the model are, in addition to estimating a MMCV risk, the ability to find risk factors and to calculate risk thresholds for these factors (section 3.3). Usually, percentiles of features are used to divide them. They are not calculated according to the feature to be predicted and therefore such thresholds are not the most appropriate. However, the thresholds calculated by our model are to be interpreted according to the weight and their standard deviation. For example, the high threshold for the mean heart rate feature is not relevant because it had a weight of 0.016 against 0.163 for the low threshold. Moreover, for the age feature, the thresholds are difficult to use despite their heavy weight, because if we look at the standard deviations (8.8 years for the low threshold and 18.5 years for the high threshold, see table 3), the thresholds overlap. Further studies with this type of model could be conducted to find new risk factors and associated thresholds.

The algorithm's limitations were investigated, as well as the reasons why some patients were not detected by the algorithm. The following scores of the mixed model were: accuracy $=$ 70.5%, sensitivity $=$ 73.5%, specificity $=$ 70.0% (table 2). The difference between patients that the model classifies as being at risk and not at risk was studied: age is significantly different $p < 0.0001$. The incidence rate in patients under 60 years of age is 4.7% compared to 21.4% for those over 60 years of age. Younger at-risk patients seem to be less well detected but this hypothesis requires additional support.

The presented model, combining clinical and sleep features, had a total score of 0.78 (table 2). The model score fell to an AUC of 0.77 with clinical features alone and an AUC of 0.67 with sleep features alone. Thus, sleep features slightly improve the score, as in Zhang's paper, where the total accuracy was 75.3% and with clinical features, 73.7%. However, the feature selection stage retains four sleep features and as shown in figure 4 the mean heart rate is the second most important feature. By looking more closely at the results, the model based only on clinical features is very sensitive but not very specific (sensitivity $=$ 84.8%; specificity $=$ 60.3%). Therefore, the full model with clinical and sleep features gave a more balanced result between sensitivity and specificity. ROC curves of the clinical model and mixed model were compared with DeLong's test (DeLong *et al* 1988), a p-value of 0.056 was reached. Moreover, before features selection, a mixed model with the 30 initial features and a clinical model with the 9 initial clinical features were significantly different (p-value=0.031).

The relevance of adding sleep features was further assessed. To achieve this, new models were created by age and sex subgroups (figure 5). Some clinical features were not used, such as hypertension and diabetes in women and patients under 60 years of age, respectively. Sleep features seemed more important for women and patients under 60 years of age, they gathered around 70% of the scores whereas they gathered around 50% for men over 60 years of age. The model seemed to look for information in the sleep features to classify patients under 60 years

of age and women. Scores should be interpreted with caution, because of the low AUCs of the models (women $\geqslant$ 60 AUC = 0.63; women $<$ 60 AUC = 0.53; men $\geqslant$ 60 AUC = 0.65; men $<$ 60 AUC = 0.57), especially for patients under 60 years of age, where the incidence was insufficient to allow good computational power. Other studies are necessary to complete this analysis.

The proposed model was also compared to established risk estimators that do not use sleep features. The AUC for the Framingham risk score was 0.76 for men and 0.79 for women (D'Agostino *et al* 2008), 0.792 for men and 0.817 for women for the QRISK2 (Hippisley-Cox *et al* 2008), and between 0.71 and 0.84 for the ESC/ESH risk score (Conroy 2003). These estimators seem to have higher performances than the proposed model, however on our cohort the Framingham score obtained was 0.73 (section 3.5). The poorer Framingham result on our cohort may be due to the nature of our population, which is a specific population of OSA patients. Framingham usually does not include all-causes mortality in its endpoint, so all-causes mortality were removed to calculate another score only on CVD. An even lower AUC of 0.71 was achieved.

Another point that could bias the model is the unbalanced distribution of the data (incidence = 13%). To check this point, the Synthetic Minority Over-sampling Technique (SMOTE) algorithm from the imbalanced-learn Python library was used (Chawla *et al* 2002). SMOTE over-samples the minority (abnormal) class and under-samples the majority (normal) class. It selects one patient from the minority class and finds its k nearest minority class neighbors. Among these neighbors, one is randomly chosen, a line is drawn between the two-minority class, and a new synthetic minority class is created on this line. An AUC of 0.78 was still achieved, with an incidence rate of 43% in a train population of 3551 patients. Test dataset was still unchanged. Therefore, a low incidence had no impact on the result.

The strength of the current study includes a multicenter design, a large sample size with a wide range of sleep-disordered breathing severity, suggesting likely generalizability of the results, a long and complete follow-up with access to comprehensive data from the National Health Insurance Information System, and the assessment of different measures of hypoxia and PRV. The French Health Care database now covers 98.8% of the French population, over 66 million persons, from birth (or immigration) to death (or emigration) (Bezin *et al* 2017). The linkage between the clinical cohort's data and SNDS data makes the long-term follow-up easier with limited attrition bias, contrary to trials (Tuppin *et al* 2017). In French hospitals, all physicians routinely contribute to PMSI data collection, with annual quality control of coding resulting in a decreased risk of misclassification bias (Fagot *et al* 2013). Our study also has its limitations, because of its observational design; potential unmeasured confounding factors cannot be excluded.

5. Conclusion

AdaBoost, an interpretable machine-learning model, based on simple trees, was applied to predict 6 year MMCV in patients investigated for clinical suspicion of OSA. A mixed set of simple clinical features, nocturnal hypoxemia and PRV features derived from single channel pulse oximetry were used. This study demonstrated that sleep parameters should also be taken into consideration to provide more accurate estimation of MMCV risk especially in patients under 60 years of age and in women. The main advantage of this approach is that the presented model can be a simple tool to use in clinical routine at end of sleep testing to help the clinician with management issues. Further studies should be conducted to find new sleep-related risk factors that could improve results.

Acknowledgments

We thank the French National Health Insurance for giving us access to the French administrative health care database.

We thank the ERMES study group: Centre Hospitalier Universitaire, Angers: Christine Person, Pascaline Priou; Centre Hospitalier, Le Mans: Olivier Molinier, Audrey Paris.

We thank Christelle Gosselin and Jean-Louis Racineux from the Institut de Recherche en Santé Respiratoire des Pays de la Loire.

We thank Julien Godey, Laetitia Moreno and Marion Vincent, sleep technicians in the Department of Respiratory and Sleep Medicine of Angers University Hospital.

This study was supported by a grant from the Pays de la Loire Respiratory Health Research Institute.

ORCID iDs

Margaux Blanchard https://orcid.org/0000-0001-8382-4037

Frédéric Gagnadoux https://orcid.org/0000-0002-4231-5102

References

Azarbarzin A *et al* 2019 The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the osteoporotic fractures in men study and the sleep heart health study *Eur. Heart J.* **40** 1149–57

Baumert M, Linz D, Stone K, McEvoy R D, Cummings S, Redline S, Mehra R and Immanuel S 2019 Mean nocturnal respiratory rate predicts cardiovascular and all-cause mortality in community-dwelling older men and women *Eur. Respir. J.* **54** 1802175

Benjafield A V *et al* 2019 Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis *Lancet Respir. Med.* **7** 687–98

Berry R B *et al* 2012 Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American academy of sleep medicine *J. Clin. Sleep Med.* **08** 597–619

Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P and Moore N 2017 The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology *Pharmacoepidemiol. Drug Saf.* **26** 954–62

Bishop C M 1995 *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press)

Blanchard M *et al* 2021a Hypoxic burden and heart rate variability predict stroke incidence in sleep apnoea *Eur. Respir. J.* **57** 2004022

Blanchard M *et al* 2021b Association of nocturnal hypoxemia and pulse rate variability with incident atrial fibrillation in patients investigated for OSA *Ann. Am. Thorac. Soc.* **18** 1043-1051

Bloom D E *et al* 2011 The Global Economic Burden of Noncommunicable Diseases *World Economic Forum*

Buuren S van and Groothuis-Oudshoorn K 2011 Mice : multivariate imputation by chained equations in R *J. Stat. Softw.* **45** 1-67

Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–57

Chen T and He T (2021) R package version 1.4.1.1 *xgboost: eXtreme Gradient Boosting*

Conroy R 2003 Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project *Eur. Heart J.* **24** 987–1003

Cox D R 1972 Regression models and life-tables *J. R. Stat. Soc.* B **34** 187–220

D'Agostino R B, Vasan R S, Pencina M J, Wolf P A, Cobain M, Massaro J M and Kannel W B 2008 General cardiovascular risk profile for use in primary care: the framingham heart study *Circulation* **117** 743–53

DeLong E R, DeLong D M and Clarke-Pearson D L 1988 Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach *Biometrics* **44** 837–45

Desquilbet L L 2020 Tutoriel sur les courbes ROC et leur création grâce au site Internet easyROC *hal-02870055* Ecole nationale vétérinaire d'Alfort

Drager L F, McEvoy R D, Barbe F, Lorenzi-Filho G and Redline S 2017 Sleep apnea and cardiovascular disease: lessons from recent trials and need for team science *Circulation* **136** 1840–50

Fagot J-P, Blotiere P-O, Ricordeau P, Weill A, Alla F and Allemand H 2013 Does insulin glargine increase the risk of cancer compared with other basal insulins? A french nationwide cohort study based on national administrative databases *Diabetes Care* **36** 294–301

Fawcett T 2006 An introduction to ROC analysis *Pattern Recognit. Lett.* **27** 861–74

Flint A C, Conell C, Ren X, Banki N M, Chan S L, Rao V A, Melles R B and Bhatt D L 2019 Effect of systolic and diastolic blood pressure on cardiovascular outcomes *New. Engl. J. Med.* **381** 243–51

Frangopoulos F, Nicolaou I, Zannetos S, Economou N-T, Adamide T and Trakada G 2020 Association between respiratory sleep indices and cardiovascular disease in sleep apnea—a community-based study in cyprus *J. Clin. Med.* **9** 2475

Freund Y and Schapire R E 1997 A decision-theoretic generalization of on-line learning and an application to boosting *J. Comput. Syst. Sci.* **55** 119–39

Gami A S, Olson E J, Shen W K, Wright R S, Ballman K V, Hodge D O, Herges R M, Howard D E and Somers V K 2013 Obstructive sleep apnea and the risk of sudden cardiac death: a longitudinal study of 10,701 adults *J. Am. Coll. Cardiol.* **62** 610–6

Gao Z, Chen Z, Sun A and Deng X 2019 Gender differences in cardiovascular disease *Med. Nov. Technol. Devices* **4** 100025

Gini C 1921 Measurement of inequality of incomes *Econ. J.* **31** 124–6

Heinzer R, Hirotsu C, Marques-Vidal P, Vollenweider P, Waeber G, Betta M, Bernardi G, Lecci S, Siclari F and Haba-Rubio J 2019 0859 Sleep determinants of incident cardiovascular events: a prospective population-based study *Sleep* **42** A344–5

Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A and Brindle P 2008 Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2 *Brit. Med. J.* **336** 1475–82

Javaheri S *et al* 2017 Types, mechanisms, and clinical cardiovascular consequences *J. Am. Coll. Cardiol.* **69** 841–58

Javaheri S, Blackwell T, Ancoli-Israel S, Ensrud K E, Stone K L, Redline S and Osteoporotic Fractures in Men Study Research Group 2016 Sleep-disordered breathing and incident heart failure in older men *Am. J. Respir. Crit. Care Med.* **193** 561–8

Justeau G *et al* 2020 Association between nocturnal hypoxemia and cancer incidence in patients investigated for OSA *Chest* **158** 2610–20

Linz D, McEvoy R D, Cowie M R, Somers V K, Nattel S, Lévy P, Kalman J M and Sanders P 2018 Associations of obstructive sleep apnea with atrial fibrillation and continuous positive airway pressure treatment: a review *JAMA Cardiol.* **3** 532–40

Mazzotti D R, Keenan B T, Urbanowicz R and Pack A I 2019 Evaluating supervised machine learning models for cardiovascular disease prediction using conventional risk factors, apnea-hypopnea index and epworth sleepiness scale *Sleep* **42** A334–334

Meurice J-C and Gagnadoux F 2010 Préambule *Rev. Mal. Respir.* **27** S113–4

Osman A M, Carter S G, Carberry J C and Eckert D J 2018 Obstructive sleep apnea: current perspectives *Nat. Sci. Sleep* **10** 21–34

Redline S *et al* 2010 Obstructive sleep apnea–hypopnea and incident stroke: the sleep heart health study *Am. J. Respir. Crit. Care Med.* **182** 269–77

Schäfer A and Vagedes J 2013 How accurate is pulse rate variability as an estimate of heart rate variability? *Int. J. Cardiol.* **166** 15–29

Segura M R, Nicolis O, Marquez B P and Carrillo Azocar J 2020 Predicting cardiovascular disease by combining optimal feature selection methods with machine learning *2020 39th Int. Conf. Chilean Computer Science Society (SCCC)* (Piscataway, NJ: IEEE) (https://doi.org/10.1109/SCCC51225.2020.9281168)

Somol P, Novovicova J and Pudil P 2010 Efficient feature subset selection and subset size optimization *Pattern Recognition Recent Advances* ed A Herout (London: IntechOpen) (https://doi.org/10.5772/9356)

Tsuji H, Larson M G, Venditti F J, Manders E S, Evans J C, Feldman C L and Levy D 1996 Impact of reduced heart rate variability on risk for cardiac events *Circulation* **94** 2850–5

Tuppin P *et al* 2017 Value of a national administrative database to guide public decisions: from the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France *Rev. DÉpidémiologie Santé Publique* **65** S149–67

Wallace M L, Coleman T S, Mentch L K, Buysse D J, Graves J L, Hagen E W, Hall M H, Stone K L, Redline S and Peppard P E 2021 Physiological sleep measures predict time to 15-year mortality in community adults: application of a novel machine learning framework *J. Sleep Res.* https://doi.org/10.1111/jsr.13386e13386 May 2021

Watson D S, Krutzinna J, Bruce I N, Griffiths C E, McInnes I B, Barnes M R and Floridi L 2019 Clinical applications of machine learning algorithms: beyond the black box *Brit. Med. J.* **364** l886

Weng S F, Reps J, Kai J, Garibaldi J M and Qureshi N 2017 Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* **12** e0174944

Zhang L, Wu H, Zhang X, Wei X, Hou F and Ma Y 2020 Sleep heart rate variability assists the automatic prediction of long-term cardiovascular outcomes *Sleep Med.* **67** 217–24

Zinchuk A and Yaggi H K 2019 Sleep apnea heterogeneity, phenotypes, and cardiovascular risk. implications for trial design and precision sleep medicine *Am. J. Respir. Crit. Care Med.* **200** 412–3