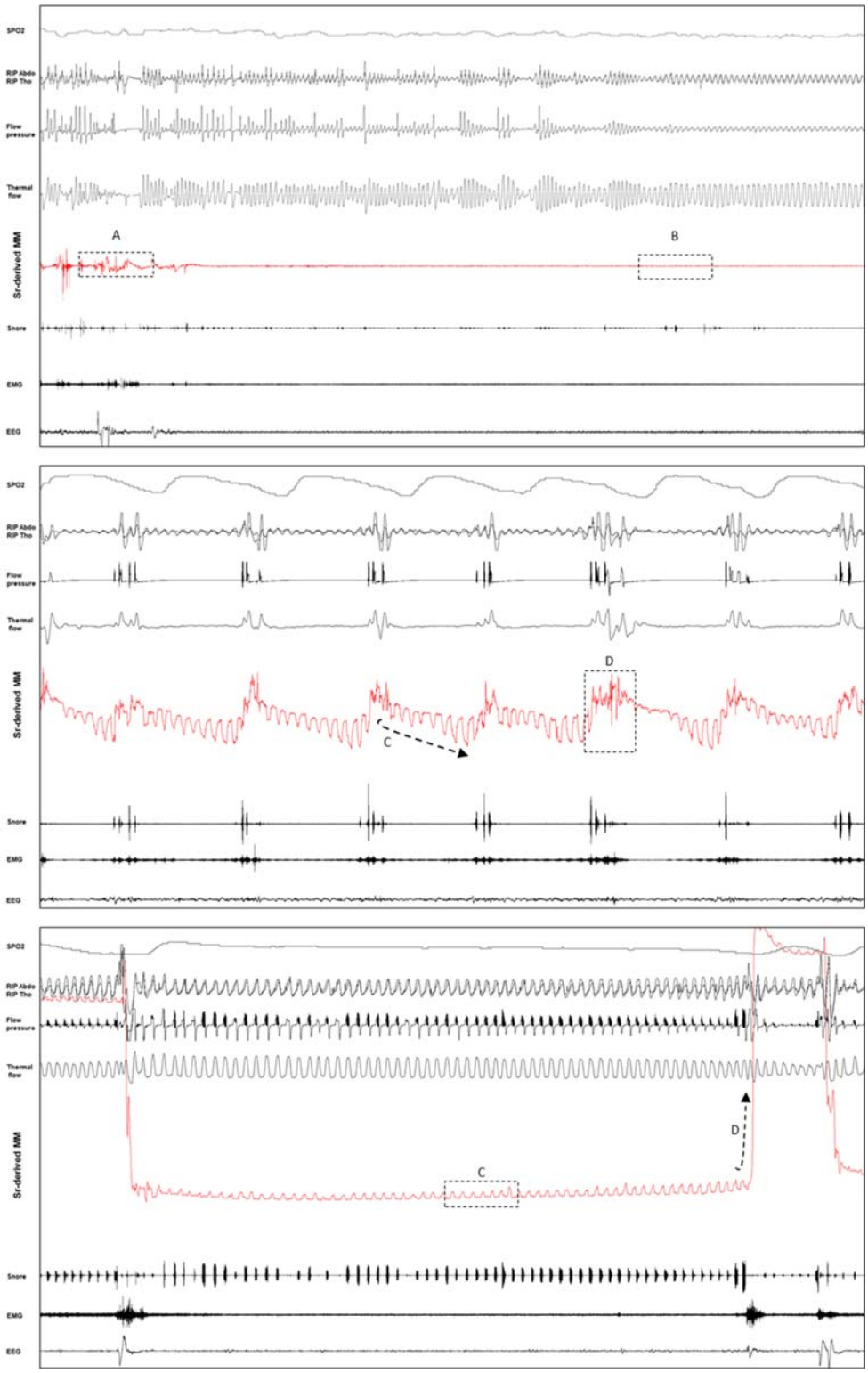
Supplementary Online Content

Pépin J-L, Letesson C, Le-Dong NN, et al. Assessment of mandibular movement monitoring with machine learning analysis for the diagnosis of obstructive sleep apnea. *JAMA Netw Open*. 2020;3(1):e1919657. doi:10.1001/jamanetworkopen.2019.19657

This supplementary material has been provided by the authors to give readers additional information about their work.

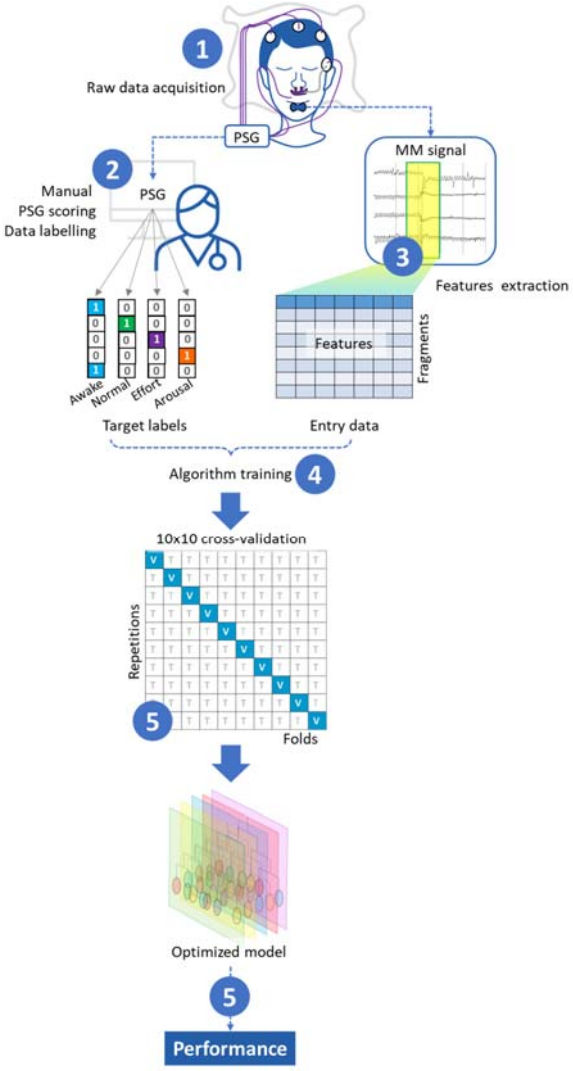
eFigure 1. Behaviour of the Sunrise MM signal (Sr-derived MM) after incorporation into PSG fragments of a transition from wake to sleep (upper panel), a sleep period marked with episodes of obstructive apnoea (middle panel) or a respiratory-effort related arousal (RERA) (lower panel).



Caption

SpO2: pulsed oxygen saturation, RIP Thx and Abd: thorax and abdominal respiratory inductive plethysmography, Flow pressure: airflow nasal pressure, Thermal flow: flow thermistor, Sr-derived MM: traces of the mandibular movements signals derived from Sunrise, Snore: microphone, EMG: electromyographic recording of the submental area, EEG: electroencephalographic recording of the C4-A1 electrode. The upper panel shows a 12-minutes fragment of a PSG with a transition from wake to sleep. Wake can be detected on Sr-derived MM through the identification of fast, irregular and non-predictable MM (see label A on the figure), whereas quiet sleep is characterized by smooth MM oscillations at the breathing frequency (see label B on the figure). The middle panel shows a 6-minutes fragment of a PSG with repeated OSA episodes. OSA can be detected on Sr-derived MM through the identification of oscillatory MM characteristic of respiratory effort (see label C on the figure) and brisk and abrupt MM indicating the closure of the mouth, a sign of arousal from sleep (see label D on the figure). The lower panel shows a 6-minutes fragment of a PSG a respiratory effort-related arousal (RERA). RERAs can be detected on Sr-derived MM through the identification of a long period of oscillatory MM indicating respiratory effort (see label C on the figure), ending by a brisk and abrupt arousal MM (see label D on the figure). A description of the labels used for characterizing MM can be found in Table e-2.

eFigure 2. Technical details on the Sunrise algorithm development phase



Caption

A multi-labels classification algorithm was developed and optimized in order to precisely identify sequential 30 seconds epochs of MM raw signals as wake, micro-arousals, sleep respiratory effort or normal sleep, based on relevant and non-redundant features.

1) The development was conducted on an independent sample of 100 patients recorded from the sleep laboratory at the CHU UCL Namur Site Ste Elisabeth in Belgium from July 2017 to October 2018. The algorithm construction and training were performed completely independently from the clinical validation described in the present paper. The training and validation samples characteristics were considered equivalent (see table e-1 for details on the clinical characteristics of the training and validation data sets).
2) For each patient, PSG data were manually and independently scored by 2 sleep experts that were blind to the study protocol and hypothesis. All sleep stages were visually scored according to the AASM rules (Berry et al., 2012). EEG arousals, sleep-related respiratory events were scored in accordance with the criteria established by the American Academy of Sleep Medicine Manual for Scoring Sleep

and Associated Events (AASM) as recommended by the ICSD-3 for OSA diagnosis. Based on the PSG scoring, one of 4 target labels (i.e. awake, quiet sleep, respiratory effort, micro arousal) was assigned for each PSG 30 seconds epoch. A label was given only when the two scorers unanimously agreed upon the scoring of the considered fragment. Ambiguous fragments were labelled as artefacts and excluded from training data.

3) At the same time, relevant features were extracted from the corresponding MM raw signal sequences and were used as input data for the algorithm training (Table e-2). The 30 seconds signal sequences were each given a true label based on the reviewed PSG scoring (i.e. wake, quiet sleep, respiratory effort, arousal).
4) The PSG-labelled MM fragments were then fed to the algorithm, crosschecking all available features to best classify/predict each 30-second fragment based solely on MM signal. The algorithm applies a classification on each 30-seconds epoch: it tests whether the processed MM signals could be classified as awake movements, and classifies "sleep" fragments into respiratory effort, arousal or quiet sleep.
5) The parameters of the algorithm were optimized through a repeated 10-folds cross validation procedure. The original data was randomly partitioned into 10 equally sized blocks. A single subsample was retained as the validation data for testing the algorithm, and the remaining 9 subsamples were used as training data. The cross-validation process was repeated 10 times to ensure that each subsample was used only once as validation data.
6) The averaged results from the cross-validation indicated a good performance of the algorithm. It could differentiate awake and sleep periods with a mean balanced accuracy of 0.80 ± 0.04; and achieved a good agreement with PSG in classification of MM sleep fragments into quiet sleep, respiratory efforts or micro arousals with a Cohen's kappa of 0.72 ± 0.04.

eTable 1. Characteristics of the training and validation samples

	Training set (n= 100)	Validation set (n=376)
Characteristics		
Gender	Male: 52 (52 %) Female: 48 (48 %)	Male: 207 (55 %) Female: 169 (45 %)
Age (yrs)	51 (15.6)	49.50 (19.33)
BMI (kg/m²)	28.90 (10.26)	30.25 (9.48)
Neck circumference (cm)	40 (8)	40 (6)
Smoking status	No: 52 (52 %) Yes: 19 (19 %) Ex-smoker: 27 (27 %) Not reported: 2 (2 %)	No: 202 (54 %) Yes: 77 (20 %) Ex-smoker: 93 (25 %) Not reported: 4 (1 %)
Sleep profile		
PSG_TST (h)	7.27 (1.47)	7.20 (1.60)
PSG_RDI (n/h)	22.8 (25.1)	18.80 (22.99)
PSG_Arl (n/h)	23.55 (18.65)	19.35 (19.99)
ESS score	10 (8)	11 (8)
Comorbidities		
Hypertension	No: 64 (64 %) Yes: 36 (36 %)	No: 264 (70 %) Yes: 107 (29 %) Not reported: 5 (1 %)
Type 2 diabetes	No: 91 (91 %) Yes: 9 (9 %)	No: 337 (90 %) Yes: 34 (9 %) Not reported: 5 (1 %)
Coronary heart disease	No: 97 (97 %) Yes: 3 (3 %)	No: 360 (96 %) Yes: 11 (3 %) Not reported: 5 (1 %)
Stroke history	No: 98 (98 %) Yes: 2 (2 %)	No: 365 (97 %) Yes: 6 (2 %) Not reported: 5 (1 %)
Atrial fibrillation	No: 97 (97 %) Yes: 3 (3 %)	No: 357 (95 %) Yes: 14 (4 %) Not reported: 5 (1 %)
Congestive heart failure	No: 98 (98 %) Yes: 2 (2 %)	No: 363 (97 %) Yes: 5 (1 %) Not reported: 8 (2 %)
Symptoms and complaints		
Insomnia	No: 85 (85 %) Yes: 15 (15 %)	No: 339 (90 %) Yes: 29 (8 %) Not reported: 8 (2 %)
Depression	No: 85 (85 %) Yes: 15 (15 %)	No: 292 (78 %) Yes: 79 (21 %) Not reported: 5 (1 %)
Mood disturbances	No: 66 (66 %) Yes: 34 (34 %)	No: 293 (78 %) Yes: 76 (20 %) Not reported: 7 (2 %)
Morning fatigue	No: 31 (31 %) Yes: 67 (67 %)	No: 217 (58 %) Yes: 151 (40 %) Not reported: 8 (2 %)
Daytime fatigue	No: 55 (55 %) Yes: 42 (42 %) Not reported: 3 (3 %)	No: 93 (25 %) Yes: 276 (73 %) Not reported: 7 (2 %)
Cognitive alterations	No: 45 (45 %) Yes: 54 (54 %) Not reported: 1 (1%)	No: 152 (40 %) Yes: 176 (47 %) Not reported: 48 (13 %)

Morning headache	No: 61 (61 %) Yes: 38 (38 %) Not reported: 1 (1 %)	No: 222 (59 %) Yes: 146 (39 %) Not reported: 8 (2 %)
Nocturia	No: 40 (40 %) Yes: 58 (58 %) Not reported: 2 (2 %)	No: 244 (65 %) Yes: 124 (33 %) Not reported: 8 (2 %)
Gasping and/or choking	No: 43 (43 %) Yes: 56 (56 %) Not reported: 1 (1 %)	No: 183 (49 %) Yes: 185 (49 %) Not reported: 8 (2 %)
Habitual snoring	No: 18 (18 %) Yes: 81 (81 %) Not reported: 1 (1 %)	No: 62 (16 %) Yes: 307 (82 %) Not reported: 7 (2 %)
Witnessed apnea	No: 44 (44 %) Yes: 55 (55 %) Not reported: 1 (1 %)	No: 183 (49 %) Yes: 186 (49 %) Not reported: 7 (2 %)

Caption

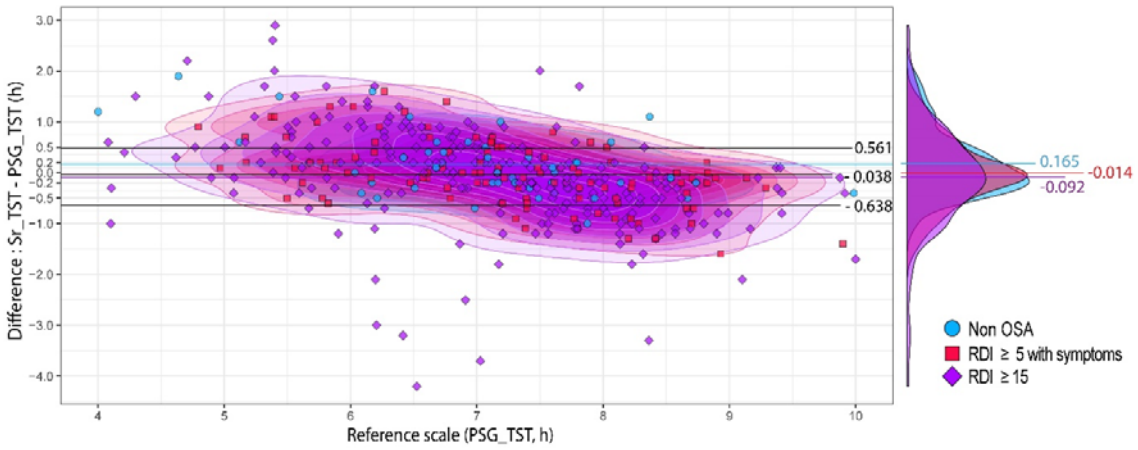
Numerical variables are described as median (interquartile range). Categorical variables were described as frequency (%). Undetermined values were labelled as "Not reported". yrs: years; cm: centimeters; h: hours; n/e: number of events per hour; ESS: Epworth Sleepiness scale; TST: Total sleep time; ArI: Arousal index; RDI: respiratory disturbance index.

eTable 2. Characteristics of interest for sleep mandibular movements (MM) analysis as shown in eFigure 1

Characteristics of interest for sleep mandibular movements (MM) analysis as shown in Figure e-1		
MM label	**Reference**	**Description**
Wake	A	Irregular and unpredictable MM with no oscillation at the breathing frequency (0.15 to 0.60 Hz) during period of 30 s.
Quiet sleep	B	Successive MM with a peak to peak amplitude < 0.3 mm at the breathing frequency (0.15 to 0.60 Hz) during period of 30 s.
Respiratory effort	C	Successive MM with a progressive increase in peak to peak amplitude $\geq$ 0.3 mm at the breathing frequency (0.15 to 0.60 Hz) during period of 30 s.
Arousal	D	Sudden MM with peak to peak amplitude > 1 mm during the respiratory mandibular cycles or brisk MM disrupting the previous respiratory mandibular cycle frequency.

Note: The four labels of interest have been illustrated on Figure e-1 using the reference 'A' (wake), 'B' (quiet sleep), 'C' (respiratory effort). and 'D' (arousal).

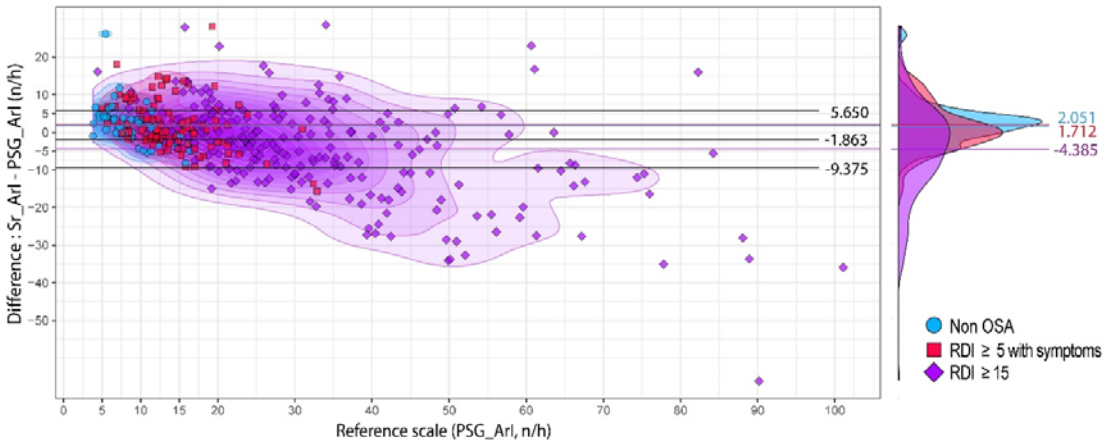
eFigure 3. Bland-Altman evaluation of the agreement between the 2 methods for TST measurement



Caption

Bland-Altman graph showing individual disagreements between PSG_TST and Sr_TST (Y-axis) as a function of PSG_TST values (X-axis); each point represents an individual case. The data were stratified into 3 groups: 1) non-OSA group (blue), 2) patients with PSG_RDI ≥ 5 with clinical symptoms (red) and /or comorbidities and 3) patients with PSG_RDI ≥ 15 (purple).

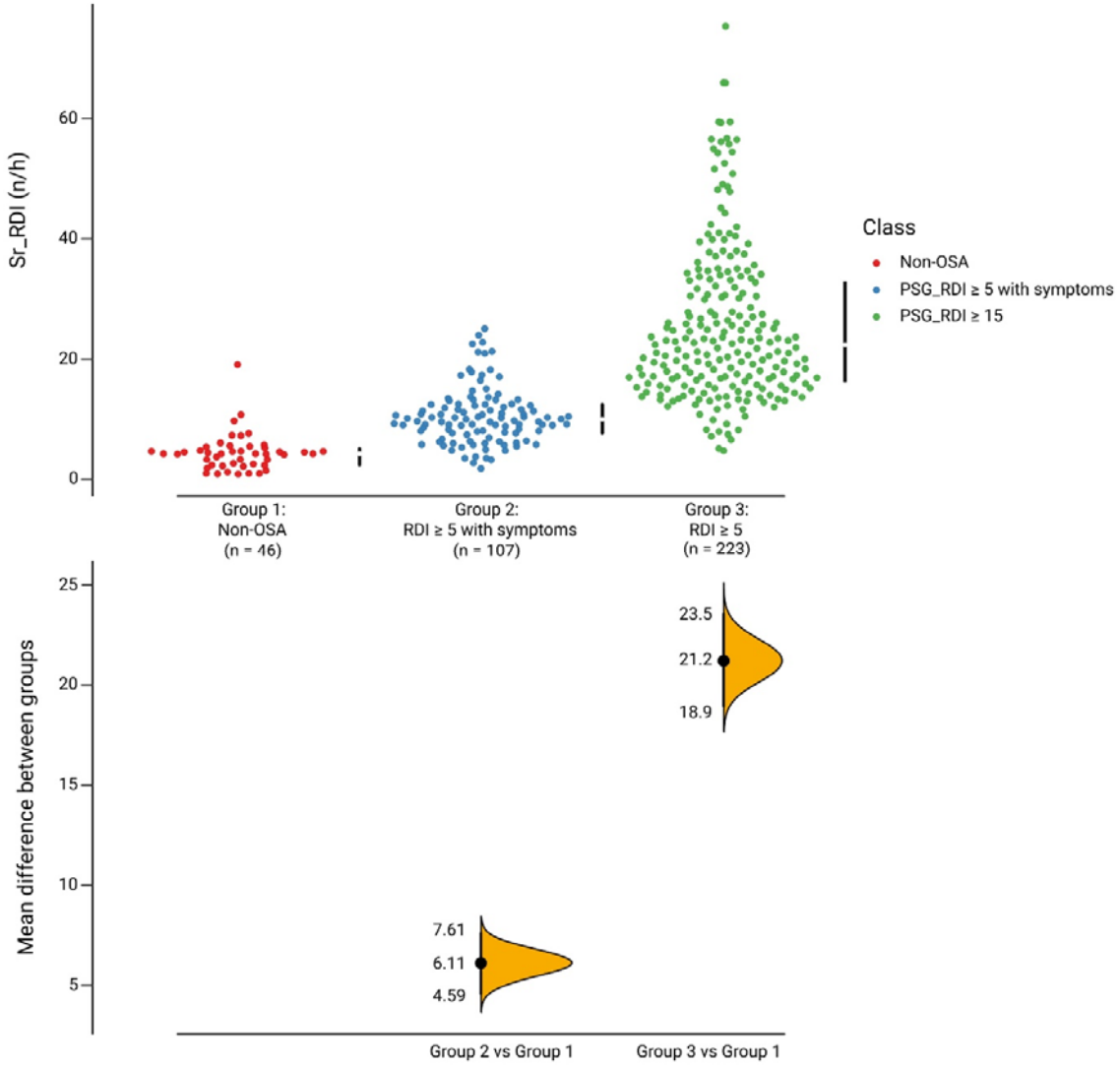
eFigure 4. Bland-Altman evaluation of the agreement between the 2 methods for the arousal index



Caption

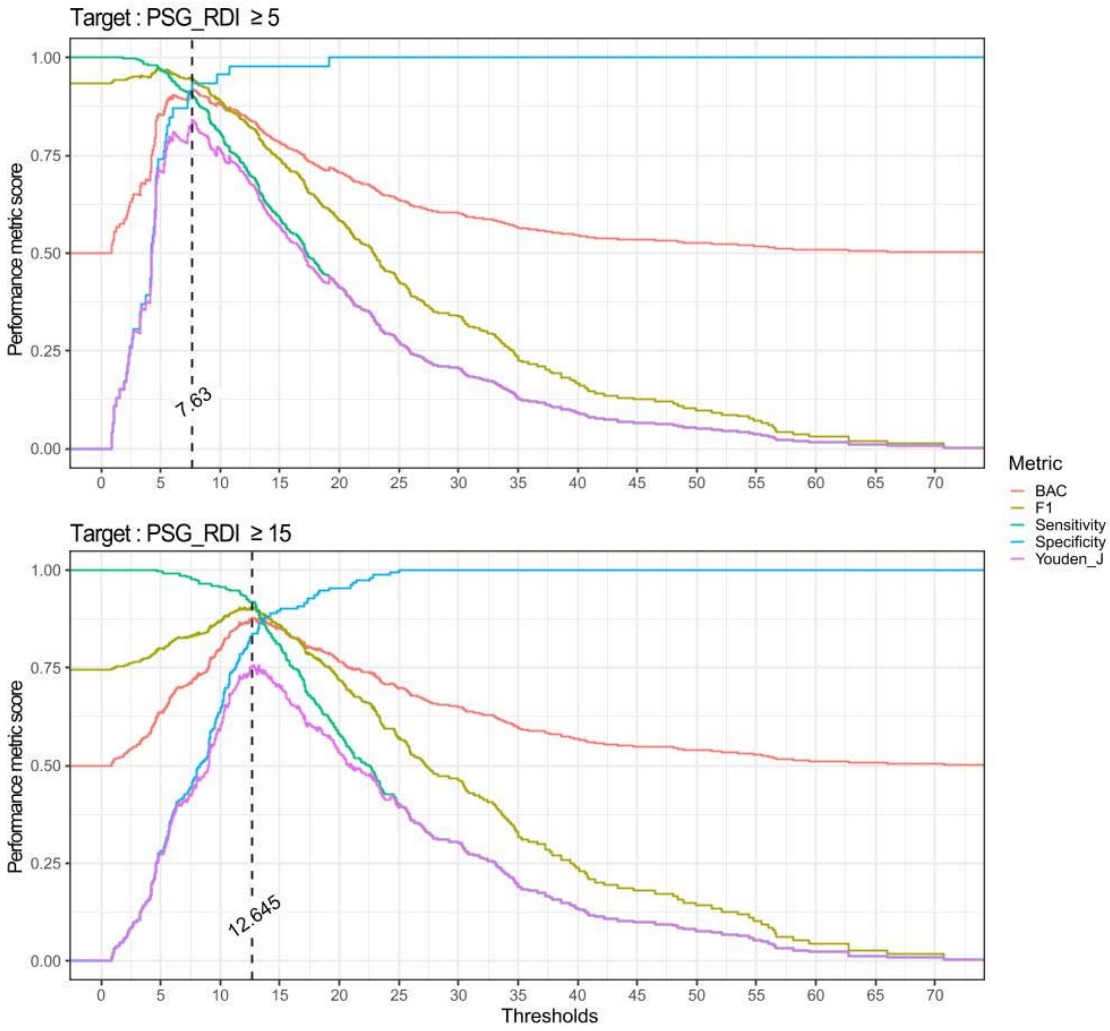
Bland-Altman graph showing individual disagreements between PSG_ArI and Sr_ArI (Y-axis) as a function of the true PSG_ArI values (X-axis); each point represents an individual case. The data were stratified into 3 groups: 1) non-OSA group (blue), 2) patients with PSG_RDI ≥ 5 with clinical symptoms (red) and /or comorbidities and 3) patients with PSG_RDI ≥ 15 (purple). Kernel density estimations plots were used to show the distribution of disagreement within each clinical sub-group. Several horizontal lines indicate the mean difference in the whole sample (black) and within each clinical group (blue, red and purple). Two dashed lines indicate the lower and upper level (Mean ± 1.96 SD) of the disagreement in the whole sample.

eFigure 5. Differences in Sr_RDI values among the 3 clinical groups diagnosed in PSG



Caption: A Gardner-Altman estimation plot showing the pairwise comparisons of Sr_RDI score between the 2 pathological groups: PSG_RDI ≥ 5 with clinical symptoms and/or comorbidities (blue), PSG_RDI ≥ 15 (green) against the non-OSA group (red). The dots represent the distribution of raw data while the yellow Kernel density plots and black dot represent the bootstrapped distribution and 95% confidence interval of the mean difference between 2 comparative groups.

eFigure 6. Variability in diagnostic performance of Sr_RDI across all possible thresholds



Caption: The graph represents 2 binary classification problems aiming to detect the patients with PSG_RDI ≥ 5 (upper panel) or PSG_RDI ≥ 15 (lower panel) using Sr_RDI score (X-axis). In each panel, there are 5 curves corresponding to 5 different statistical metrics that measure the diagnostic performance of the diagnostic rule. These include: sensitivity (or true positive rate, green) = true positive / (true positive + false negative); specificity (or true negative rate, sky blue) = true negative / (true negative + false positive); BAC (pink): balanced accuracy = (true positive rate + true negative rate) / 2; F1_score (light green) is defined as harmonic mean between sensitivity and PPV; Youden's J index (purple) is defined as (sensitivity + specificity) – 1. These metrics could vary from 0 (the worst value) to 1 (the best value) in function of the diagnostic threshold. The curves show the variability of the performance metrics and the trade-off between them over all possible thresholds on Sr_RDI scale. The dotted vertical lines indicate the optimal threshold for each target, determined by the optimization of Youden index.

eAppendix. Sample size calculation

The sample size estimation aimed at determining the minimum number (n) of subjects needed to detect a mean difference δ (delta) and a standard deviation of difference σ (sigma) for a particular population (subgroups with RDI < 15 or ≥ 15) , with a statistical power (or 1 – β) of 0.9 and one-sided type I error probability α (significance threshold) of 0.01, using Dupont's formula (1990):

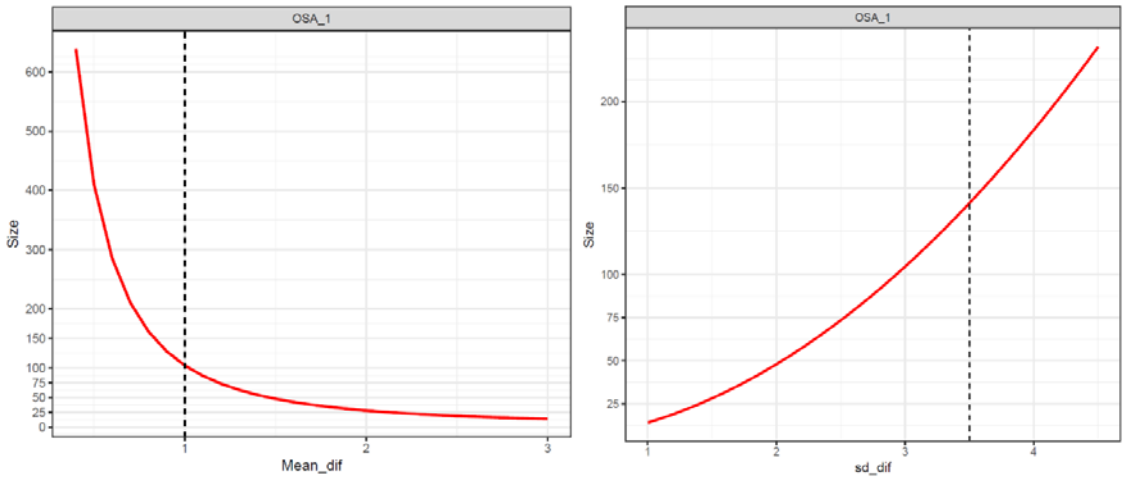
$$n = \frac{(t_{n-1,\alpha/2} + t_{n-1,\beta/2})^2}{d^2}$$

Where d is the Cohen's d effect-size, determined as d = delta/sigma, and $t_{v,p}$ is the quantile of Student's t distribution with v degrees of freedom and probability p.

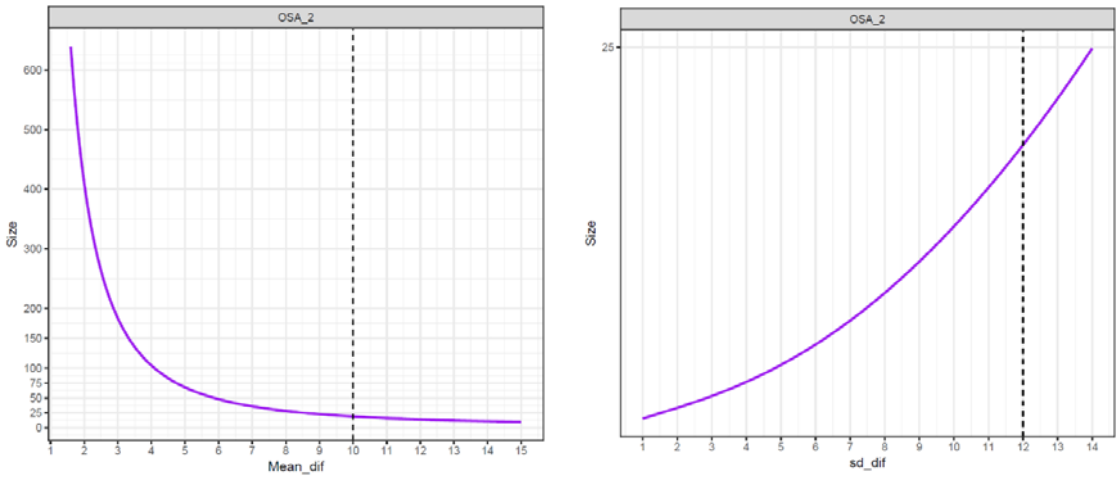
Prior pilot studies of the MM monitoring technology allowed us to expect the disagreements in RDI of 1.0 ± 3.5 and 10.4 ± 12.1 for two classes (RDI < 15 or ≥15). These were used as priors for delta and sigma parameters in our sample size estimation.

But, to ensure robustness of the estimations, we ran sample size simulations covering all possible values of delta and sigma for each subgroup:

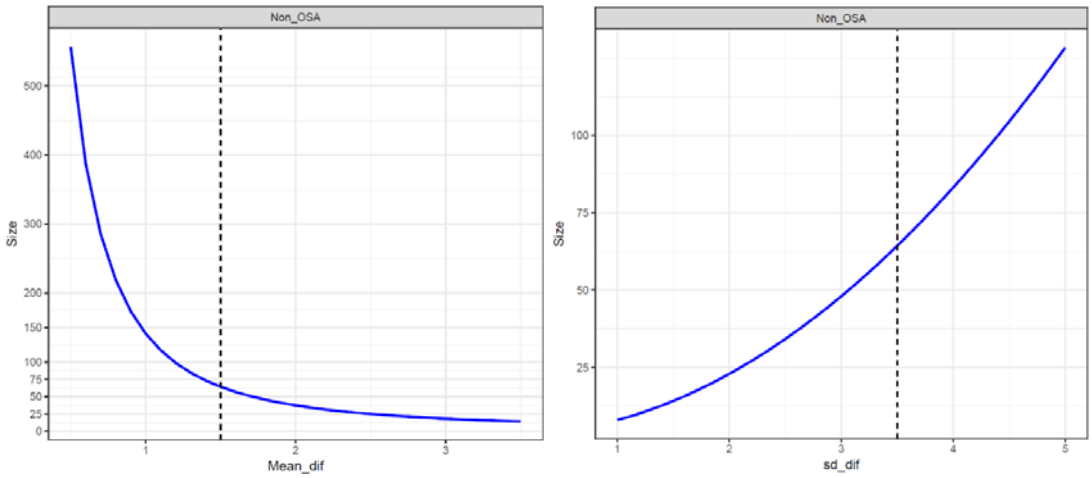
A. For the group of RDI < 15, we need about 100 to 120 patients to detect a mean difference around 1 and standard differences of 3.5, with a statistical power of 0.9 and significance threshold of 0.01



B. For the group of RDI ≥ 15, we need 25 to 200 patients to cover a mean difference of 3 to 13 and standard differences of 1 to 14, with a statistical power of 0.9 and significance threshold of 0.01



C. As we also needed Non-OSA subjects with RDI values lower than 5, we decided to perform a specific simulation for this value range, by assuming that the subjects would have the same mean difference of 1 and standard differences of that could not be higher than 3. Based on this simulation, we would need about 50 to 75 Non-OSA subjects.



Our estimated sample size thus ranged from a minimum of 175 to a maximum of 395. Thus, our sample size of 376 subjects was large enough to meet such requirement.

References

Dupont, W.D. Power and sample size calculations. Controlled Clinical Trials, 1990, 11(2):116-128.

eTable 3. Comparison with alternate techniques for automated diagnosis of OSA

					≥ 5 n/h			≥ 15 n/h		
Device	**Ref**	**Characteristics**	**SCOPER**	**#**	**AUC**	**Sen**	**Spe**	**AUC**	**Sen**	**Spe**
Sunrise	-	Fully automated Wireless	$S_3P_2E_4R_5$	376	95	91	94	93	92	84
WristOx 3100	1	Fully automated Wireless	O_1	154	-	89	94	-	88	90
Remmers sleep recorder	2	Fully automated	$O_{1x}P_2R_5A_x$	94	85	75	81	91	63	96
APNiA	3	Fully automated	$S_3C_4O_1P_2E_4R_2$	28	96	88	73	97	70	94
NOX T3	4	Fully automated	$S_3O_{1x}P_2E_1R_2A_x$	32	-	100	70	-	92	85
Somnocheck	5	Semi-automated	$C_4O_{1x}P_2R_2A_x$	121	96	96	65	91	81	83
SleepView	6	Semi-automated	$C_4O_{1x}R_2A_x$	93	92	80	95	92	87	85
ARES Unicorder	7	Semi-automated	$S_3C_4O_2P_2E_3R_2A_1$	92	-	98	84	-	92	95
Stardust II	8	Semi-automated	$C_4O_{1x}P_2E_4R_2$	80	95	95	62	95	86	78
Watch PAT200	9	Semi-automated Wireless	$S_3C_2O_{1x}$	75	91	96	43	92	92	77
Sonomat	10	Semi-automated Contactless	E_4A_2	60	94	94	77	97	88	91

Caption: The following table compares the performances and characteristics of the presently reviewed device with 10 comparators, reviewed in Mendonça et al. (2018). The following criteria were used to select comparators: a) the level of evidence is qualified as I or II, as per Flemons et al.'s criteria (2003); b) the quality of evidence is qualified as 'a' or 'b', as per Flemons et al.'s criteria (2003); c) the analysis for the test method is at least semi-automated; d) the device is commercially available; e) polysomnography is the reference method;

f) similar cut-offs were tested (5 and 15 n/h) and sensitivity and specificity are reported. The application of these criteria led to the selection of 10 devices. Unlike most of the reviewed comparators, the Sunrise system relies on a single-channel recording. The system was validated on the largest sample and supported by fully automated analysis with no manual intervention. Abbreviations: Ref, reference; #, number of patients in the population studied; AUC, area under the ROC curve; Sen, sensitivity; Spe, specificity.

eReferences

1. Nigro C, Aimaretti S, Gonzalez S, Rhodius E. Validation of the WristOx 3100™ oximeter for the diagnosis of sleep apnea/hypopnea syndrome. Sleep Breath 2009;13(2):127e36.

2. Jobin V, Mayer P, Bellemare F. Predictive value of automated oxygen saturation analysis for the diagnosis and treatment of obstructive sleep apnoea in a home-based setting. Thorax 2007;62(5):422e7.

3. Duan-Cantolla J, Almeida G, Guere~nu V, Rotaeche L, Alkhraisat M, Carro J, et al. Validation of a new domiciliary diagnosis device for automatic diagnosis of patients with clinical suspicion of OSA. Respirology 2017;22(2):378e85.

4. Cairns A, Wickwire E, Schaefer E, Nyanjom D. A pilot validation study for the NOX T3 portable monitor for the detection of OSA. Sleep Breath 2014;18(3):609e14.

5. Oliveira A, Martinez D, Vasconcelos L, Gonçalves S, Lenz M, Fuchs S, et al. Diagnosis of obstructive sleep apnea syndrome and its outcomes with home portable monitoring. Chest 2009;135(2):330e6.

6. Zou, J., Meng, L., Liu, et al. Evaluation of a 2-Channel Portable Device and a Predictive Model to Screen for Obstructive Sleep Apnea in a Laboratory Environment. Respiratory Care 2015; 60(3): 356.

7. Ayappa I, Norman RG, Seelall V, Rapoport DM. Validation of a self-applied unattended monitor for sleep disordered breathing. J Clin Sleep Med. 2008;4(1):26–37.

8. Santos-Silva R, Sartori D, Truksinas V, Truksinas E, Alonso F, Tufik S, et al. Validation of a portable monitoring system for the diagnosis of obstructive sleep apnea syndrome. Sleep 2009;32(5):629e36.

9. Garg N, Rolle A, Lee T, Prasad B. Home-based diagnosis of obstructive sleep apnea in an urban population. J Clin Sleep Med 2014;10(8):879e85.

10. Norman N, Middleton S, Erskine O, Middleton P, Wheatley J, Sullivan C. Validation of the sonomat: a contactless monitoring system used for the diagnosis of sleep disordered breathing. Sleep 2014;37(9):1477e87.

11. Flemons W, Littner M, Rowley J, Gay P, AndersonW, Hudgel D, et al. Homediagnosis of sleep apnea: a systematic review of the literature. Chest 2003;124(4):1543e79.