Machine learning and geometric morphometrics to predict obstructive sleep apnea from 3D craniofacial scans

Fabrice Monna[a*], Raoua Ben Messaoud[b*], Nicolas Navarro[c,d], Sébastien Baillieul[b,e], Lionel Sanchez[f], Corinne Loiodice[b,e], Renaud Tamisier[b,e], Marie Joyeux Faure[b,e§], Jean Louis Pepin[b,e§]

a- ARTEHIS, UMR CNRS 6298, Université de Bourgogne Franche-Comté, 6 boulevard Gabriel, Bât. Gabriel, F-21000 Dijon, France

b- HP2 Laboratory, Inserm U1300, Grenoble Alpes University, Grenoble, France

c- Biogéosciences UMR CNRS 6282, Université de Bourgogne Franche-Comté, 6, boulevard Gabriel, Bat. Gabriel, F-21000 Dijon, France

d- EPHE, PSL University, 4-14 rue Ferrus, F-75014 Paris, France

e- EFCR Laboratory, Thorax and Vessels division, Grenoble Alpes University Hospital, Grenoble, France

f- ARCTIC, 18 chemin Cadet, F-97411 Saint-Paul, France

*Co-first authors

§Co-last authors

Address for correspondence.

Jean-Louis Pépin

Laboratoire EFCR, CHU de Grenoble, CS 10217 38043 Grenoble – France

e-mail : jpepin@chu-grenoble.fr

Highlights

- Maxillofacial geometry is shown to be associated with the risk of sleep apnea
- Maxillofacial shape is first processed by geometric morphometrics
- Machine learning analysis gives better results than traditional questionnaires
- Digital medicine strategies for OSA diagnosis improve access to care

Abbreviations

AASM: American Academy of Sleep Medicine

AHI: Apnea-Hypopnea Index

auROC: area under the Receiver Operating Characteristic curve

BMI: Body Mass Index

GPA: Generalized Procrustes Analysis

HTA: Hypertension

ICSD: International Classification of Sleep Disorders

OSA: Obstructive Sleep Apnea

PCA: Principal Component Analysis

PCs: Principal components

PSG: Polysomnography

Abstract

Background. Obstructive sleep apnea (OSA) remains massively underdiagnosed, due to limited access to polysomnography (PSG), the highly complex gold standard for diagnosis. Performance scores in predicting OSA are evaluated for machine learning (ML) analysis applied to 3D maxillofacial shapes.

Methods. The 3D maxillofacial shapes were scanned on 280 Caucasian men with suspected OSA. All participants underwent single night in-home or in-laboratory sleep testing with PSG (Nox A1, Resmed, Australia), with concomitant 3D scanning (Sense v2, 3D systems corporation, USA). Anthropometric data, comorbidities, medication, BERLIN, and NoSAS questionnaires were also collected at baseline. The PSG recordings were manually scored at the reference sleep center. The 3D craniofacial scans were processed by geometric morphometrics, and 13 different supervised algorithms, varying from simple to more advanced, were trained and tested. Results for OSAS recognition by ML models were then compared with scores for specificity and sensitivity obtained using BERLIN and NoSAS questionnaires.

Results. All valid scans (*n*=267) were included in the analysis (patient mean age: 59±9 years; BMI: 27±4 kg/m^2). For PSG-derived AHI≥15 events/h, the 56% specificity obtained for ML analysis of 3D craniofacial shapes was higher than for the questionnaires (Berlin: 50%; NoSAS: 40%). A sensitivity of 80% was obtained using ML analysis, compared to nearly 90% for NoSAS and 61% for the BERLIN questionnaire. The auROC score was further improved when 3D geometric morphometrics were combined with patient anthropometrics (auROC=0.75).

Conclusion. The combination of 3D geometric morphometrics with ML is proposed as a rapid, efficient, and inexpensive screening tool for OSA.

Trial registration number: NCT03632382; Date of registration: 15-08-2018

Keywords. obstructive sleep apnea, craniofacial scan, machine learning, 3D geometric morphometrics

1. Introduction

Obstructive sleep apnea (OSA) is defined as recurrent episodes of upper airway obstruction during sleep [1]. Diagnosis of OSA is mainly based on respiratory indices, such as apneas and hypopneas, measured by full polysomnography (PSG). Considered as the gold standard, PSG remains a cumbersome diagnostic method, resulting in limited and therefore inequitable access to care. The PSG method requires scoring expertise, and data interpretation is time-consuming. Due to these limitations, the worldwide health system faces a challenging situation, with an OSA population of approximately 1 billion patients, and a still undiagnosed population, estimated at over 30 million in Europe alone [2-4]. The considerable social and economic impacts of OSA [5] create a pressing need to resolve this diagnostic bottleneck.

The pathophysiology of OSA results from anatomical upper airway narrowing, and from reduced pharyngeal dilator muscle activity during sleep [6, 7]. Specific craniofacial profiles have been identified as being associated with reduced upper airway size, so that patient morphology may be considered a relevant predictor for OSA [8-10]. Abnormal maxillofacial characteristics typically linked to OSA are a long face [11, 12], together with mandibular prognathism or retrognathism [12-16]. Imaging techniques including cephalometry [13, 16], computed tomography [17,18], magnetic resonance imaging [9,19-21], and digital photography have been developed to map these craniofacial structures [22]. Recently, 2D and 3D scans have been used to characterize maxillofacial structure [23-25]. It has been shown that 3D photography allows the assessment of facial characteristics as an alternative to MRI [26]. Another study confirmed that 3D photography is strongly correlated with 3D computed tomography (CT) [25]. With the emergence of artificial intelligence, several tools have been developed to improve OSA prediction [27, 28]. Machine learning (ML) and deep learning models have been used to identify OSA patients based on 2D and 3D photographs [28, 29].

Here, taking as reference the results obtained by PSG, the primary objective was to investigate the reliability of OSA diagnosis obtained by the 3D geometric morphometric analysis of maxillofacial scans, combined with ML analysis. The novelty of this study is that the dataset consists of the entire 3D surface, not only the frontal and profile images, or depth maps. Another originality is that a data-driven approach based on ML analysis can be performed in 10 minutes, economizing both time and resources. The secondary objective was to compare these results with the performance scores obtained with two questionnaires: the Berlin [30, 31], and the NoSAS [32, 33].

2. Materials and methods

2.1. Study Design

The EPISAS monocentric prospective study was conducted at the Grenoble Alpes University Hospital from 2018 to 2020. The study was registered on Clinicaltrials.gov (NCT03632382). Consecutive adult Caucasian men (age≥40) with suspected OSA, referred for gold standard PSG, were invited to participate. Exclusion criteria were previous maxillofacial interventions or dental malocclusion, body mass index (BMI) ≥35 kg/m^2, or a thick beard (impeding maxillofacial characterization). All participants signed written informed consent.

2.2. Data collected at inclusion

Patient anthropometric and comorbidity data were collected. Berlin and NoSAS questionnaires were also completed. Scores of at least two out of three for BERLIN [30, 31], and eight for NoSAS [32, 33], were considered as OSA predictors (e-Fig. 1 and e-Fig. 2).

2.3. Polysomnography (PSG)

Nocturnal in-home or in-lab PSG was performed with a Nox A1 polygraph (ResMed, Australia). Sleep measurements were recorded using sensors for airflow, respiratory effort, snoring, SaO2, eye and leg movements, chin electromyography (EMG), the electrical activity of the heart (ECG) and brain (EEG), following American Academy of Sleep Medicine (AASM) recommendations for good practices [34]. The PSG signals were manually scored by experts from the Grenoble Alpes University Hospital, France, following the criteria recommended by the AASM [35]. Apnea was defined as a complete cessation of airflow lasting 10 s or longer and was classified as obstructive, central, or mixed, depending on the presence or absence of respiratory effort. Hypopnea was scored using the AASM definition, requiring at least a 30% reduction in airflow lasting 10 s or longer, and associated with a decrease of at least 3% in oxygen saturation, as measured by pulse oximetry, or arousal [35, 36]. Diagnosis of OSA was established according to the International Classification of Sleep Disorders, 3rd edition [37]. The sleep apnea diagnostic threshold was set at 15 events/h.

2.4. 3D geometric morphometrics of the craniofacial and submandibular structure

2.4.1. Acquisition

To characterize properly the geometry of the neck and the submandibular area, all 3D scans were acquired at the Grenoble Alpes Hospital by the same clinical research assistant, who was specifically trained for the present study. A hand-held commercial scanner, Sense v2 (3D systems USA), was used to generate 3D maxillofacial models with a resolution and precision of 1-2 mm. Seven landmarks were established on the 3D models (Fig. 1). Four were easily identified: one at each earlobe (LM1, LM3), one at the nasal bridge (LM2), and one at the tip of the chin (LM4). As no obvious landmarks could systematically be identified to constrain the lower part of the area of interest, a colored target was placed on each acromioclavicular joint (LM5, LM7), with a final target on the sternal fork (LM6).

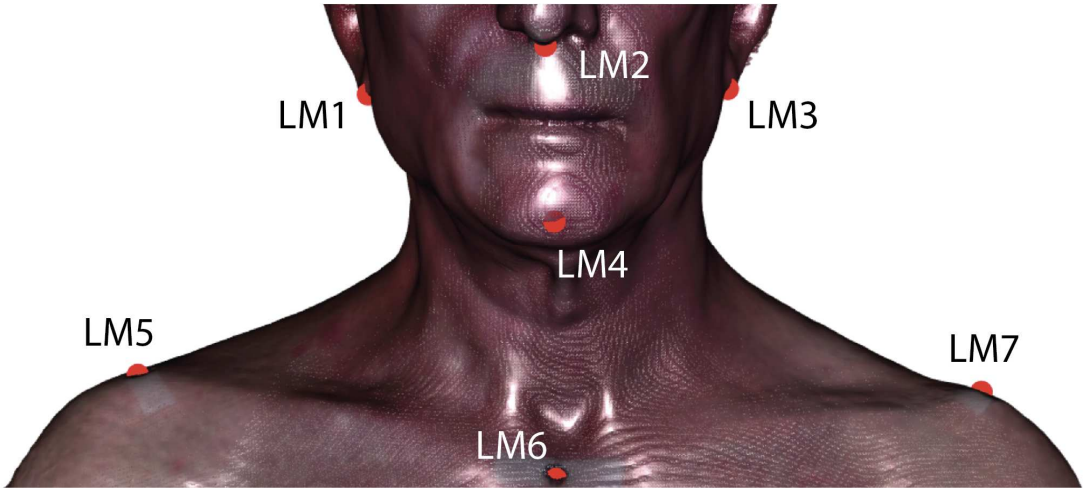

Figure 1: Position of the 7 landmarks on a typical 3D model. Colored targets are manually placed at LM5, LM6, and LM7 before scanning.

Between-subject consistency for head position was ensured using a mount equipped with bubble levels (Fig. 2a) to maintain the horizontality of the plane passing through the upper part of the ears and the eyes (Fig. 2b). A slot in the eyeglass frame was used to align the eyes (Fig. 2c).

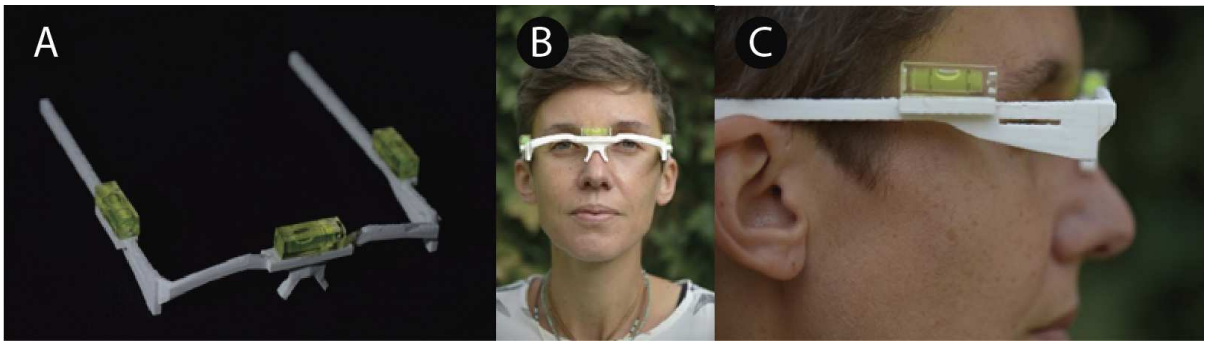

Figure 2: The prototype ensuring the horizontality of the reference plane on both axes. A: the mount with bubble levels; B: an example of use; C: the slot on the side of the frame allows the eyes to be aligned.

Each participant had to raise, lower, or incline his head until the horizontality of this plane was reached, and a neutral facial expression was required. At 18 different times during the entire period of data collection, a member of the technical staff was scanned to compare the

repeatability of the acquisition procedure with the naturally occurring variations observed within the patient cohort. The entire 3D scanning procedure lasted about 10 minutes.

2.4.2. Data preparation

The 3D scans were cleaned and repaired to eliminate any unreferenced vertices, non-manifold edges, and small disconnected parts. The resulting models were saved in PLY format, which is a common polygon file format describing 3D objects as a collection of vertices, faces, and other associated elements, such as normal direction, color, etc. Structurally, it encompasses a file header, the vertex, the face lists, and the attached elements [38]. The seven landmarks (LM1 to LM7, Fig. 1) were manually positioned on the textured meshes. Their 3D coordinates were then used to align all patient meshes in a common space, using a Generalized Procrustes Analysis (GPA) [39], applying bilateral symmetry for LM1 and LM3, the pair located under the ears, and for LM5 and LM7, the pair on the acromioclavicular joints [40]. Note that size is eliminated at this step. A mesh close to the mean shape was used to build the atlas of semi-landmarks (n=500) on a surface constrained by the seven landmarks (Fig. 3a). This atlas was projected onto the surface of the other meshes using the seven landmarks as reference (Fig. 3b), to capture soft tissue geometry as accurately as possible [41], in particular for the neck, where no clear landmarks are present. To favor homology, the semi-landmarks were allowed to slide, minimizing the total bending energy of the thin plate splines, before being reprojected onto the mesh surfaces [42]. A GPA was then applied to re-align all meshes, using not only the seven original landmarks but also the 500 slid semi-landmarks.

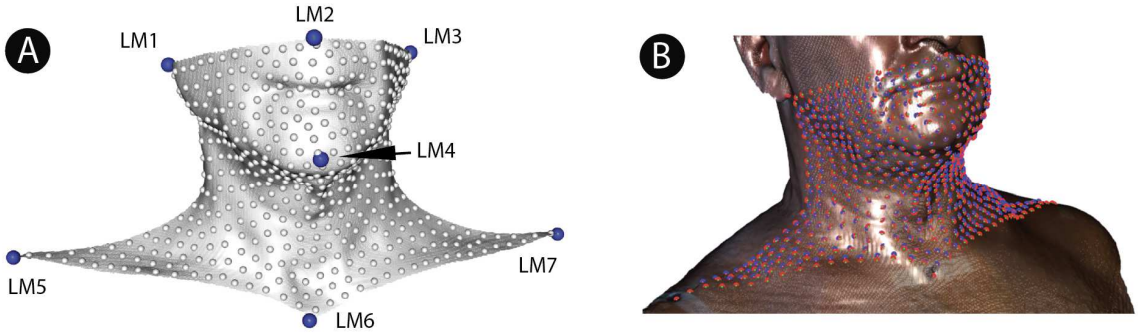

Figure 3: Atlas production and projection onto the mesh surfaces. A: restriction of the area of interest using the 7 landmarks, template for an atlas with 500 semi-landmarks; B: the atlas is projected onto all meshes (in red) and allowed to slide (in blue).

2.4.3. Dimensionality reduction

A total of 507 points was available to characterize each patient's maxillofacial shape. To reduce the number of variables, while maintaining a proper description of between-individual geometric variation, a principal component analysis (PCA) was computed on the Procrustes-aligned coordinates after their projection onto the space tangent to the mean shape. The first 3 principal components (PCs) expressed over two-thirds of the total variance, while 95% of the total variance was captured with only 20 PCs. In the following, several tests are presented to assess the smallest number of PCs required for optimal OSA prediction.

2.5. Machine learning algorithms and performance scores

2.5.1. Underlying idea

An AHI threshold of 15 events/h (measured by PSG) was used to define the presence or absence of OSA. The goal was to build a supervised mathematical model (i.e. a decision rule) where this binary condition (y_c, with $c \in [0,1]$ for negative OSA and positive OSA, respectively) must be predicted from $\mathbf{x}$, a vector corresponding to the m shape descriptors retained, expressed in the form of PCs, $\mathbf{x} = \{PC_1,..., PC_m\}$. A total of 11 classifiers was tested,

from the simplest to the most sophisticated: naive Bayes, linear and quadratic discriminant analyses (LDA & QDA), *k*-nearest neighbors (*k*-NN), support vector machine (SVM) with different kernel types (linear, polynomial, or using a radial basis function, namely rbf), extra trees, random forest (RF), artificial neural network (ANN), adaptative boosting (AdaBoost), and extreme gradient boosting (XGBoost). The underlying principles of these ML algorithms applied to binary classification can be found in many textbooks [43-45]. Note that other descriptors, such as anthropometric variables and symptoms, were later introduced into the analysis using XGBoost because this algorithm can also process categorical variables.

2.5.2. Model performance

A nested cross-validation method was used to evaluate model performance [46]. It divides the cohort into several parts (here 4), used for the outer loop. Sequentially, each of these parts is used for evaluation, with the remainder used for training. For training and hyperparameter tuning, the data are further divided into several parts (here also 4), used for the inner loop. All individuals are thus used alternately either for training or for evaluation, so that performance estimates are expected to be almost without bias [47]. The area under the receiver operating characteristic curve (auROC) was used as the main performance metric for classification, but specificity and sensitivity are also reported, for information.

2.6. Practical implantation

Morphometric data were prepared with the R v3.3 programming language (https://www.r-project.org/), using mainly the Morpho [48] and geomorph packages [49]. The ML phase was carried out with a homemade snippet programmed under Python 3.7 (https://www.python.org/), using the scikit-learn (https://scikit-learn.org) and XGBoost (https://xgboost.readthedocs.io/) libraries. Mesh cleaning was performed using the free Meshlab software (https://www.meshlab.net/).

3. Results

3.1. Study flow and population

Of the 1251 patients screened, only 280 were suitable for inclusion in the study (e-Fig. 3). The poor quality of some 3D scans (under 4%) reduced the study cohort to 267 participants with valid data. Table 1 presents the characteristics of the study population.

Table1. Description of the population at baseline (*n*=267). AHI: Apnea-Hypopnea Index; BMI: Body Mass index;

Variables	n (%) mean [min; max]
Sex: M	267 (100%)
Age (yr)	59.2 [40-75]
BMI (kg/m^2)	27 [18.3-35.1]
AHI (events/h)	23.7 [0.5-99.5]
Mallampati Class	
1	26 (9.7%)
2	96 (35.9%)
3	95 (35.5%)
4	50 (18.7%)
Neck circumference (cm)	40.3 [34-48]
Waist circumference (cm)	101.4 [71-182]
Hip circumference (cm)	101.9 [78-131]
AHI (events/h)	23.7 [0.5-99.3]
BERLIN score ≥2	153 (57.3%)
NoSAS score ≥8	63 (23.6%)
ESS	9 [0-24]
Comorbidities	
Hypertension	107 (40.1%)
Coronary heart disease	20 (7.5%)
Stroke or transient ischemic attacks	49 (18.4%)
Heart failure	23 (8.6%)
IDM	37 (13.8%)
Arrhythmia	24 (9%)
Other cardiovascular disease	10 (3.7%)
Cancer	10 (3.7%)
Type 1 diabetes	6 (2.2%)
Type 2 diabetes	35 (13.1%)
Renal failure	8 (3%)
Chronic Obstructive Pulmonary Disease (COPD)	13 (4.9%)
Treatments	
Diabetic drugs (A10)	46 (17.2%)
Cardiovascular system agents (C01)	14 (5.2%)
Anti-hypertensives (C02)	8 (3%)
Diuretics	39 (14.6%)

COPD: Chronic Obstructive Pulmonary Disease; IDM: myocardial infarction; ESS: Epworth Sleepiness Scale.

3.2. OSA prediction based on BERLIN and NoSAS questionnaires

A BERLIN score ≥ 2 corresponded to a sensitivity of 61% and a specificity of 50%, with a positive predictive value of 0.57. A NoSAS score ≥ 8 was associated with an auROC of 0.7, with a sensitivity of nearly 90%, a specificity of less than 40%, and a high positive predictive value of 0.77.

3.3. OSA prediction from morphometric data alone

3.3.1. Reproducibility of the procedure

The 18 replicates for the member of the technical staff are well clustered in the PCA morphospace, by comparison with the shape variation observed in the cohort (Fig. 4).

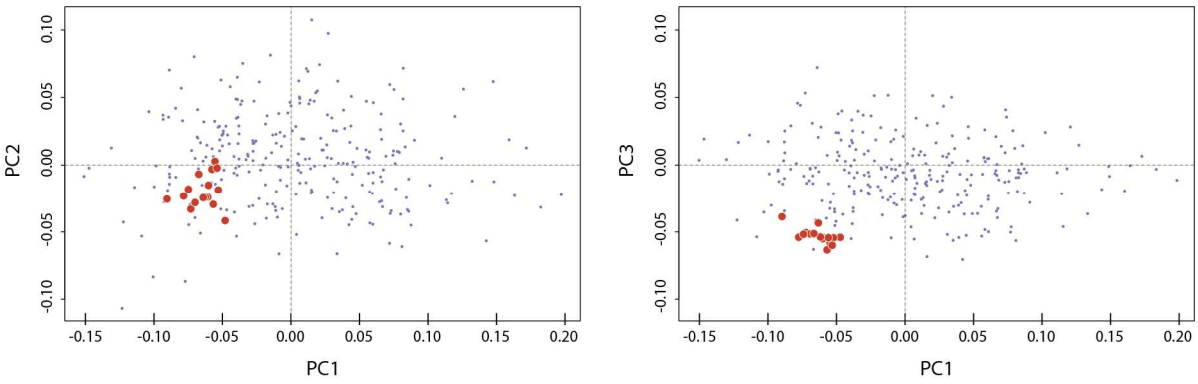

Figure 4: Projection onto the morphospace of the cohort (in blue) and of a member of the technical staff acquired 18 times (in red); left: PC2 *vs.* PC1; right: PC3 *vs.* PC1.

3.3.2. Influence of the number of PCs retained

A preliminary test involving LDA was undertaken with an increasing number of PCs (from 2 to 49) as input data. Since this classifier is linear by nature, it might not be optimal for our case study, but it is nevertheless considered sufficient to examine the discriminating power of the PCs. A simple Leave-One-Out Cross-Validation (LOOCV) procedure was applied for model evaluation to preserve the headcount. For up to 10 PCs (Fig. 5), the auROC was about 0.69 (e.g. with a sensitivity of 74%, and a specificity of 54% when 4 PCs were

processed). After that point, the auROC decreased as more PCs were added; this is a classic cost of dimensionality in classification tasks based on morphometric data.

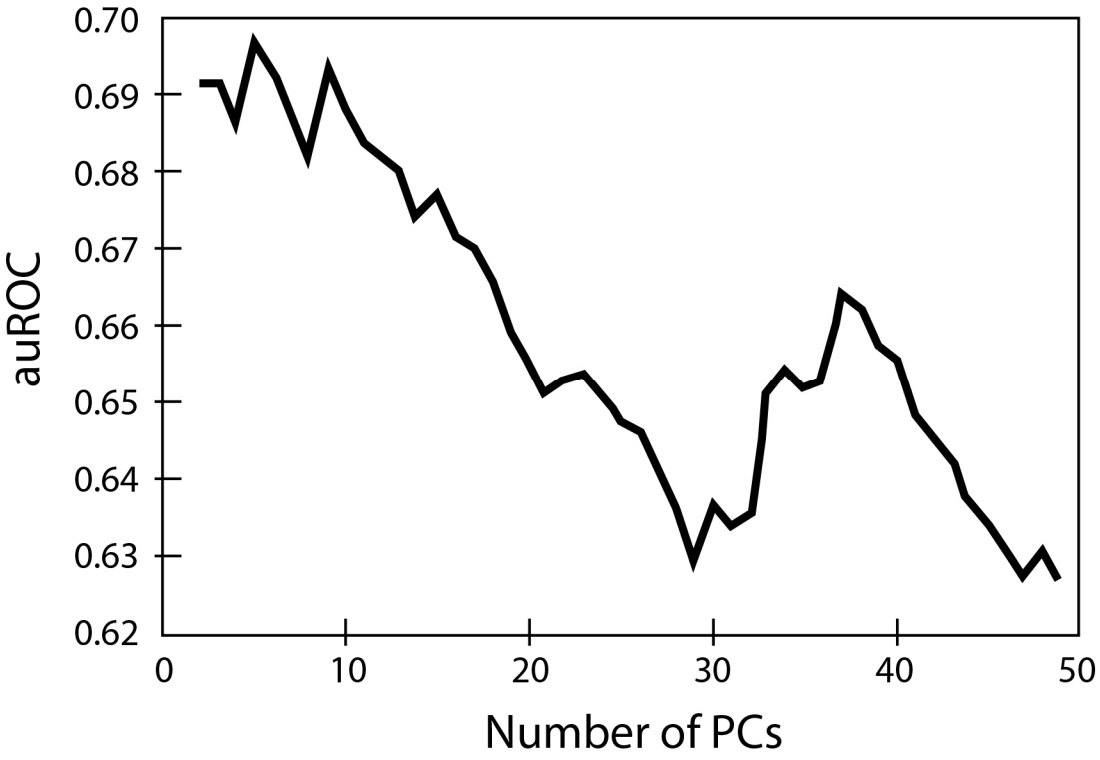

Figure 5: Influence of the number of PCs retained on the auROC score.

3.3.3. Algorithm testing

Further tests were carried out with the first 2 to 5 PCs, by applying 13 different supervised algorithms, varying from simple to more advanced ML techniques (Table 2). Given the cost of dimensionality with LDA and the higher complexity of most of the other ML algorithms tested, a parsimonious strategy concerning the number of feature inputs was adopted. Further experiments therefore took into account only 5 PCs at most. Whatever the model and the number of PCs retained, the auROC values were within the 0.62-0.70 range (i.e. always clearly better than 0.5, corresponding to random classification). In more detail, the ANN and RF algorithms probably suffer from a lack of data to be fully effective. The most efficient were LDA, Adaboost, extra trees classifier, XGBoost, and LR, with from 3 to 5 PCs. The LR classifier (including the first 3 PCs as feature inputs) was preferred at this step because it is

faster to compute. Scores obtained by nested cross-validation yielded an auROC of 0.70, with a sensitivity of 74%, and a specificity of 60% (Fig. 6).

	Number of PCs included			
	2	3	4	5
Naive Bayes	0.689	0.668	0.660	0.664
LDA	0.699	**0.703**	0.697	**0.703**
QDA	0.690	0.671	0.645	0.629
LR	0.699	**0.704**	0.697	**0.703**
k-NN	0.694	0.673	0.674	0.655
SVM				
rbf	0.690	0.651	0.655	0.617
poly.	0.699	0.669	0.639	0.615
linear	0.699	**0.701**	0.689	**0.702**
RF	0.669	0.680	0.671	0.662
Extra trees	0.694	**0.701**	0.688	0.673
ANN	0.648	0.645	0.671	**0.702**
Adaboost	0.699	**0.700**	**0.703**	0.695
XGboost	0.691	0.697	0.682	0.675

Table 2. The auROC evaluated by nested cross-validation for various ML models, and incorporating from 2 to 5 PCs as explanatory variables. In bold, the scores above 0.7. ANN: Artificial Neural Network; *k*-NN: *k*-nearest neighbors; LDA: Linear Discriminant Analysis; LR: Logistic Regression; poly.: polynomial; QDA: Quadratic Discriminant Analysis; rbf: Radial Basis Function; RF: Random Forest; SVM: Support Vector Machine

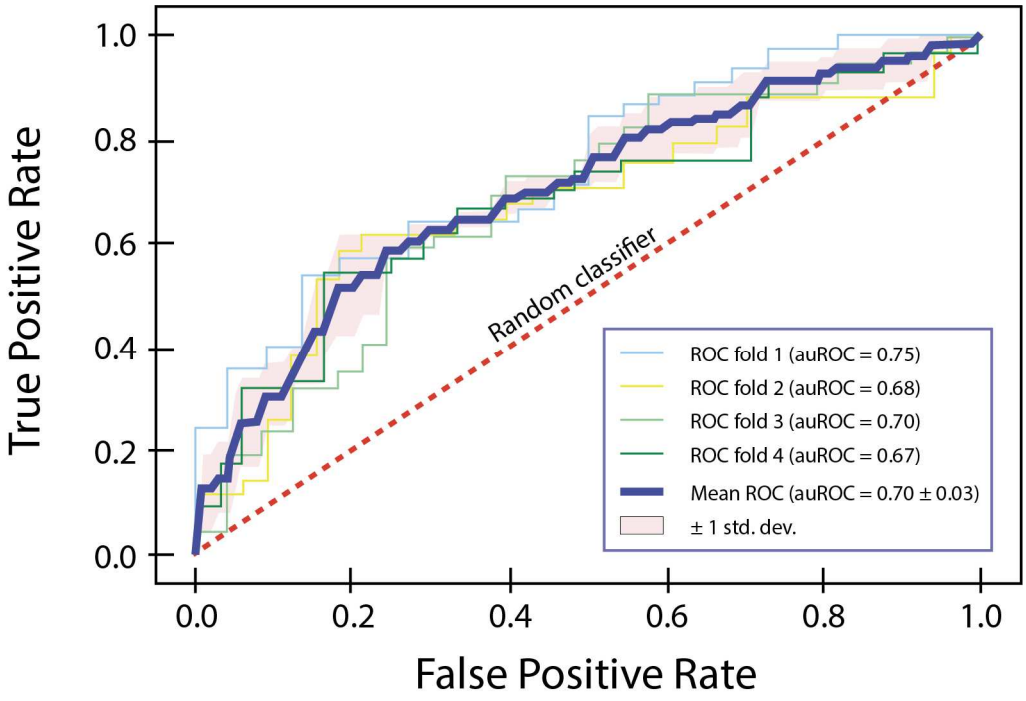

Figure 6: The ROC curve and associated 95% confidence interval for the LR classifier, with nested cross-validation, considering the first three PCs derived from the morphological data alone. The auROC scores are also reported.

3.4. OSA prediction from morphometric data, including questionnaires and anthropometric data

Using simple descriptive statistics (Student's t-test or Mann-Whitney U test, depending on the nature of the variables), ten anthropometric variables and symptoms were identified as significantly discriminating for OSA: hip, neck, and waist circumferences, age, BMI, Mallampati class, hypertension (HTA), witnessed apnea, and sleepiness while driving. These features were therefore processed together with the probability of belonging to the OSA risk group, provided by the LR model previously built from morphometric data alone. An XGBoost algorithm was applied at this step since it accepts both categorical and numerical variables as feature inputs. The inclusion of the above-mentioned variables in the model slightly boosted performance scores: the auROC reached 0.75, with a sensitivity of 80% and a specificity of 56% (Fig. 7).

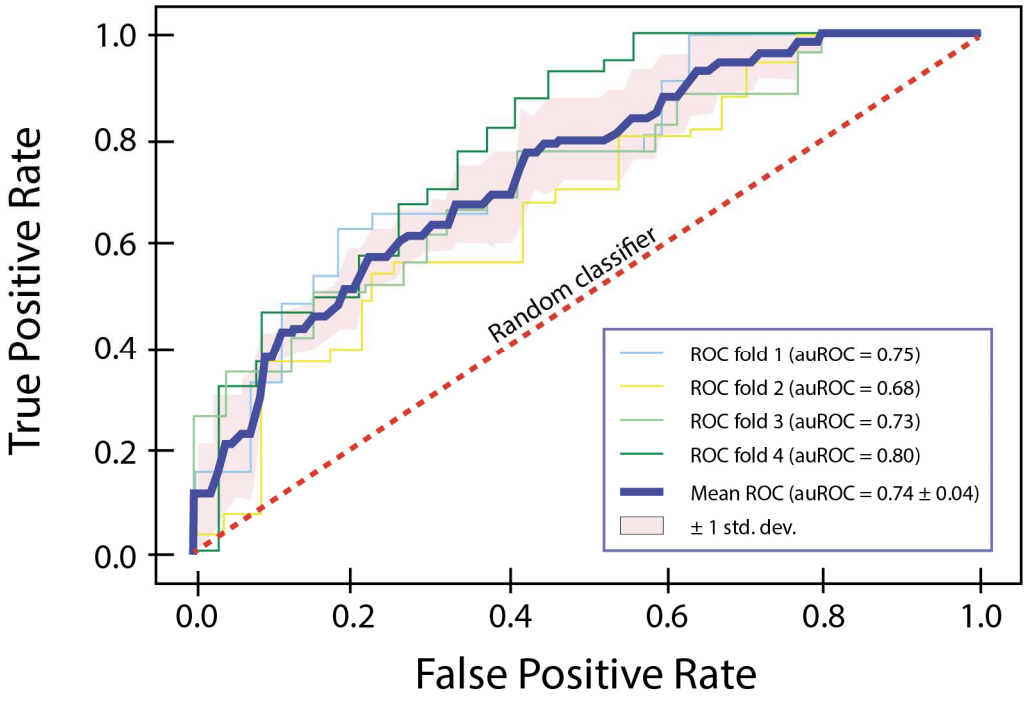

Figure 7: The ROC curve and associated 95% confidence interval for the XGBoost classifier, with nested cross-validation, considering the probabilities provided by the LR model together with a set of selected variables. The auROC scores are also reported.

3.5. Visual prediction of OSA

Shape differences between OSA and non-OSA groups were visualized using the linear discriminant function computed on 5 PCs. Predicted means for the two groups were unscaled and back-transformed on the coordinate scale. The reference mesh used to build the template was warped accordingly, using TPS. Shape changes predicted by the LDA (Fig. 8) show that, on average, people belonging to the OSA group have relatively shorter and thicker necks, together with stronger retrognathism than those in the non-OSA group. This result is in conformity with the classical physical characteristics of OSA patients observed by clinicians [7, 50].

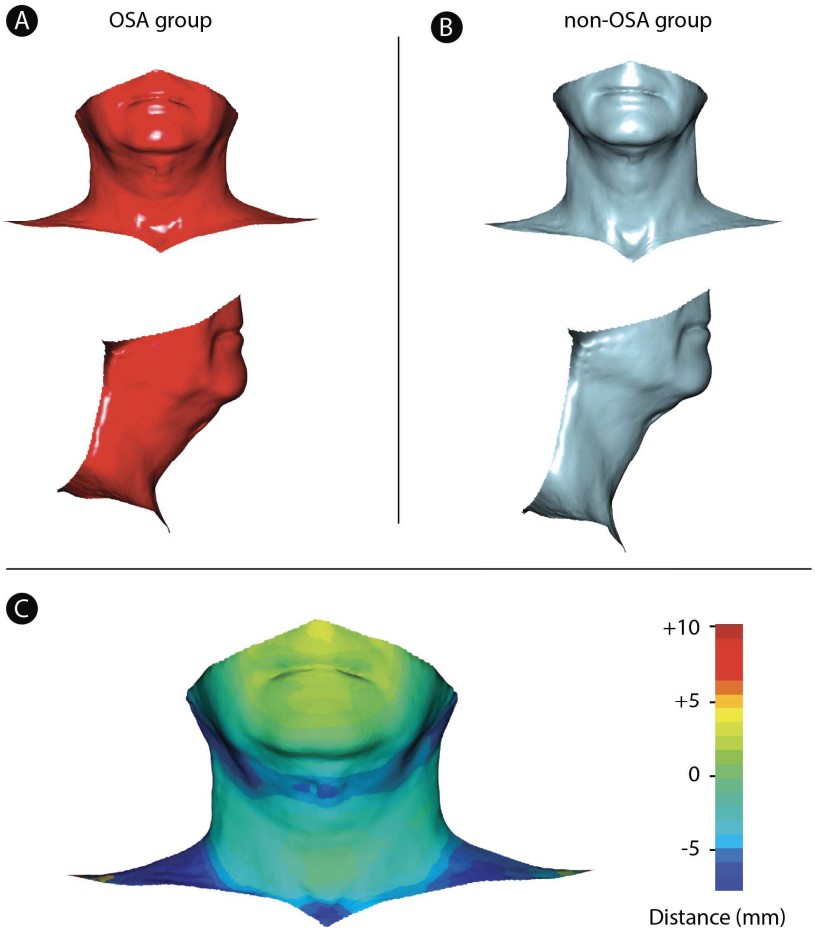

Figure 8: Predicted shape changes along the linear discriminant function between the OSA and non-OSA groups. A) Predicted shapes for OSA and B) Non-OSA groups. C) Colors represent the distances from the predicted OSA shape and the non-OSA shape.

4. Discussion

4.1. Overall evaluation

Several studies have previously assessed the performance of screening tools and scales, including the Epworth sleepiness scale (ESS), and the Berlin and NoSAS questionnaires [51-53]. Although these methods are widely used nowadays for OSA identification, their low specificity leads to a high burden of subsequent negative PSG [53-55]. The accuracy of questionnaires depends on population characteristics [56, 57]. In a recent meta-analysis evaluating the Berlin questionnaire in different settings (*n*= 8222), pooled specificity varied

from 33% to 47% [51], a range consistent with our findings. Interestingly, the present study demonstrates that 3D geometric morphometrics, combined with an appropriate ML algorithm, exhibits a predictive performance for OSA diagnosis similar to that obtained from questionnaires and data routinely collected in sleep centers. When a selected set of anthropometric characteristics and symptoms complement the 3D maxillofacial data, the performance scores surpass those of the BERLIN questionnaire and advantageously complete those of the NoSAS questionnaire. Both these traditional methods present too low a specificity to be fully operational. This new screening tool therefore possesses the potential to bypass the complexity of current OSA diagnostic procedures, thus improving access to care, and reducing medical misclassifications. Another strength of our diagnostic tool is its accessibility; 3D scanning can be conducted in a few minutes during the daytime, in different settings, and even at home, at low cost. Data acquisition does not require a high level of expertise, and the analysis is fully automated. Note that the entire process can easily be implemented as an end-to-end digital solution.

Existing studies have essentially been based on 2D analyses and/or photographs [58-60]. To the best of our knowledge, only one study has assessed anthropometrics combined with questionnaires and 3D scanning, reporting a sensitivity of 74% and a specificity of 63% [28]. Our study consistently improved performance, by 10 to 15%, by implementing 3D scans together with ML compared to BERLIN and NoSAS questionnaires. Future developments should explore the most effective combination of items to be included to diagnose different OSA phenotypes.

4.2. Perspectives and limitations

4.2.1. Potential sources of error ~~and their influence~~

It should be kept in mind that, as with any physical measurement, the 3D landmarks introduced in the ML algorithms are potentially flawed. These errors may occur during acquisition, mainly due to patient movements, but also due to the accuracy/resolution of the scanning device itself. Even without any patient movement, there may be homology defects

between patients. Finally, the manual placement by the operator of the seven landmarks is also subject to minor errors of interpretation. Nevertheless, the acquisition procedure appears fairly reproducible with respect to the variability observed within the cohort (Fig. 4). As a result, although all these sources of error coexist, any impact on the problem at hand should be minor.

4.2.2. The impact of established models

Special attention should be paid to the learning curve of the model established from shape data combined with anthropometric data and questionnaire responses (Fig. 9), which depicts both training and cross-validation scores (together with their 95% confidence interval) as a function of the size of the training dataset. Once 75 to 90 patients have been included during the learning phase, the training scores are high, with an auROC of about 0.95, whereas cross-validation scores are much lower, around 0.68.

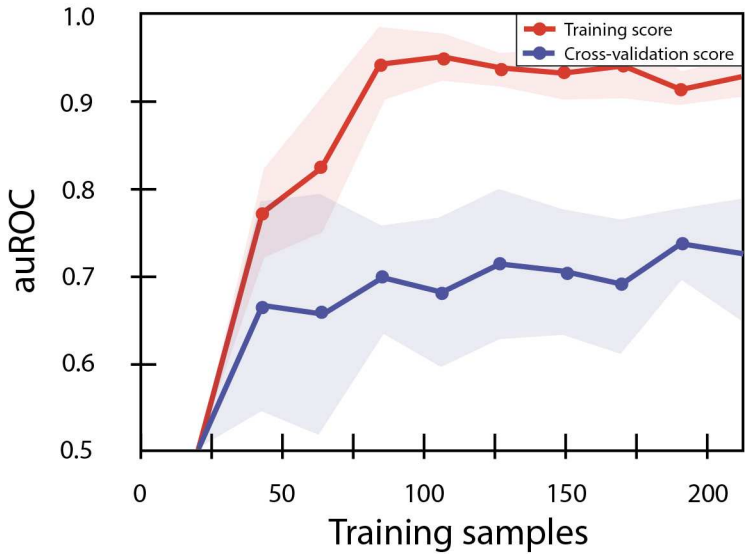

Figure 9: The learning curve of the XGBoost model established from morphological data combined with questionnaire responses, and anthropometric data; the training auROC score (in red) and cross-validation auROC score (in blue).

This result is not surprising, as the model overfits this relatively small learning dataset, while instances are lacking to produce high validation scores and appropriate generalization

capabilities for unseen instances. As expected, as the training set increases in size, the training score decreases (the model starts to underfit the dataset), while the validation score increases. Ultimately, both curves are supposed to converge when there are enough data to train the model optimally. From that point on, adding more data should no longer be beneficial, but better model performance might eventually be obtained by implementing new engineering features, or by building another, more complex model. This point of convergence may be associated with the limit of irreducible error, which is intrinsically linked to the problem at hand. What is important here is that the convergence point is currently far from being attained, suggesting great opportunities for improvement in the prediction of OSA risk. It is difficult to estimate the number of patients that would be required to reach this convergence point, but it is reasonable to assume that the inclusion of more than a thousand patients would be necessary. In any case, a learning curve obtained by the XGBoost model, established from morphological data combined with questionnaire responses and anthropometric data (Fig. 9) is slightly superior to the best models computed from the morphological data alone (e-Fig. 4 in the Supplementary Material), either because these later models quickly reached a limit auROC score of ca. 0.7 or less, or because the room for improvement when including more patients seems to be smaller.

As this study was limited to Caucasian men only, further studies should investigate differences in upper airway and craniofacial structures in relation to sex and ethnicity [61, 62]. As obesity is one of the main risk factors for OSA, due to fat deposits around the upper airways that narrow the airway during sleep [63], obese patients (BMI≥35 kg/m^2) were excluded from our study, so as to focus more specifically on maxillofacial characteristics. It would be interesting in future studies to include different BMI profiles.
A cohort suitable for such studies is realistic, despite its impressive size, as the epidemiology of sleep apnea concerns over one billion people worldwide.

5. **Conclusion**

The present study clearly demonstrates a link between maxillofacial geometry and the risk of sleep apnea. Although the cohort under study is large (almost 300 patients), it is not sufficient to encompass the entire range of maxillofacial shape diversity, so that model outputs will only be partially successful in predicting a syndrome as complex as sleep apnea. Nevertheless, the tool proposed in this study (combining the 3D geometry of patient scans processed by geometric morphometrics with machine learning) already presents a capacity for discrimination beyond that of the tools currently available (i.e. NoSAS and BERLIN questionnaires), an encouraging result on which to base further studies.

The informative morphometric data retained here is contained in the first 2-5 PCs (i.e. the overall geometry observable in Fig. 8). With a larger cohort, three further gains become possible. Substantial improvement should be observed in predictive performance, as suggested by the training curves. More PCs capturing finer details of maxillofacial shape can be included and their influence comprehensively evaluated. As the OSA / non-OSA groups are probably not linearly separable, more complex algorithms (such as artificial neural networks or random forests) are potentially better adapted, as greater quantities of data become available.

Clinicaltrials.gov:NCT03632382

Conflicts of interest/Competing interests

The authors declare a financial interest consisting of a patent application related to the material discussed in this manuscript.

Authors' contributions

FM: Conceptualisation, Methodology, Formal analysis, Writing Original Draft; RBM and MJF: Writing, Reviewing & Editing Original Draft, SB: Design, Reviewing & Editing Original Draft; NN: Methodology, Formal analysis and Reviewing Original Draft; LS: Technical and material development and support; CL: Administrative, and clinical support; RT: Design, Reviewing & Editing Original Draft; JLP: Funding acquisition, Conceptualisation, Design, Methodology, Supervision, Writing Original Draft

Ethical standards

This human study has been approved by an Independent Ethics Committee ("Comité de Protection des Personnes", Grenoble, France, ID-RCB: 2018-A00440-55) and has been performed in accordance with ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. The study was conducted in accordance with Good Clinical Practice, and all applicable laws and regulations following the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline. All the participants gave their informed consent prior to their inclusion in the study.

Acknowledgment

The authors thank Alison Foote (Grenoble Alpes University Hospital, France) and Carmela Chateau-Smith (CPTC, University of Burgundy, France) for editing the manuscript.

Funding

This work was supported by the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02) and the "e-health and integrated care and trajectories medicine and MIAI artificial intelligence" Chairs of excellence, from the C

This work was also supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003) and SATT Linksium.

References

[1]. Levy P, Kohler M, McNicholas WT, et al. Obstructive sleep apnoea syndrome. Nat Rev Dis Primers. 2015;1:15015.

[2]. Shokoueinejad M, Fernandez C, Carroll E, et al. Sleep apnea: a review of diagnostic sensors, algorithms, and therapies. Physiol Meas. 2017;38(9): R204-R52.

[3]. Sia CH, Hong Y, Tan LWL, van Dam RM, Lee CH, Tan A. Awareness and knowledge of obstructive sleep apnea among the general population. Sleep Med. 2017;36:10-7.

[4]. Peppard PE, Hagen EW. The Last 25 Years of Obstructive Sleep Apnea Epidemiology-and the Next 25? Am J Respir Crit Care Med. 2018;197(3):310-2.

[5]. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med. 2019;7(8):687-98.

[6]. Myers KA, Mrkobrada M, Simel DL. Does this patient have obstructive sleep apnea?: The Rational Clinical Examination systematic review. JAMA. 2013;310(7):731-41.

[7]. Neelapu BC, Kharbanda OP, Sardana HK, et al. Craniofacial and upper airway morphology in adult obstructive sleep apnea patients: A systematic review and meta-analysis of cephalometric studies. Sleep Med Rev. 2017;31:79-90.

[8]. Lee RW, Petocz P, Prvan T, Chan AS, Grunstein RR, Cistulli PA. Prediction of obstructive sleep apnea with craniofacial photographic analysis. Sleep. 2009;32(1):46-52.

[9]. Lee RW, Sutherland K, Chan AS, et al. Relationship between surface facial dimensions and upper airway structures in obstructive sleep apnea. Sleep. 2010;33(9):1249-54.

[10]. Di Francesco R, Monteiro R, Paulo ML, Buranello F, Imamura R. Craniofacial morphology and sleep apnea in children with obstructed upper airways: differences between genders. Sleep Med. 2012;13(6):616-20.

[11]. Ferguson KA, Ono T, Lowe AA, Ryan CF, Fleetham JA. The relationship between obesity and craniofacial structure in obstructive sleep apnea. Chest. 1995;108(2):375-81.

[12]. Kushida CA, Efron B, Guilleminault C. A predictive morphometric model for the obstructive sleep apnea syndrome. Ann Intern Med. 1997;127(8 Pt 1):581-7.

[13]. Guilleminault C, Riley R, Powell N. Obstructive sleep apnea and abnormal cephalometric measurements. Implications for treatment. Chest. 1984;86(5):793-4.

[14]. Tsai WH, Remmers JE, Brant R, Flemons WW, Davies J, Macarthur C. A decision rule for diagnostic testing in obstructive sleep apnea. Am J Respir Crit Care Med. 2003;167(10):1427-32.

[15]. Mayer P, Pepin JL, Bettega G, et al. Relationship between body mass index, age and upper airway measurements in snorers and sleep, apnoea patients. Eur Respir J. 1996;9(9):1801-9.
[16]. Jamieson A, Guilleminault C, Partinen M, Quera-Salva MA. Obstructive sleep apneic patients have craniomandibular abnormalities. Sleep. 1986;9(4):469-77.
[17]. Lowe AA, Fleetham JA, Adachi S, Ryan CF. Cephalometric and computed tomographic predictors of obstructive sleep apnea severity. Am J Orthod Dentofacial Orthop. 1995;107(6):589-95.
[18]. Ogawa T, Enciso R, Shintaku WH, Clark GT. Evaluation of cross-section airway configuration of obstructive sleep apnea. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2007;103(1):102-8.
[19]. Okubo M, Suzuki M, Horiuchi A, et al. Morphologic analyses of mandible and upper airway soft tissue by MRI of patients with obstructive sleep apnea hypopnea syndrome. Sleep. 2006;29(7):909-15.
[20]. Schwab RJ, Pasirstein M, Pierson R, et al. Identification of upper airway anatomic risk factors for obstructive sleep apnea with volumetric magnetic resonance imaging. Am J Respir Crit Care Med. 2003;168(5):522-30.
[21]. Welch KC, Foster GD, Ritter CT, et al. A novel volumetric magnetic resonance imaging paradigm to study upper airway anatomy. Sleep. 2002;25(5):532-42.
[22]. Schwab RJ, Leinwand SE, Bearn CB, et al. Digital Morphometrics: A New Upper Airway Phenotyping Paradigm in OSA. Chest. 2017;152(2):330-42.
[23]. Zinser MJ, Zachow S, Sailer HF. Bimaxillary 'rotation advancement' procedures in patients with obstructive sleep apnea: a 3-dimensional airway analysis of morphological changes. Int J Oral Maxillofac Surg. 2013;42(5):569-78.
[24]. Sutherland K, Schwab RJ, Maislin G, et al. Facial phenotyping by quantitative photography reflects craniofacial morphology measured on magnetic resonance imaging in Icelandic sleep apnea patients. Sleep 2014;37(5):959–968.
[25]. Lin S-W, Sutherland K, Liao Y-F, et al. Three-dimensional photography for the evaluation of facial profiles in obstructive sleep apnoea. Respirology 2018;23(6):618–625.
[26]. Kau CH, Richmond S, Incrapera A, et al. Three-dimensional surface acquisition systems for the study of facial morphology and their application to maxillofacial surgery. Int J Med Robot 2007;3(2):97–110.
[27]. Ryu S, Kim JH, Yu H, et al. Diagnosis of obstructive sleep apnea with prediction of flow characteristics according to airway morphology automatically extracted from medical images: Computational fluid dynamics and artificial intelligence approach. Comput Methods Programs Biomed. 2021;208:106243.

[28]. Hanif U, Leary E, Schneider L, et al. Estimation of Apnea-Hypopnea Index Using Deep Learning On 3-D Craniofacial Scans. IEEE J Biomed Health Inform 2021;25(11):4185–4194.

[29]. Tsuiki S, Nagaoka T, Fukuda T, et al. Machine learning for image-based detection of patients with obstructive sleep apnea: an exploratory study. Sleep Breath 2021;25(4):2297–2305.

[30]. Tan A, Yin JD, Tan LW, van Dam RM, Cheung YY, Lee CH. Using the Berlin Questionnaire to Predict Obstructive Sleep Apnea in the General Population. J Clin Sleep Med. 2017;13(3):427-32.

[31]. Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. Ann Intern Med. 1999;131(7):485-91.

[32]. Marti-Soler H, Hirotsu C, Marques-Vidal P, et al. The NoSAS score for screening of sleep-disordered breathing: a derivation and validation study. Lancet Respir Med. 2016;4(9):742-8.

[33]. Duarte RL, Magalhaes-da-Silveira FJ, Oliveira ESTS, Silva JA, Mello FC, Gozal D. Obstructive Sleep Apnea Screening with a 4-Item Instrument, Named GOAL Questionnaire: Development, Validation and Comparative Study with No-Apnea, STOP-Bang, and NoSAS. Nat Sci Sleep. 2020;12:57-67.

[34]. Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline. J Clin Sleep Med. 2017;13(3):479-504.

[35]. Berry RB, Budhiraja R, Gottlieb DJ, et al. 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. 2012;8(5):597-619.

[36]. Berry RB, Brooks R, Gamaldo C, et al. AASM Scoring Manual Updates for 2017 (Version 2.4). J Clin Sleep Med. 2017;13(5):665-6.

[37]. Sateia MJ. International classification of sleep disorders-third edition: highlights and modifications. Chest. 2014;146(5):1387-94.

[38]. Botsch M, Kobbelt L, Pauly M et al. Polygon Mesh Processing, A K Peters/CRC Press. 2011. 250 pp.

[39]. Dryden I, Mardia KV. Statistical shape analysis: with applications in R 2016;2nd Edition. John Wiley & Sons Ltd, Chichester, UK:384

[40]. Klingenberg CP, Barluenga M, Meyer A. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. Evolution. 2002;56(10):1909-20.

[41]. Schlager S. Morpho and Rvcg – Shape Analysis in R: R-Packages for Geometric Morphometrics, Shape Analysis and Surface Manipulations. In Zheng, G., Li, S., Székely, G. (eds) Statistical Shape and Deformation Analysis. 2017(Chapter 9):217-56.

[42]. Gunz P, Mitteroecker P, Bookstein FL. Semilandmarks in Three Dimensions. In: Slice D.E. (eds) Modern Morphometrics in Physical Anthropology. M A Boston and U S Springer. 2005;Gunz, P., Mitteroecker, P., and Bookstein, F.L. 2005. Semilandmarks in three dimensions. In Slice D.E. (ed) Modern Morphometrics in Physical Anthropology. Springer, Boston, MA. pp 73-98.:73-98.

[43]. Kung SY. Kernel Methods and Machine Learning. Cambridge University Press. 2014:572.

[44]. Bishop CM. Pattern Recognition and Machine Learning2006. Springer-Verlag NewYork Inc. 2006:738.

[45]. Lantz B. Machine Learning With R. Packt Publishing.2nd edition:452.

[46]. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform. 2014;6(1):10.

[47]. Raschka S. Python Machine Learning. Packt Publishing. 2015;2nd edition:454.

[48]. Schlager S. Soft-tissue reconstruction of the human nose: population differences and sexual dimorphism. 2013.

[49]. Adams DC, Otarola-Castillo E. Geomorph: an R package for the collection and analysis of geometric morphometric shape data. Methods Ecol Evol. 2013;4:393-9.

[50]. Wong ML, Sandham A, Ang PK, Wong DC, Tan WC, Huggare J. Craniofacial morphology, head posture, and nasal respiratory resistance in obstructive sleep apnoea: an inter-ethnic comparison. Eur J Orthod. 2005;27(1):91-7.

[51]. Bernhardt L, Brady EM, Freeman SC, et al. Diagnostic accuracy of screening questionnaires for obstructive sleep apnoea in adults in different clinical cohorts: a systematic review and meta-analysis. Sleep Breath. 2021.

[52]. Chiu HY, Chen PY, Chuang LP, et al. Diagnostic accuracy of the Berlin questionnaire, STOP-BANG, STOP, and Epworth sleepiness scale in detecting obstructive sleep apnea: A bivariate meta-analysis. Sleep Med Rev. 2017;36:57-70.

[53]. Senaratna CV, Perret JL, Matheson MC, et al. Validity of the Berlin questionnaire in detecting obstructive sleep apnea: A systematic review and meta-analysis. Sleep Med Rev. 2017;36:116-24.

[54]. de Menezes Junior LAA, Fajardo VC, do Nascimento Neto RM, et al. Diagnostic accuracy of the Berlin questionnaire and the NoSAS score in detecting risk for obstructive sleep apnea in rotating shift workers. Sleep Breath. 2021.

[55]. Herschmann S, Berger M, Haba-Rubio J, Heinzer R. Comparison of NoSAS score with Berlin and STOP-BANG scores for sleep apnea detection in a clinical sample. Sleep Med. 2021;79:113-6.
[56]. Oktay Arslan B, Ucar Hosgor ZZ, Orman MN. Which Screening Questionnaire is Best for Predicting Obstructive Sleep Apnea in the Sleep Clinic Population Considering Age, Gender, and Comorbidities? Turk Thorac J. 2020;21(6):383-9.
[57]. Giampa SQC, Pedrosa RP, Gonzaga CC, et al. Performance of NoSAS score versus Berlin questionnaire for screening obstructive sleep apnoea in patients with resistant hypertension. J Hum Hypertens. 2018;32(7):518-23.
[58]. Tabatabaei Balaei A, Sutherland K, Cistulli P, Chazal P de. Prediction of obstructive sleep apnea using facial landmarks. Physiol Meas 2018;39(9):094004.
[59]. Lee RWW, Chan ASL, Grunstein RR, Cistulli PA. Craniofacial phenotyping in obstructive sleep apnea-a novel quantitative photographic approach. Sleep 2009;32(1):37–45.
[60]. Eastwood P, Gilani SZ, McArdle N, et al. Predicting sleep apnea from three-dimensional face photography. J Clin Sleep Med 2020;16(4):493–502.
[61]. Xu L, Keenan BT, Wiemken AS, et al. Differences in three-dimensional upper airway anatomy between Asian and European patients with obstructive sleep apnea. Sleep 2020;43(5):zsz273.
[62]. Sutherland K, Keenan BT, Bittencourt L, et al. A Global Comparison of Anatomic Risk Factors and Their Relationship to Obstructive Sleep Apnea Severity in Clinical Samples. J Clin Sleep Med 2019;15(4):629–639.
[63]. Schwartz AR, Patil SP, Laffan AM, et al. Obesity and obstructive sleep apnea: pathogenic mechanisms and therapeutic approaches. Proc Am Thorac Soc 2008;5(2):185–192.