EClinicalMedicine 11 (2019) 81–88

Contents lists available at ScienceDirect

EClinicalMedicine

journal homepage: https://www.journals.elsevier.com/eclinicalmedicine

EClinicalMedicine
Published by THE LANCET

Feasibility of Single Channel Oximetry for Mass Screening of Obstructive Sleep Apnea

Joachim A. Behar [a,*,1], Niclas Palmius [b,1], Qiao Li [c], Silverio Garbuio [d], Fabìola P.G. Rizzatti [d], Lia Bittencourt [d,e], Sergio Tufik [d,2], Gari D. Clifford [c,2]

[a] Faculty of Biomedical Engineering, Technion, Israel Institute of Technology, Haifa, Israel
[b] Wolfson College, Oxford OX2 6UD, UK
[c] Departments of Biomedical Informatics & Biomedical Engineering, Emory University & Georgia Institute of Technology, Atlanta, GA, USA
[d] Departamento de Psicobiologia, Universidade Federal de São Paulo, São Paulo, Brazil
[e] Departamento de Medicina, Universidade Federal de São Carlos, São Carlos, Brazil

ARTICLE INFO

Article history:
Received 2 January 2019
Received in revised form 30 May 2019
Accepted 30 May 2019
Available online 7 June 2019

Keywords:
Obstructive sleep apnea screening
Oxygen saturation
Sleep questionnaires
Machine learning

ABSTRACT

Background: The growing awareness for the high prevalence of obstructive sleep apnea (OSA) coupled with the dramatic proportion of undiagnosed individuals motivates the elaboration of a simple but accurate screening test. This study assesses, for the first time, the performance of oximetry combined with demographic information as a screening tool for identifying OSA in a representative (i.e. non-referred) population sample.
Methods: A polysomnography (PSG) clinical database of 887 individuals from a representative population sample of São Paulo's city (Brazil) was used. Using features derived from the oxygen saturation signal during sleep periods and demographic information, a logistic regression model (termed OxyDOSA) was trained to distinguish between non-OSA and OSA individuals (mild, moderate, and severe). The OxyDOSA model performance was assessed against the PSG-based diagnosis of OSA (AASM 2017) and compared to the NoSAS and STOP-BANG questionnaires.
Findings: The OxyDOSA model had mean AUROC = 0.94 ± 0.02, Se = 0.87 ± 0.04 and Sp = 0.85 ± 0.03. In particular, it did not miss any of the 75 severe OSA individuals. In comparison, the NoSAS questionnaire had AUROC = 0.83 ± 0.03, and missed 23/75 severe OSA individuals. The STOP-BANG had AUROC = 0.77 ± 0.04 and missed 14/75 severe OSA individuals.
Interpretation: We provide strong evidence on a representative population sample that oximetry biomarkers combined with few demographic information, the OxyDOSA model, is an effective screening tool for OSA. Our results suggest that sleep questionnaires should be used with caution for OSA screening as they fail to identify many moderate and even some severe cases. The OxyDOSA model will need to be further validated on data recorded using overnight portable oximetry.

1. Introduction

Obstructive sleep apnea (OSA) is a public health problem that affects a large part of the general adult population with [1,2] up to 50% of the general adult male population and 23% of the general adult female population [2]. It is estimated that a large proportions of individuals with OSA are undiagnosed and untreated [3–5]. Several studies have shown that if untreated, OSA increases the risk of cardiovascular diseases [6], metabolic syndrome [7] and diabetes [8]. Undiagnosed and untreated OSA is associated with behavioral alteration [9], workplace productivity losses and increased motor vehicle accidents due to sleepiness on the road, leading to high clinical and economical costs [10].

OSA screening is a priority so that treatment can be instituted before a major health effect of OSA develops [11]. While polysomnography (PSG) is the gold standard for diagnosing OSA, it is not suitable for mass screening due to its high cost and lack of accessibility. A screening tool must be less expensive and more convenient than the standard PSG albeit some loss of accuracy [11]. Tools that have been studied for OSA screening include questionnaires, analysis of upper airway morphology [12] and monitoring of biosignals using portable technologies [13–16]. OSA screening can be performed on the general population or in a clinical setting for a specific group of patients.

Overnight drops in oxygen saturation is characteristic of individuals with OSA. In addition, it is possible that the repetitive nocturnal hypoxemia in OSA causes oxidative stress contributing to the pathogenesis of cardiovascular morbidity [17,18]. Thus, oximetry is a good candidate for

* Corresponding author.
E-mail address: jbehar@technion.ac.il (J.A. Behar).
[1] Equal contribution.
[2] Co-senior authorship.

https://doi.org/10.1016/j.eclinm.2019.05.015

Research in context

Evidence before this study

The need for a mass screening tool for OSA has motivated the research and development of sleep questionnaires (e.g. STOP-BANG, NoSaS) and single channels monitors in identifying patients at risk of OSA. Oximetry has been studied as a candidate for single channel monitoring of OSA. However, the performance of these screening options on a representative population sample has not been studied and their robustness against using different hypopnea rules (recommended/alternative) and scoring indexes (AHI/RDI) has not been assessed.

Added value of this study

This research shows, for the first time, that biomarkers derived from oximetry are accurate predictors for mass OSA screening on a large (n = 887) representative population sample. This research demonstrates that sleep questionnaires miss important clinical cases of OSA. By adding oxygen saturation based biomarkers to the predictive model (OxyDOSA), all the important clinical cases are identified. Finally, we demonstrate the robustness of the OxyDOSA model when using the recommended and acceptable rules for hypopneas as well as when using different diagnostic indexes (AHI/RDI).

Implication of all available evidence

The OxyDOSA model that combines some demographic information and oxygen saturation biomarkers is an accurate and robust test for mass OSA screening. In comparison, the reliability of sleep questionnaires for OSA screening is limited because they fail to identify some serious cases of OSA.

OSA screening, both because overnight oxygen desaturations are biomarkers of the disease and because they might best reflect the consequences of the disease on cardiovascular function. Oximetry has been studied as a screening tool in sleep clinic patients [19–22], in surgical patients [23] or group of individuals with specific comorbidities – see del Campo et al. [24] and Uddin et al. [25] for comprehensive reviews. However, no study evaluated the diagnostic performance of oximetry combined with demographic information on a representative population sample. In addition, no study assessed the stability of such model across different diagnostic hypopnea rules (recommended and alternative) and scoring indexes (apnea hypopnea index and respiratory disturbance index).

This study focuses on assessing the potential of automated oximetry analysis as an accurate screening tool for OSA on a representative population sample. For that purpose, oximetry derived features as well as demographic information were used to train a logistic regression classifier. The predictive value of this model, termed the Oxygen saturation and Demographics based model for OSA (OxyDOSA) was evaluated against the reference polysomnographic and clinical diagnosis (AASM 2017 [26]) and compared to the performance of the STOP-BANG [27] and the recently introduced NoSAS questionnaire for OSA screening [28]. The NoSAS is based on a subset of the STOP-BANG questionnaire features, namely: age, sex, BMI, neck circumference and snoring. The predictive performances of the OxyDOSA are evaluated for different hypopnea rules and scoring indexes.

2. Methods

2.1. Database

We used the São Paulo Epidemiologic Sleep Study (EPISONO) cohort study database [1,29] which consists of a sleep survey with a probabilistic three-stage cluster sample of São Paulo inhabitants representative of the population according to gender, age (20–80 years), and socioeconomic status. Face-to-face interviews and in-lab full-night PSG using a nasal cannula and a thermistor were performed. A total of 1042 volunteers underwent PSG (refusal rate = 5.4%). The data were recorded using the Embla system (Embla S7000, Embla Systems, Inc., Broomfield, CO., USA). The Nonin XPOD pulse oximeter (Nonin Medical, Inc., Plymouth, Minnesota, USA) was used by the Embla system for recording. The signal was sampled at 3 Hz and with accuracy ±2%. Since the time of the original EPISONO study in 2007 [1] the diagnostic recommendations provided by the American Academy of Sleep Medicine (AASM) have changed. Thus, in order to use the most recent diagnostic guidelines the PSG recordings were fully rescored according to the newest guidelines. The updated diagnosis used in this work follows the AASM 2017 guidelines [30]. We used the recommended rule for the definition of hypopneas. OSA severity was defined with respect to the AHI, as mild ($5 \le AHI < 15$), moderate ($15 \le AHI < 30$) and severe ($AHI \ge 30$). The data from 887 patients (Table 1) could successfully be rescored using the 2017 AASM [30] guidelines (see Supplement 1.1 for more details).

2.2. Classes

Because the end goal of the screening test is to identify patients with OSA versus non-OSA, the following binary classification task was considered: OSA (mild, moderate, and severe) versus non-OSA. Although the identification of moderate and severe OSA individuals is critical, an OSA screening test should also aim at identifying mild OSA. Moderate and severe OSA patients are the individuals a screening test should not miss because these patients will need to be treated. By opposition, a mild OSA patient may not be systematically treated with CPAP [24]. However, mild identification is important because the condition can further develop and they might benefit from making lifestyle changes such as sleeping on the side or changing their diet. In particular, Tuomilehto et al. [25] showed that early weight reduction was a curative treatment for the vast majority of patients with mild OSA. In addition, there is some evidence that even patients suffering from mild OSA may be at risk of hypertension [26], car accident [27] and that they can benefit from treatment [28]. However, in order to take into account, the relative importance between mild versus moderate versus severe individuals, the cost function of the logistic regression classifier was penalized with weights 1, 5 and 10 for misclassifying mild, moderate and severe respectively. This emphasizes the logistic regression model to recognize moderate and severe cases.

2.3. Features

We used the following oxygen saturation features: the 3% oxygen desaturation index (ODI), the mean oxygen saturation ($MSpO_2$), lowest

Table 1
Study database. The diagnosis is based on the ICSD-3 and AASM 2017 guidelines and using the recommended rule for hypopnea.

Diagnosis	Number	Percentage (%)
Non-OSA	503	56.7
Mild OSA	206	23.2
Moderate OSA	103	11.6
Severe OSA	75	8.5
Total	887	

value of oxygen saturation (SpO_2 Nadir), and the proportion of time spent with oxygen saturation under 90% (T90). The locations of the desaturations returned by the Embla monitor were used to compute the ODI and we computed the other oxygen saturation features from the raw oxygen saturation time series. We defined the SpO_2 Nadir feature as the first percentile of the valid oxygen saturation time series i.e. excluding areas with abnormal values as returned by the oximeter. The raw demographic features used for the STOP-BANG questionnaire were also available from the study (Tables 2 and 3). Fig. S1 and S2 show the distributions/bar plots obtained for each feature and Tables 2 and 3 show their median and interquartile ranges.

2.4. Machine Learning

The classifier model must identify the highest number of individuals with OSA even to the detriment of having a higher proportion of false positive (i.e. we seek a high sensitivity). However, too many false positive (i.e. a low specificity) will overload sleep clinics with non-OSA individuals which is time consuming and costly. Typically, OSA sleep questionnaires or oximetry based algorithms are evaluated using heuristics or simple thresholding over a number of scored answers or a hard ODI threshold. In this work, we used logistic regression to elaborate our machine learning models. We performed nested cross fold validation by rotating the test set using 5-folds with stratification of the individuals falling in the two classes (non-OSA/OSA). This was done in order to be able to report the average performances of the models on the whole dataset. Feature selection was performed using least absolute shrinkage and selection operator (Lasso) [31]. Feature selection enables to select the combination of oxygen saturation and demographic features that give the best predictive value to a given logistic regression model. For each model being trained the dataset is divided into: 64% training, 16% validation and 20% test set.

Repeated random sub-sampling validation [4] (100 runs) with stratification was performed for each model on the training set (see Fig. S3). In short, nested cross-validation consists of: [1] an outer k-fold cross-validation loop that is used to split the data into training and test folds. We used 5-folds cross-validation for the outer loop and the 5-folds were divided the same way for all models evaluated; [2] an inner loop which is used to select the model using cross-validation on the training set. We used repeated random sub-sampling validation for the inner loop. Model parameters are set by the analysis of the models prediction on the validation sets (inner loop). Then, the model is trained on the whole training set and evaluated on the separate test set (outer loop). Based on the outer test folds, the average and variance performance of the models are reported. This allows to evaluate the predictive capacity of a model type on the whole database. Stratification consisted in ensuring that each fold had the same percentage of OSA and non-OSA individuals. Repeated random sub-sampling validation used in the inner loop consists of randomly selecting part of the training set data for training the model and part for validation while repeating this process a number of times (100 times in this study). The statistics of the classifiers are reported on the test sets for a threshold at 0.5. In order to train the LR model, the missing data were replaced by the average value of the corresponding missing feature across all the training set

Table 3
Ordinal valued features.

Feature	Healthy (n = 503)	Mild (n = 206)	Moderate (n = 103)	Severe (n = 75)
Freq. daytime fatigue				
– Never	166	89	36	37
– 1–2×/month	33	9	8	5
– 1–2×/week	127	49	25	16
– 3–4×/week	48	10	10	1
– Daily	129	49	24	16
	(503/503)	(206/206)	(103/103)	(75/75)
Snoring Level				
– Never	245	42	17	3
– As loud as breathing	100	54	23	14
– As loud as talking	75	39	25	20
– Louder than talking	24	17	11	10
– Can be heard in another room	25	41	24	27
	(469/503)	(193/206)	(100/103)	(74/75)
Freq. Observed Stop Breathing				
– Never	278	106	54	28
– 1–2×/month	25	17	7	9
– 1–2×/week	13	10	3	5
– 3–4×/week	10	5	4	4
– Daily	25	22	17	17
	(351/503)	(160/206)	(85/103)	(63/75)
High BP				
– Yes	77	57	40	42
– No	393	133	57	32
	(470/503)	(190/206)	(97/103)	(74/75)
Gender				
– Female	321	94	43	29
– Male	182	112	60	46
	(503/503)	(190/190)	(103/103)	(75/75)

The numbers in parenthesis indicate the number of individuals for whom the information was recorded out of the given subset. BP: blood pressure.

Table 2
Median (MED) and interquartile range (±IQR) statistics on oxygen saturation based features and demographic features.

Type	Feature	Non-OSA (n = 503) MED (±IQR)	Mild OSA (n = 206) MED (±IQR)	Moderate OSA (n = 103) MED (±IQR)	Severe OSA (n = 75) MED (±IQR)	p-value Kruskal-Wallis
SpO_2	ODI	0.42 ± 1.00 (503/503)	3.32 ± 3.55 (206/206)	9.40 ± 5.82 (103/103)	23.32 ± 18.77 (75/75)	1.1e-116
	$MSpO_2$	96.4 ± 1.6 (503/503)	94.9 ± 1.8 (206/206)	94.7 ± 2.2 (103/103)	93.5 ± 2.2 (75/75)	1.1e-52
	SpO_2 Nadir	94.0 ± 2.0 (503/503)	92.0 ± 3.0 (206/206)	90.0 ± 3.0 (103/103)	86.0 ± 7.0 (75/75)	5.5e-84
	T90	0.000 ± 0.0001 (503/503)	0.0010 ± 0.0062 (206/206)	0.0090 ± 0.0239 (103/103)	0.0435 ± 0.0914 (75/75)	3.5e-83
DE	Age	34.00 ± 17.00 (503/503)	47.00 ± 18.00 (206/206)	50.00 ± 20.00 (103/103)	57.00 ± 16.50 (75/75)	4.8e-46
	Neck Circ.	34.00 ± 4.70 (486/503)	37.00 ± 5.00 (202/206)	38.10 ± 5.35 (98/103)	39.00 ± 7.15 (74/75)	3.1e-32
	BMI	24.56 ± 5.30 (502/503)	27.34 ± 5.59 (205/206)	28.39 ± 5.92 (103/103)	29.24 ± 8.07 (75/75)	2.9e-29

The 3% oxygen desaturation index (ODI), the mean oxygen saturation ($MSpO_2$), lowest value of oxygen saturation (SpO_2 Nadir), and the proportion of time spent with $SpO_2 < 90\%$ (T90). The numbers in parenthesis indicate the number of individuals for whom the information was recorded out of the given subset. BMI: body mass index, Neck Circ.: neck circumference. The ODI and other oxygen saturation features are computed over the total recording time.

individuals. Features were normalized by subtracting the mean and dividing by the standard deviation (z-transform) computed on the training sets.

Four sets of classifiers were evaluated for comparison: LR-SB for which classifiers were trained using all the raw demographic features provided by the STOP-BANG questionnaire [27]; LR-ODI for which classifiers were trained using the oxygen desaturation index as the sole feature; LR-SpO_2 for which classifiers were trained using the four oxygen saturation features available; the OxyDOSA model for which classifiers were trained using features selected (using Lasso) among all oxygen saturation and the raw demographic features available. Table 4 describes in more details the list of the features used for each set of models evaluated. In order to make an objective comparison between the different models, the same divisions between training and test sets (i.e. division of the folds) were used for all models.

2.5. Statistics

The statistics used were sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), accuracy (Ac) and the harmonic mean between the Se and PPV termed F_1 measure. In the context of this study these are defined as: Se, the percentage of individuals with OSA that have been correctly identified as OSA out of the whole OSA population; Sp, the percentage of individuals without OSA that have been identified as such out of the whole non-OSA population; PPV, the percentage of individuals correctly identified as having OSA out of all the individuals that were predicted as having OSA; NPV, the percentage of individuals correctly identified as non-OSA out of all the individuals that were predicted as not having OSA; Ac, the percentage of individual accurately classified. We also report the area under the receiver operator curve (AUROC). Finally, we also report the per-subclass sensitivities (Se-mild, Se-moderate and Se-severe) since misclassification of mild individuals does not have the same clinical implication than the misclassification of moderate and severe OSA individuals.

Table 4
List of the features used for each of the models evaluated.

	Feature	SB	NoSAS	LR-SB	LR-ODI	LR-SpO_2	OxyDOSA
1	Score Snore	×					
2	Score Tired	×					
3	Score Stop Breathing	×					
4	Score High BP	×		×			×
5	Score BMI	×					
6	Score Age	×					
7	Score Neck Circ.	×					
8	Score Gender	×	×	×			×
9	Raw Snoring Level		×*	×			×
10	Raw Freq. Daytime Fatigue			×			×
11	Raw Freq. Observed Stop Breathing			×			×
12	Raw BMI		×	×			×
13	Raw Age		×	×			×
14	Raw Neck Circ.		×	×			×
15	ODI				×	×	×
16	T90					×	×
17	$MSpO_2$					×	×
18	SpO_2 Nadir					×	×

SB: STOP-BANG, BP: blood pressure, BMI: body mass index, ODI: oxygen desaturation index, T90: time spent with $SpO_2 < 90\%$. $MSpO_2$: mean oxygen saturation, SpO_2 Nadir: lowest value of oxygen saturation. Score values relate to the yes/no scored answer to the STOP-BANG questionnaire whereas raw values represent the raw values of the features used to answer the STOP-BANG questions. * For the NoSAS, snoring was considered positive if the answer to the snoring level was at least "As loud as breathing".

2.6. Role of the funding sources

The original clinical trial of the EPISONO study [1,29] was supported by the Associação Fundo de Incentivo a Pesquisa (AFIP), Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) and São Paulo Research Foundation (FAPESP).

3. Results

Fig. 1 shows the overall distributions of the logistic regression classifiers outputs by OSA severity. Any individual having a probability superior to the threshold represented by the dotted horizontal line will be predicted as having OSA. Relative to this threshold, Table 5 summarizes the performance of the four logistic regression models, STOP-BANG and NoSAS questionnaires on the whole dataset (n = 887); the LR-SB obtained AUROC = 0.87 ± 0.04, the LR-ODI obtained AUROC = 0.92 ± 0.01, the LR-SpO_2 obtained AUROC = 0.92 ± 0.02 and the OxyDOSA model obtained AUROC = 0.94 ± 0.02. The STOP-BANG and NoSAS questionnaires obtained AUROC = 0.77 ± 0.04 and AUROC = 0.83 ± 0.03 respectively. Fig. 2 shows the receiver operating characteristic curves obtained on the rotated test sets. The maximal AUROC and Se were obtained for the OxyDOSA (Table 5). Analysis of the feature importance of OxyDOSA showed that ODI, Age and the Snoring Level were the most relevant features (Fig. 3) in elaborating the predictive model. Study of the false negatives for the NoSAS (Table 5) showed that a total of 208 patients with OSA were not identified among which 139 were mild (67% of all mild cases), 46 were moderate (45% of all moderate cases) and 23 had severe OSA (31% of all severe cases). Similarly the STOP-BANG missed many moderate and severe cases (Table 5). The AUROC for the STOP-BANG and NoSAS were AUROC = 0.77 ± 0.04 and AUROC = 0.83 ± 0.03 respectively. LR-SB missed very few moderate (3/103) and severe (2/75) cases and it had an AUROC = 0.87 ± 0.04. However, LR-SB had a low Sp = 0.64 ± 0.05.

There were no significant differences between using the ODI only (LR-ODI) and using all the four available oxygen saturation features (LR-SpO_2). The LR-ODI had AUROC = 0.92 ± 0.01, Se = 0.82 ± 0.03 and Sp = 0.88 ± 0.03 (Table 5). A total of 69 OSA individuals were not identified by LR-ODI among which 63 were mild (31% of all mild) and 6 were moderate (6% of all moderate).

The OxyDOSA model had AUROC = 0.94 ± 0.02, Se = 0.87 ± 0.04 and Sp = 0.85 ± 0.03. The study of the false negatives for OxyDOSA showed that 49 patients with OSA were not identified among which 48 were mild (23% of all mild cases), and only one was moderate (0.97% of all moderate cases). The standard deviations of all the performance statistics were small (Table 5) thus demonstrating the stability of the model.

Table S1 summarizes the individuals' reference diagnosis when using the acceptable rule for hypopnea definition (AHI A2017) and when using the respiratory disturbance index (RDI). Depending on the hypopnea rule used and the index (AHI/RDI), the prevalence of OSA varied from 28.1% to 51.2%.

The Kruskal–Wallis test was conducted to examine the differences between the features reported in Table 2 for patients in the different subgroups (non-OSA, mild, moderate, severe). For all the features of Table 2 the test rejected the null hypothesis that the sample data come from the same distribution under the null hypothesis: Age ($\chi^2 = 213.6$, p = 4.8e−46, df = 3), Neck Circ. ($\chi^2 = 149.6$, p = 3.1e-32, df = 3), BMI ($\chi^2 = 135.9$, p = 2.9e−29, df = 3), ODI ($\chi^2 = 539.9$, p = 1.1e−116, df = 3), MSpO2 ($\chi^2 = 539.9$, p = 1.1e−52, df = 3), SpO2Nadir ($\chi^2 = 389$, p = 5.5e−84, df = 3), T90 ($\chi^2 = 385$, p = 3.5e−83, df = 3).

4. Discussion

Our first major finding is that questionnaires (NoSAS and STOP-BANG) performances are limited for accurate OSA screening as they fail to identify some moderate and severe individuals (Table 5). It is

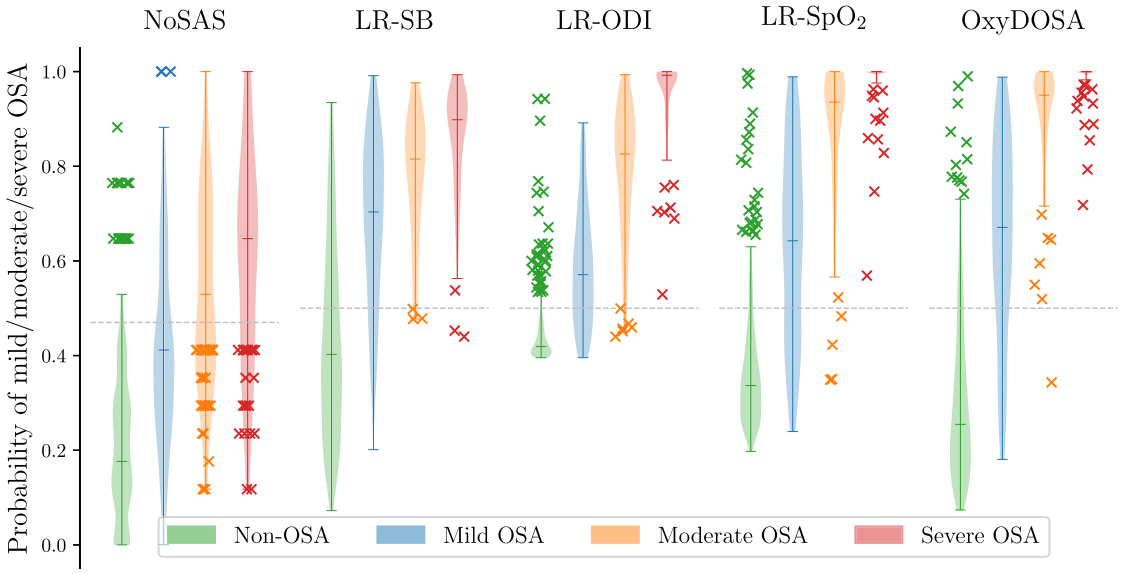

Fig. 1. Overall distributions ('violin plots') of logistic regression models outputs for the different groups of individuals (Non-OSA, mild OSA, moderate OSA, severe OSA). The threshold at 0.5 is displayed in dotted horizontal line. Any individual having a probability superior to this threshold will be predicted as having OSA by the LR model. The individual crosses highlight the outlier individuals. In particular, note that the OxyDOSA model only misses one moderate out of all the moderate and severe patients. For the NoSAS the distributions were obtained by normalizing the NoSAS score by the total number of points (i.e. NoSAS score divided by 17).

important to note that the NoSAS questionnaire was originally designed for identifying individuals with AHI > 20 [28]. It is thus not surprising that within the context of our study it had a low overall Se since we consider as OSA all individuals having AHI > 5. Yet, the NoSAS missed over 45% of the moderate (15 ≤ AHI < 30) and 31% of the severe (AHI ≥ 30) OSA cases which demonstrates some limitations of the model. In Craig et al. [32], the authors showed on the MOSAIC randomized controlled trial that continuous positive airway pressure improves sleepiness but not calculated vascular risk in minimally symptomatic OSA individuals. We thus decided to investigate whether the 14 severe cases missed by the STOP-BANG were individuals with no symptoms reported as part of the STOP-BANG. Table S2 summarizes the STOP-BANG answers for these individuals together with the AHI and ODI. Out of the 14 patients: two presented symptoms of snoring and 9 of daytime tiredness. Overall out of the 14 misclassified severe, 11 had at least one symptom. In addition, we investigated if the 14 STOP-BANG misclassified individuals were "borderline" moderate. The mean AHI was 42.6 among these patients and four had an AHI above or equal to 50. One individual even had an AHI of 77. These results suggest that the STOP-BANG missed clinically relevant cases. Interestingly, most of these cases were women with neck-size<40 cm and BMI < 35 kg/m^2. As a conclusion, overall, questionnaires miss moderate and even some severe OSA cases which highlight their limitation for accurate OSA screening in a representative population sample.

Second, we show that by training a logistic regression model over the raw demographic features (by opposition to binary answers to the STOP-BANG/NoSAS questionnaires) of the STOP-BANG, the LR-SB model outperform the STOP-BANG and NoSAS questionnaires significantly: AUROC was 0.87 ± 0.04 for the LR-SB against 0.77 ± 0.04 and 0.83 ± 0.03 for the STOP-BANG and NoSAS respectively. This result encourages using raw features (rather than using ordinal valued features) and weighting them with decimal weights (rather than with integer weights). Although, the LR-SB loses some human 'interpretability', in that it is not anymore the result of a sum of integers weighting each of the questions, its higher performances should encourage the adoption of such data-driven models. Investigation of the feature weights for the LR-SB (Fig. S7) revealed that Age, Snoring Level, Neck Circumference, BMI and Gender were the most predictive in accordance with

Table 5
Performance of the models (average and standard deviation for the test sets) evaluated against the AHI R2017.

Statistics/model	AUROC	Ac	F_1	NPV	PPV	Se	Se-mild	Se-moderate	Se-severe	Sp
NoSAS	0.83 ± 0.03	0.72 ± 0.03	0.58 ± 0.07	0.69 ± 0.03	0.81 ± 0.04	0.46 ± 0.08 (176/384)	0.33 ± 0.04 (67/206)	0.54 ± 0.14 (57/103)	0.69 ± 0.14 (52/75)	0.92 ± 0.02 (463/503)
STOP-BANG	0.77 ± 0.04	0.72 ± 0.02	0.65 ± 0.04	0.73 ± 0.03	0.70 ± 0.03	0.61 ± 0.05 (233/384)	0.47 ± 0.06 (97/206)	0.71 ± 0.13 (75/103)	0.81 ± 0.11 (61/75)	0.81 ± 0.02 (405/503)
LR-SB	0.87 ± 0.04	0.75 ± 0.04	0.76 ± 0.04	0.89 ± 0.04	0.66 ± 0.04	0.90 ± 0.04 (345/384)	0.84 ± 0.04 (172/206)	0.97 ± 0.03 (100/103)	0.97 ± 0.06 (73/75)	0.64 ± 0.05 (324/503)
LR-ODI	0.92 ± 0.01	0.85 ± 0.03	0.83 ± 0.03	0.87 ± 0.02	0.84 ± 0.04	0.82 ± 0.03 (315/384)	0.70 ± 0.04 (143/206)	0.94 ± 0.04 (97/103)	1.00 ± 0.00 (75/75)	0.88 ± 0.03 (443/503)
LR-SpO_2	0.92 ± 0.02	0.85 ± 0.02	0.82 ± 0.03	0.87 ± 0.03	0.82 ± 0.01	0.83 ± 0.05 (317/384)	0.70 ± 0.07 (143/206)	0.96 ± 0.04 (99/103)	1.00 ± 0.00 (75/75)	0.86 ± 0.01 (435/503)
OxyDOSA	0.94 ± 0.02	0.86 ± 0.03	0.84 ± 0.04	0.90 ± 0.03	0.82 ± 0.03	0.87 ± 0.04 (335/384)	0.77 ± 0.05 (158/206)	0.99 ± 0.02 (102/103)	1.00 ± 0.00 (75/75)	0.85 ± 0.03 (428/503)

Four sets of classifiers were evaluated for comparison. These are denoted: LR-SB for which classifiers were trained using all the demographic features used for the STOP-BANG questionnaire; LR-ODI for which classifiers were trained using the oxygen desaturation index as the sole feature; LR-SpO_2 for which classifiers were trained using all the oxygen saturation features; OxyDOSA for which classifiers were trained using features selected from all oxygen saturation and the demographic features available from the STOP-BANG questionnaire. Statistics are reported for the test sets.

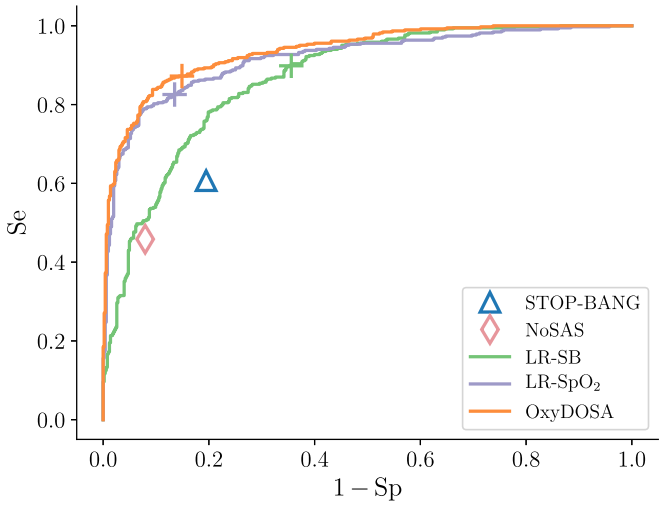

Fig. 2. The receiver operating characteristic (ROC) curves obtained on the validation sets for the following evaluated models: LR-SB, LR-SpO_2 and OxyDOSA. The statistics obtained for the NoSAS and STOP-BANG are also plotted as symbols. Crosses represent the points on the ROC curves which were selected for the different logistic regression models. Corresponding AUROC and other statistics are summarized in Table 5.

the results of Marti-Soler et al. [28] in elaborating the NoSAS questionnaire.

The third major finding is that using oximetry (LR-ODI and LR-SpO_2) it was possible to significantly outperform the questionnaires and the LR-DE performances. This demonstrates that night oximetry is an accurate predictor for OSA. However, although no severe cases were missed, oximetry alone still missed six (LR-ODI) and four (LR-SpO_2) moderate cases.

The fourth major finding is that by combining both the oximetry and the demographic features we could further improve the prediction accuracy of the logistic regression classifier (Table 5). We termed the resulting model OxyDOSA. OxyDOSA did not missed any of the severe cases and only one moderate case. Thus we showed that by using the OxyDOSA model, we could identify all the important cases of OSA while keeping a reasonable specificity ($Sp = 0.85 \pm 0.03$).

In addition to the overall statistic performance of OxyDOSA, the model has the merit to output a probability value of being OSA

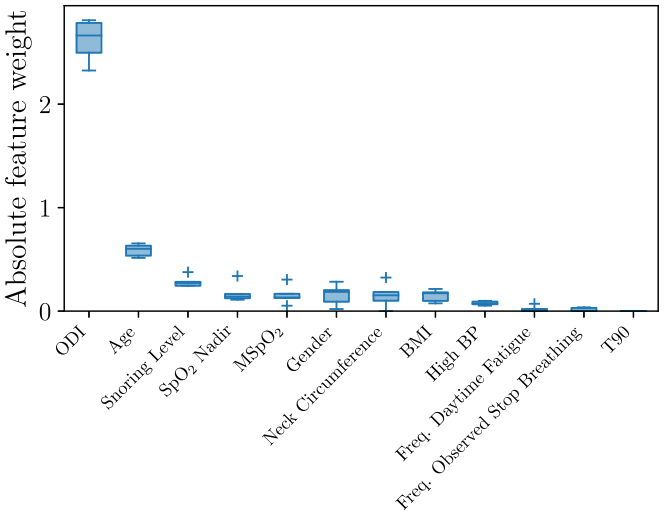

Fig. 3. Box plots of the feature weights on the outer loop folds. This figure highlights the relative importance of the different features in identifying individuals with OSA. This shows that the OxyDOSA prediction mainly relies on the ODI, Age, Snoring Level, SpO_2 Nadir, $MSpO_2$, Gender, Neck Circumference and BMI.

(Fig. 1). This probabilistic output enables to distinguish between individuals who might be suffering from mild OSA from the more severe cases. In particular, individuals with a clear positive test for OSA would be diagnosed based on the oximetry test i.e. without the need for a PSG. Conversely, individuals where oximetry and symptoms are equivocal would be recommended to have a PSG. In order to evaluate whether the prediction could be improved by using more advanced machine learning models we tried a random forest approach. No improvement was obtained with respect to logistic regression model. In addition, we noted that combining oxygen saturation features (LR-SpO_2) versus using the ODI only feature (LR-ODI) did not significantly improved the prediction of the model with respect to LR-ODI. We interpret the lack of improvement reached by using more advanced machine learning models and additional oxygen saturation features, to be due to the intrinsic definition of sleep apnea i.e. the medical condition was originally defined as an event count thus favoring very specific features and their count rather than other characteristics embedded in the signal.

Finally, to test whether our conclusions for the logistic regression models were robust across different hypopnea rules (recommended/alternative) and diagnostic indexes (AHI/RDI), we re-trained all the models against the AASM 2017 guidelines using the alternative rule (AHI A2017) and against the RDI based diagnosis for the recommended and acceptable rules (RDI R2017 and RDI A2017). Table S1 summarizes the number of individuals in each category with respect to the diagnostic indexes and hypopnea scoring rule. Performance statistics and distributions are provided in Tables S3–S5 and Figs. S4–S6. The OxyDOSA had the largest AUROC in all cases and missed less moderate and severe cases than the alternative models. The LR-SB provided improved results over the NoSAS or STOP-BANG questionnaires in all instances. We also noted that questionnaires performance varied significantly with respect to the diagnostic index and the hypopnea rule used; for example, the STOP-BANG sensitivity varied from 0.61 with AHI R2017 up to 0.73 with AHI A2017. This highlights that the questionnaires are tuned against a particular diagnostic guideline, index and hypopnea rule. We showed that the OxyDOSA can be trained against alternative guidelines or diagnostic rules and that its comparative performances to other models stays superior.

Untreated OSA increases healthcare utilization and reduces work performance [33]. At present, connected oximetry for mass OSA screening may be particularly worthwhile for places with barriers to symptomatic patients accessing diagnostic services but where those with a positive diagnosis would be able to access CPAP treatment and long-term follow-up.

4.1. Comparison with Other Studies

The usage of oximetry as a test for OSA screening has been explored by a number of other studies [19–22,34] – see del Campos et al. [24] and Uddin et al. [25] for comprehensive reviews. However, often these studies suffer from some important limitations: low sample size, old AASM diagnostic guidelines, hard ODI threshold (instead of learning it from the data) and database biased toward a specific group of individuals (e.g. individuals referred to the sleep clinic, preoperative patients or patients with no other known pulmonary or cardiac condition). See Supplement 2.1 for more discussion. Our study offers a data-driven approach to create an algorithm capable of screening OSA individuals from a single channel sensor and few demographic features. We prove the viability of this approach on a large dataset ($n = 887$) of a representative population sample and using the latest AASM guidelines. We also show that our conclusions are robust to the hypopnea rule and the diagnostic index.

4.2. Limitations

The Embla monitor we used only considers desaturations that happen during sleep periods. In order to use oximetry as a single channel

test, all desaturation (i.e. including the ones that could be detected during wake periods) should be considered. This point represents the main limitation of this study. Second, although we used a relatively large dataset of 887 patients, the data were recorded for a South American population and thus it will be valuable to validate our conclusions on a separate dataset from a different ethnic group such as the HypnoLaus study [2]. Third, the oxygen saturation features that were used in this study were obtained from oximetry of sleep studies performed in the sleep clinic. It will be important to repeat a similar analysis on data recorded using a portable oximeter used within the patient home. Last, we estimated the AHI A2017 and RDI A2017 (Table S1) using the original oxygen desaturations annotations from the EPISONO study i.e. 3% desaturations although the acceptable rule considers 4% desaturations. Thus the AHI A2017 and RDI A2017 used are estimates.

4.3. Outlook on the Future of OSA Screening

The creation of intelligent algorithms combined with the ongoing innovations in designing novel wearable biosensors offers an unprecedented opportunity to monitor and manage patients remotely. In particular, it is realistic to expect that in a near future oximetry will become available in smartwatches such as the Apple watch (which already includes an ECG sensor). At this point the efforts necessary for performing a night oximetry test for sleep apnea will become even less than filling a sleep questionnaire online.

5. Conclusion

Our study shows on a representative population sample that oximetry combined with some minimal demographic information (the OxyDOSA model) is a viable option for accurate mass OSA screening. The OxyDOSA model had an overall Se $= 0.87 \pm 0.04$ and Sp $= 0.85 \pm 0.03$ and identify all the most important OSA cases. In comparison, the reliability of sleep questionnaires for OSA screening is limited because they fail to identify some serious cases of OSA (moderate and severe). The elaboration of data-driven screening tests in combination to the development of wearable biosensors will enable mass remote screening and monitoring of OSA. The OxyDOSA has been implemented as a web app and is available at: https://aim-lab.github.io/oxydosa.html

Contributors

JB and NP conceived and designed the research; NP, JB, QL and GC analyzed the data; LB, ST and FPGR provided clinical guidance on the conduct of the analysis and interpretation of the results. JB and NP drafted the manuscript; NP and JB prepared the figures; JB, NP, QL, SG, FPGR, LB, ST and GC edited and revised the manuscript, and approved the final version.

Funding

National research grants (see acknowledgments).

Acknowledgements

The work was supported by the Associação Fundo de Incentivo a Pesquisa (AFIP), Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) and São Paulo Research Foundation (FAPESP)- (LB and ST), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grants number 401569/2016-0 and 309336/2017-1 (LB) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grant number 306138/2017-4 (ST).

Declaration of Competing Interest

JB and NP hold shares in SmartCare Analytics Ltd.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eclinm.2019.05.015.

References

[1] Tufik S, Santos-Silva R, Taddei JA, Bittencourt LRA. Obstructive sleep apnea syndrome in the Sao Paulo epidemiologic sleep study. Sleep Med 2010;11:441–6.
[2] Heinzer R, Vat S, Marques-Vidal P, et al. Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study. Lancet Resp Med 2015;3:310–8.
[3] Finkel KJ, Searleman AC, Tymkew H, et al. Prevalence of undiagnosed obstructive sleep apnea among adult surgical patients in an academic medical center. Sleep Med 2009;10:753–8.
[4] Devaraj U, Rajagopala S, Kumar A, Ramachandran P, Devereaux PJ, D'Souza GA. Undiagnosed obstructive sleep apnea and postoperative outcomes: a prospective observational study. Respiration 2017;94:18–25.
[5] Young T, Evans L, Finn L, Palta M. Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. Sleep 1997;20:705–6.
[6] Drager LF, McEvoy RD, Barbe F, Lorenzi-Filho G, Redline S. Sleep apnea and cardiovascular disease: lessons from recent trials and need for team science. Circulation 2017;136:1840–50.
[7] Hirotsu C, Haba-Rubio J, Togeiro SM, et al. Obstructive sleep apnoea as a risk factor for incident metabolic syndrome: a joined Episono and Hypnolaus prospective cohorts study. Eur Respir J 2018;52:1801150.
[8] Reutrakul S, Babak M. Obstructive sleep apnea and diabetes: a state of the art review. Chest 2017;152:1070–86.
[9] Tufik S, Andersen ML, Bittencourt LR, de Mello MT. Paradoxical sleep deprivation: neurochemical, hormonal and behavioral alterations. Evidence from 30 years of research. An Acad Bras Cienc 2009;81:521–38.
[10] Knauert M, Naik S, Gillespie MB, Kryger M. Clinical consequences and economic costs of untreated obstructive sleep apnea syndrome. World J Otorhinolaryngol Neck Surg 2015;1:17–27.
[11] Pack A. Sleep apnea: Pathogenesis, diagnosis and treatment. CRC Press; 2016.
[12] Zonato AI, Bittencourt LR, Martinho FL, Ferreira Santos J, Gregório LC, Tufik S. Association of systematic head and neck physical examination with severity of obstructive sleep apnea-hypopnea syndrome. Laryngoscope 2003;133:973–80.
[13] Behar J, Roebuck A, Domingos JS, Gederi E, Clifford GD. A review of current sleep screening applications for smartphones. Phys Meas 2013;34.
[14] Roebuck A, Monasterio V, Gederi E, et al. A review of signals used in sleep analysis. Phys Meas 2014;35.
[15] Penzel T, Schobel C. New technology to assess sleep apnea: wearables, smartphones, and accessories. F1000Research 2018;7:413.
[16] Roebuck A, Clifford GD. Comparison of standard and novel signal analysis approaches to obstructive sleep apnea classification. Front Bioeng Biotech 2015;3:114.
[17] Lavie L. Obstructive sleep apnoea syndrome — an oxidative stress disorder. Sleep Med Rev 2003;7:35–51.
[18] Dyugovskaya L, Lavie P, Lavie L. Increased adhesion molecules expression and production of reactive oxygen species in leukocytes of sleep apnea patients. Am J Resp Crit Care Med 2002;165:934–9.
[19] Gurubhagavatula I, Maislin G, Pack A. An algorithm to stratify sleep apnea risk in a sleep disorders clinic population. Am J Resp Crit Care Med 2001;164:1904–9.
[20] Chiner E, Signes-Costa J, Arriero JM, Marco J, Fuentes I, Sergado A. Nocturnal oximetry for the diagnosis of the sleep apnoea hypopnoea syndrome: a method to reduce the number of polysomnographies? Thorax 1999;54:968–71.
[21] Jung DW, Hwang SH, Cho JG, et al. Real-time automatic apneic event detection using nocturnal pulse oximetry. IEEE Trans Biomed Eng 2018;65:706–12.
[22] Morillo DS, Gross N, León A, Crespo LF. Automated frequency domain analysis of oxygen saturation as a screening tool for SAHS. Med Eng Phys 2012;34:946–53.
[23] Chung F, Liao P, Elsaid H, Islam S, Shapiro CM, Sun Y. Oxygen desaturation index from nocturnal oximetry: a sensitive and specific tool to detect sleep-disordered breathing in surgical patients. Anesth Analg 2012;114:993–1000.
[24] del Campo F, Crespo A, Cerezo-Hernández A, Gutiérrez-Tobal GC, Hornero R, Álvarez D. Oximetry use in obstructive sleep apnea. Expert Rev Resp Med 2018;12:665–81.
[25] Uddin MB, Chow CM, Su SW. Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: a systematic review. Physiol Meas 2018;39:2018.
[26] Berry R, Budhiraja R, Gottlieb D. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. J Clin Sleep Med 2012;8:597–619.
[27] Chung F, Subramanyam R, Liao P, Sasaki E, Shapiro C, Sun Y. High STOP-Bang score indicates a high probability of obstructive sleep apnoea. Brit J Anaesth 2012;108:768–75.
[28] 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 Resp Med 2016;4:742–8.
[29] Santos-Silva R, Tufik S, Conway SG, Taddei JA, Bittencourt LRA. Sao Paulo epidemiologic sleep study: rationale, design, sampling, and procedures. Sleep Med 2009;10:679–85.
[30] Berry RB, Brooks R, Gamaldo C, et al. AASM scoring manual updates for 2017 (version 2.4). J Clin Sleep Med 2017;13:665–6.
[31] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 1996:267–88.

[32] Craig SE, Kohler M, Nicoll D, et al. Continuous positive airway pressure improves sleepiness but not calculated vascular risk in patients with minimally symptomatic obstructive sleep apnoea: the MOSAIC randomised controlled trial. Thorax 2012; 67:1090–6.
[33] Kapur V, Blough DK, Sandblom RE, et al. The medical cost of undiagnosed sleep apnea. Sleep 1999. https://doi.org/10.1093/sleep/22.6.749.
[34] Ben-Israel N, Tarasiuk A, Zigel Y. Obstructive apnea hypopnea index estimation by analysis of nocturnal snoring signals in adults. Sleep 2012;35:1299–305.