Online Data Supplement

Identifying the Obstructive Sleep Apnea Phenotype Responsive to Supplemental Oxygen Therapy

Scott A Sands[1,2*], Bradley A Edwards[1,3,4], Philip I Terrill[5], James P Butler[1], Robert L Owens[1,6], Luigi Taranto-Montemurro[1], Ali Azarbarzin[1], Melania Marques[1], Lauren B Hess[1], Erik T Smales[1], Camila M de Melo[1], David P White[1], Atul Malhotra[1,6], Andrew Wellman[1]

[1]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital and Harvard Medical School, Boston, USA.
[2]Department of Allergy, Immunology and Respiratory Medicine and Central Clinical School, The Alfred and Monash University, Melbourne, Australia.
[3]Sleep and Circadian Medicine Laboratory, Department of Physiology Monash University, Melbourne, VIC, Australia.
[4]School of Psychological Sciences and Monash Institute of Cognitive and Clinical Neurosciences, Monash University, Melbourne.
[5]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia.
[6]Division of Pulmonary, Critical Care and Sleep Medicine, University of California San Diego, La Jolla CA, USA.

Supplemental Methods

Participants

Detailed characteristics of participants are described in Table S1.

A minimum AHI of 20 events/hr was chosen to minimize the possibility that a clinically-important response to treatment (50% reduction in AHI) could occur by chance due to night-to-night variability (SD approximately 9 events/hr [S1]).

Power analysis

The study was powered (alpha = 0.05, power ~80%) to find a 1.0 SD difference in the response (by 33±33% percent reduction in AHI) between high loop gain and low loop gain subgroups (high: LG_1>0.7, i.e. "hypersensitivity") based on a prevalence of 1:2 (high:low). Ultimately the observed difference between groups was just +10.5±37.6% [mean±SD] (95%CI: -16 to 37%), a 0.28 SD difference; i.e. the best estimate of the difference was small and there was no more than a 37% greater reduction in patients with higher loop gain (using LG_1).

Percent change in AHI was chosen over absolute change in AHI because the percent change is typically least-strongly correlated with the baseline (sham) AHI (here: r=0.07 [p=0.7] versus r=0.4 [p=0.016] for absolute reduction). If we had used absolute reduction in AHI as the outcome variable, the observed difference would have been borderline non-significant (p=0.051), with a group difference of 13.3±18.9 events/hr [mean±SD] (95%CI: 0.1 to 26.6 events/hr). However, this difference can be explained by an increased baseline AHI in the high LG_1 group; after adjusting for baseline AHI the difference between groups became +6.0 events/hr (95%CI: -9.2 to 21.2 events/hr).

Power for multivariable analysis. Given the absence of appropriate existing data for a formal power analysis, we estimated that approximately (10×M)+10=50 subjects would be necessary (56 were used) to build a prediction model that would use at least M=4 terms. Robustness was assessed based on the loss in predictive value via cross-validation. We emphasize that the primary goal of the multivariable analysis was not to show that each trait contributes significantly to responses (there was no minimum detectable odds ratio). Rather, the objective was to define two subgroups that would have significantly different responses (after cross-validation); since these subgroups need to be powered to detect a difference in response, the power considerations are the same as for the initial analysis, i.e. 36 patients would provide ~80% power to detect a difference in the reduction in AHI by 33±33% (1 SD). Ultimately the difference was 46.3±30.3% [mean±SD] (95%CI: 24.8 to 67.7 %), or 1.5 SD, which was largely unchanged after adjusting for baseline AHI (46.8%; 95%CI: 26.0 to 67.6%). The difference in absolute reduction in AHI between subgroups was also significant: 23.2±16.4 events/hr [mean±SD] (95%CI: 11.6 to 34.7 events/hr), or 1.4 SD; adjusting for baseline AHI had a minimal effect (24.2 events/hr; 95%CI: 14.6 to 33.8 events/hr).

Procedure

Studies were performed a week apart to facilitate between-study consistency of work and lifestyle factors that might contribute to sleepiness or blood pressure levels (e.g. exercise, diet, caffeine use). Participants were asked to keep routine medication use consistent between studies.

Medications for hypertension, when applicable, were administered at home on the morning prior to the overnight study, and then were not taken until after morning blood pressure measurements were made.

At arrival (~7pm), seated blood pressure measurements were made that served to familiarize participants with the measurement experience, reducing the chance of possible "first measurement" effects influencing the evening blood pressure values.

After study completion, patients were asked if they knew which night was oxygen and which was sham: 26% guessed correctly, 20% guessed incorrectly, and 54% were unsure (signed rank test $P=0.7$; correct=1, unsure=0, incorrect=−1) indicating that subjects were effectively blinded.

Ventilatory control tests (dynamic inspired CO_2) were also performed before and after sleep on both nights [S2]; data are not provided here to focus on polysomnographic predictors.

Polysomnographic setup

Care was taken to ensure high quality nasal pressure signals were recorded: a cannula without evidence of mechanical damping effects was selected [prongs 3.5 mm diameter] (Hudson RCI standard "over the ear" cannula, Teleflex, Morrisville NC). Cannulas were secured to the face with tape to minimize displacement (Tegaderm, 3M, Maplewood MN); signal amplification was DC coupled to preserve the baseline (Validyne, Northridge CA) and unfiltered signals were exported for analysis.

Hypopneas were scored based on a 30% reduction in airflow, avoiding the desaturation criterion given the use of supplemental oxygen.

An epiglottic pressure catheter (Millar Instruments, Houston TX) was used to adjudicate central versus obstructive hypopneas to confirm obstructive pathophysiology.

EEG arousals were scored using standard criteria (≥3-s change in EEG frequencies θ, α, β). All patients analyzed also had AHI>20 by standard hypopnea criteria (3% desaturation or arousal) [S3]. At baseline, events not associated with desaturation or arousal made up just 7.7±8.6% (mean±S.D.) of the scored events.

Quantifying the pathophysiological traits using polysomnography

Eupneic ventilation during OSA is inferred from the mean ventilation for each window of data on the basis that mean PCO_2 is not greatly deranged during this time. This assumption did not adversely affect chemical drive and loop gain measurements in our model simulations [S4]. Eupneic ventilation on CPAP also compares closely with the mean value of ventilation during sleep in patients with OSA [S5].

To construct each phenotypic summary plot of ventilation versus ventilatory drive during sleep, the following process was automated:

1. Values for ventilation and ventilatory drive were tabulated for each breath that appeared during windows of non-REM sleep [S4].
2. Breaths were also labelled based on whether or not a scored EEG arousal was present within the breath (from start inspiration to end expiration). Breaths within an arousal or ≤2 breaths after an arousal ended (after sleep onset) were excluded from analysis to minimize the possibility of including data influenced by wakefulness in the assessment of behavior during sleep.
3. Ventilatory drive data were sorted and divided into 10 groups or bins (deciles). For each decile, the median ventilation was measured and plotted against the median ventilatory drive for each decile.
4. Linear interpolation was used between deciles to find a) the value of ventilation at eupneic ventilatory drive ($V_{passive}$), and b) the value of ventilation at the arousal threshold (called V_{active}); compensation is given by V_{active} minus $V_{passive}$.

Definition of predictive model

The term "model" here is used to indicate a classifier plus the necessary coefficients/cutoffs for predicting responders/non-responders: Univariable models consist of a cutoff alone (threshold). Multivariable models comprise a set of selected features (phenotypic variables), a set of coefficients, as well as a cutoff. In all cases (univariable and multivariable), we sought to maximize sensitivity and specificity [S6]. Also in all cases, we employed leave-one-out cross validation to provide generalizable measures of performance.

Assessment of predictive value

Cross-validation. When assessing the performance or predictive value of a model (defined above) that has been developed (trained) on available data, it is best practice to use unseen data for model validation (testing) to prevent over-estimation of the predictive value for future applications. With the modest sample size available in our study (i.e. N=9 responders), use of a fully-separate dataset for development and validation was considered inefficient use of available data. Rather, we used a common procedure called (leave-one-out) "cross-validation". This procedure was used throughout the study for univariable and multivariable analyses (except for LG_1, the *a priori* primary predictor).

The procedure first involves developing a predictive model using all patients. To test the performance of this model, the entire process of developing the model was repeated but using all subjects except one who was left out. This "modified" predictive model was then tested on the unseen individual who was left out; we recorded the outcome of the prediction (true positive, false negative, etc.). This leave-one-out process was then repeated N times (here, N=36). A new modified model was developed each time a new individual was left out. Each time the model changes slightly, but we can be certain of the independence of the development and validation data. Of note, the actual model presented is that which is based on all subjects, since this is the best model we can present for future use based on all available data.

Multivariate model analysis

We chose logistic regression as a simple "machine learning" tool that is easily interpreted. Quadratic model terms were used given the observation of interactions between variables, i.e. a simpler linear model was not sufficient to explain responses. The process was designed to be simple and transparent. In brief, the process for identifying the model was as follows:

1. Terms are initially included in the model: all variables (N=4), their squares (N=4) and interaction terms (N=6). The number of terms starts with M=14.
2. A logistic regression model was fit to the data using M terms.
3. The term with the highest p-value (Wald test) was removed (if p>0.157) [S7-9], and Step 2 was repeated for the remaining terms.
4. Once no further terms were removed, a logistic regression model cutoff was selected to maximize sensitivity plus specificity (receiver operating characteristic analysis) [S10].

Additional considerations. Because two highly-correlated measures of loop gain were available (LG_1, LG_n), we tested the model performance with LG_1 and LG_n separately; LG_n was consistently the most predictive of these two variables and therefore chosen over LG_1. Vpassive was forced into the model because of (1) expert knowledge that collapsibility should contribute to responses [S11], and (2) its removal varied the model coefficients (betas) for other key traits (loop gain and compensation) considerably (by >25%). Model weights were used to balance the influence of patients per subgroup. To estimate the performance of this model when applied to unseen data, we repeated the above procedure using leave-one-out cross-validation (described above).

Including additional published data to build a robust multiple logistic regression model. Data from a previous study [S12] were used to help build the multiple regression model, but we did not seek to test outcomes in the additional individuals. Hence, during cross-validation, we used N=55 (20+36-1) patients to develop a regression model to predict the outcome for each of the 36 patients in the current study. There were, however, some differences in study design between the current and previous one: Edwards et al used the same inspired oxygen concentration as the current study but in fact tested the combination of oxygen and 3 mg eszopiclone versus sham/placebo: However, we argue that it is likely that eszopiclone had relatively little impact on the AHI in that study. Recent data [S13] illustrated that a similar dose of zopiclone (i.e. 3.75 mg of eszopiclone, plus 3.75 mg of its inactive stereoisomer) had no impact on AHI overall and none of the 8 patients with AHI>20 events/hr exhibited more than a 20% reduction in AHI with this treatment.

Sensitivity analysis also proved that the additional data from Edwards et al. were useful in building a robust multivariable regression model. Without these additional data, (1) the quadratic model was underspecified and could not be used, (2) a linear model identified all four traits as contributors but no parameter was significant suggesting findings may not be robust (indeed poor compensation tended to predict a positive outcome, which is likely to be erroneous), (3) the linear model performance was similar before cross-

validation but after cross-validation was slightly weaker (ΔAHI = 53±8% versus 15±7%, p=0.002; PPV = 54±14% [p=0.04], NPV = 91±6% [p<0.006], accuracy = 78±6% [p=0.003]) indicating reduced robustness compared with the inclusion of the additional data. However, findings relating to improved secondary outcomes in the predicted responder subgroup were all upheld.

Statistical analysis

Statistical analyses were performed using MATLAB (Statistics and Machine Learning Toolbox, Mathworks, Natick MA, USA).

$V_{passive}$ and arousal threshold data failed normality tests; their skewness were therefore minimized using square root transforms centered around the value of 100% (see manuscript for equations).

Adjustments were not made for multiple secondary outcomes; all outcomes assessed were presented regardless of significance. Exploratory outcomes that were significantly improved (e.g. percentage time in stage 1 non-REM sleep; Table S1) were not emphasized.

Use of clinical variables. A variety of clinical variables are available from which one might potentially build a separate predictive model that does not require the use our endophenotype traits. While we consider that such an effort would be highly-valuable, we caution that the use of endophenotype traits has a distinct advantage: there is a highly-plausible mechanistic basis for the association with the response to treatment. The use of patient characteristics that have little-to-no mechanistic basis is challenging statistically (spurious associations are expected when using a large number of variables in a relatively small dataset) and thereby requires a far greater number of patients. There is also the concern that any change in population characteristics (e.g. age, race) would likely require recalibration of the predictive model.

Supplemental References

S1. White LH, Lyons OD, Yadollahi A, Ryan CM, Bradley TD. Night-to-night variability in obstructive sleep apnea severity: relationship to overnight rostral fluid shift. *J Clin Sleep Med* 2015: 11(2): 149-156.

S2. Sands SA, Mebrate Y, Edwards BA, Nemati S, Manisty CH, Desai AS, Wellman A, Willson K, Francis DP, Butler JP, Malhotra A. Resonance as the Mechanism of Daytime Periodic Breathing in Patients with Heart Failure. *Am J Respir Crit Care Med* 2016.

S3. Berry RB, Budhiraja R, Gottlieb DJ, Gozal D, Iber C, Kapur VK, Marcus CL, Mehra R, Parthasarathy S, Quan SF, Redline S, Strohl KP, Davidson Ward SL, Tangredi MM, American Academy of Sleep M. 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.

S4. Terrill PI, Edwards BA, Nemati S, Butler JP, Owens RL, Eckert DJ, White DP, Malhotra A, Wellman A, Sands SA. Quantifying the ventilatory control contribution to sleep apnoea using polysomnography. *Eur Respir J* 2015: 45(2): 408-418.

S5. Sands SA, Terrill PI, Edwards BA, Taranto Montemurro L, Azarbarzin A, Marques M, de Melo CM, Loring SH, Butler JP, White DP, Wellman A. Quantifying the Arousal Threshold Using Polysomnography in Obstructive Sleep Apnea. *Sleep* 2018: 41(1).

S6. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950: 3(1): 32-35.

S7. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. *Source code for biology and medicine* 2008: 3: 17.

S8. Heinze G, Dunkler D. Five myths about variable selection. *Transplant international : official journal of the European Society for Organ Transplantation* 2017: 30(1): 6-10.

S9. Dunkler D, Plischke M, Leffondre K, Heinze G. Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. *PLoS One* 2014: 9(11): e113677.

S10. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics* 2006: 7: 91.

S11. Eckert DJ, White DP, Jordan AS, Malhotra A, Wellman A. Defining phenotypic causes of obstructive sleep apnea. Identification of novel therapeutic targets. *Am J Respir Crit Care Med* 2013: 188(8): 996-1004.

S12. Edwards BA, Sands SA, Owens RL, Eckert DJ, Landry S, White DP, Malhotra A, Wellman A. The Combination of Supplemental Oxygen and a Hypnotic Markedly Improves Obstructive Sleep Apnea in Patients with a Mild to Moderate Upper Airway Collapsibility. *Sleep* 2016: 39(11): 1973-1983.

S13. Carter SG, Berger MS, Carberry JC, Bilston LE, Butler JE, Tong BK, Martins RT, Fisher LP, McKenzie DK, Grunstein RR, Eckert DJ. Zopiclone Increases the Arousal Threshold without Impairing Genioglossus Activity in Obstructive Sleep Apnea. *Sleep* 2016: 39(4): 757-766.

Supplemental Table

Table S1. Characteristics and Impact of Treatment

Characteristic	All patients (N=36)		Responders* (N=9)		Non-Responders (N=27)		Predicted Responders** (N=13)		Pred. Non-Responders (N=23)	
	Sham	Oxygen	Sham	Oxygen	Sham	Oxygen	Sham	Oxygen	Sham	Oxygen
Demographics										
Age (years)	55±2		53±4		55±2		53±3		55±3	
Sex (M:F)	26:10		6:3		20:7		8:5		18:5	
Race (black:white:asian:other)	9:25:0:1		5:3:0:0		4:22:0:1 ¶ ∆		7:6:0:0		2:19:0:1 ¶¶ ∆	
Body mass index (kg/m^2)	31.1±0.7		32.3±1.2		30.6±0.8		31.6±1.0		30.7±0.9	
Neck circumference (cm)	40.6±0.5		40.2±1.0		40.7±0.7		39.9±0.8		41.0±0.7	
Current treatment (CPAP:oral appliance:untreated)	12:2:22		1:0:8		11:2:14		4:0:9		8:2:13	
Medications										
Anti-hypertensives, N (%)	12 (33)		1 (11)		11 (41)		3 (21)		9 (41)	
Proton pump inhibitors, N (%)	5 (14)		1 (11)		4 (15)		1 (7)		4 (18)	
Statins	4 (11)		1 (11)		3 (11)		1 (7)		3 (14)	
Antidepressants/anti-anxiety	4 (11)		1 (11)		3 (11)		1 (7)		3 (14)	
Aspirin	3 (8)		0 (0)		3 (11)		1 (7)		2 (9)	
Levothyroxine	3 (8)		0 (0)		3 (11)		0 (0)		3 (14)	
Zolpidem	1 (3)		0 (0)		1 (4)		0 (0)		1 (5)	
Metformin	1 (3)		0 (0)		1 (4)		0 (0)		1 (5)	
Polysomnography										
Time in bed (min)	421±8	422±11	416±16	446±20	423±10	414±12	421±13	424±17	421±11	421±14
Apnea-hypopnea index† (events/hr)	57.9±3.7	40.5±3.8	56.6±7.7	17.6±4.6	58.3±4.3	48.1±3.9	56.1±5.7	23.9±4.0	58.9±4.9	49.8±4.6
Effect of oxygen (%)	−29.0±6.2 ###		−71.8±4.6 ¶¶ ###		−14.8±4.9 #		−58.6±5.6 ¶¶ ###		−12.3±7.2	
Arousal index† (events/hr)	50.3±3.7	35.9±3.4	46.0±8.1	23.1±3.9	51.7±4.2	40.1±4.0	47.1±6.0	22.5±3.3	52.0±4.7	43.4±4.3
Effect of oxygen (%)	−25.5±5.0 ###		−48.3±3.7 ¶¶ ###		−17.9±6.1 ##		−47.5±6.5 ¶¶ ###		−13.1±5.5 #	
Nadir oxygen saturation (%Hb)	87.1±0.8	97.1±0.4	89.2±1.5	97.9±0.5	86.4±0.9	96.9±0.6	88.2±1.4	97.5±0.6	86.5±1.0	96.9±0.6
Effect of oxygen (%Hb)	10.0±0.8 ###		8.7±1.5 ###		10.4±0.9 ###		9.2±1.0 ###		10.4±1.0 ###	
Stage 1 sleep (% total sleep time)	25.9±3.7	23.8±3.7	22.3±5.1	12.4±4.3	27.1±4.6	27.7±4.5	20.1±4.5	9.6±2.4	29.2±5.1	31.9±4.9
Effect of oxygen (%total sleep time)	−0.3[−8.1 to 3.7]		−10.5[−14.5 to −1.3] ¶		+0.8[−4.8 to 10.6]		−7.2[−14.5 to −0.1] ¶¶ #		1.1[−1.9 to 13.8]	
Additional outcomes										
ΔSystolic blood pressure‡ (mmHg)	+3.0±1.9	−0.8±1.4	+3.4±2.7	−4.1±1.8	+2.9±2.3	+0.3±1.7	+3.2±3.1	−2.5±2.2	+2.9±2.4	+0.2±1.8
Effect of oxygen (mmHg)	−3.8±2.1		−7.6±2.2 ##		−2.6±2.8		−5.8±2.0 #		−2.7±3.1	
ΔDiastolic blood pressure‡ (mmHg)	+4.1±1.5	+0.9±0.9	+6.6±2.8	−0.6±1.3	+3.2±1.8	+1.4±1.1	+6.4±2.5	−0.7±1.2	+2.7±1.9	+1.8±1.2
Effect of oxygen (mmHg)	−3.1±1.5 #		−7.2±2.9 #		−1.8±1.6		−7.1±2.3 ¶ ##		−0.9±1.8	
Slept Better:Same:Worse on oxygen ††	19:9:7 #		5:2:1		14:7:6		9:2:1 ¶ ###		10:7:6	
Alertness, Stanford Sleepiness Scale £	2.0±0.2	2.1±0.2	2.3±0.6	2.3±0.5	2.0±0.2	2.1±0.2	2.2±0.4	2.4±0.4	1.9±0.2	2.0±0.2
Effect of oxygen (points)	0.1±0.2		0.0±0.5		0.1±0.2		0.2±0.4		0.0±0.2	
OSA severity, alternate										
AHI standard scoring, supine non-REM (events/hr)	54.2±3.9	34.8±3.7	49.8±7.6	14.7±4.2	55.6±4.6	41.4±3.9	49.8±5.8	19.6±3.8	56.6±5.1	43.3±4.5
Effect of oxygen (%)	−36.2±5.7 ###		−73.9±4.8 ¶¶ ###		−23.6±5.7 ###		−62.8±5.3 ¶¶ ###		−21.0±6.6 ##	
AHI standard scoring, all states/positions	52.5±3.7	33.5±3.3	44.7±5.6	16.1±3.5	55.1±4.5	39.3±3.6	46.1±4.8	20.9±3.5	56.1±5.0	40.6±4.1
Effect of oxygen (%)	−37.5±4.4 ###		−66.1±5.5 ¶¶ ###		−28.0±4.3 ###		−56.8±4.9 ¶¶ ###		−26.7±5.2 ###	

Values are mean±S.E.M. or median[interquartile range]. *Responders are defined by a ≥50% reduction in apnea-hypopnea index. **Predicted responders are based on the cross-validated logistic regression model analysis. †Reported during non-REM supine

sleep. ‡Morning minus evening values are taken to reflect sleep apnea burden (supine). For 'Polysomnography' and 'Additional outcomes', statistical comparisons are shown for the "Effect of oxygen'. ¶P<0.05, ¶¶P<0.01, ¶¶¶P<0.001 responders versus non-responders. #P<0.05, ##P<0.01, ###P<0.001 oxygen vs sham. ΔFisher exact test (Black vs not Black). ††Not collected in one individual (responder) due to >1 month between studies (rescheduling difficulties); statistical differences were compared using ranks: Better=+1, Same=0, Worse=−1. £Taken >30 mins after lights on. Medication use was unchanged prior to each overnight study and there were no statistically differences between subgroups; antihypertensives included hydrochlorothiazide, lisinopril, losartan, labetolol, atenolol, amplodipine, verapamil, doxasozin; antidepressants and anti-anxiety medications included selective serotonin reuptake inhibitors and aripiprazole. "Standard scoring" denotes the definition of hypopneas based on ≥3% oxygen desaturation or arousal.

Table S2: Two-trait simplified logistic regression model for predicting responses to oxygen therapy

Variable	β	SEM	odds ratio*	p	Interpretation
Constant	−0.23	0.41		0.6	
Loop gain	9.73	4.89	2.7	0.046	Higher loop gain→success
$V_{passive}$	7.24	2.54	6.0	0.004	Reduced collapsibility→success
Loop gain × $V_{passive}$	-32.1	6.62	0.43	0.14	Low loop gain and greater collapsibility→failure

The Table describes the results (3 terms) after backward stepwise elimination (p-to-remove=0.157) which began with the two key traits (loop gain [LG_n], $V_{passive}$), an interaction term (included but not significant), and two squared terms (excluded since p>0.157). SEM = standard error of the mean. *Odds ratio describes the increase in likelihood of being a responder per SD increase in each term. Traits were mean-subtracted before application to the regression model: mean Vpassive*=62.8%, mean loop gain [LG_n]=0.42. To promote normality, Vpassive values were square-root transformed around 100% using $y=1+(x-1)^{0.5}$ (n.b. x=1 describes 100%). Patients were considered a "predicted responder" here if $Y = -0.23 + 9.73[\text{Loop gain}] + 7.24[V_{passive}] - 32.1[\text{Loop gain} \times V_{passive}] > 0.25$ (use of this equation requires transformed, mean-subtracted traits). The model included data from Edwards et al. [S12] such that N=56 (36+20). Predictive value (cross-validated) for patients in the current study (N=36): (ΔAHI = 53±7% in predicted responders versus 10±7% in predicted non-responders [p=0.0002]; PPV = 56±12% [p=0.01], NPV = 100±0% [p<0.0001], accuracy = 81±7% [p<0.0001]).

Supplemental Figures

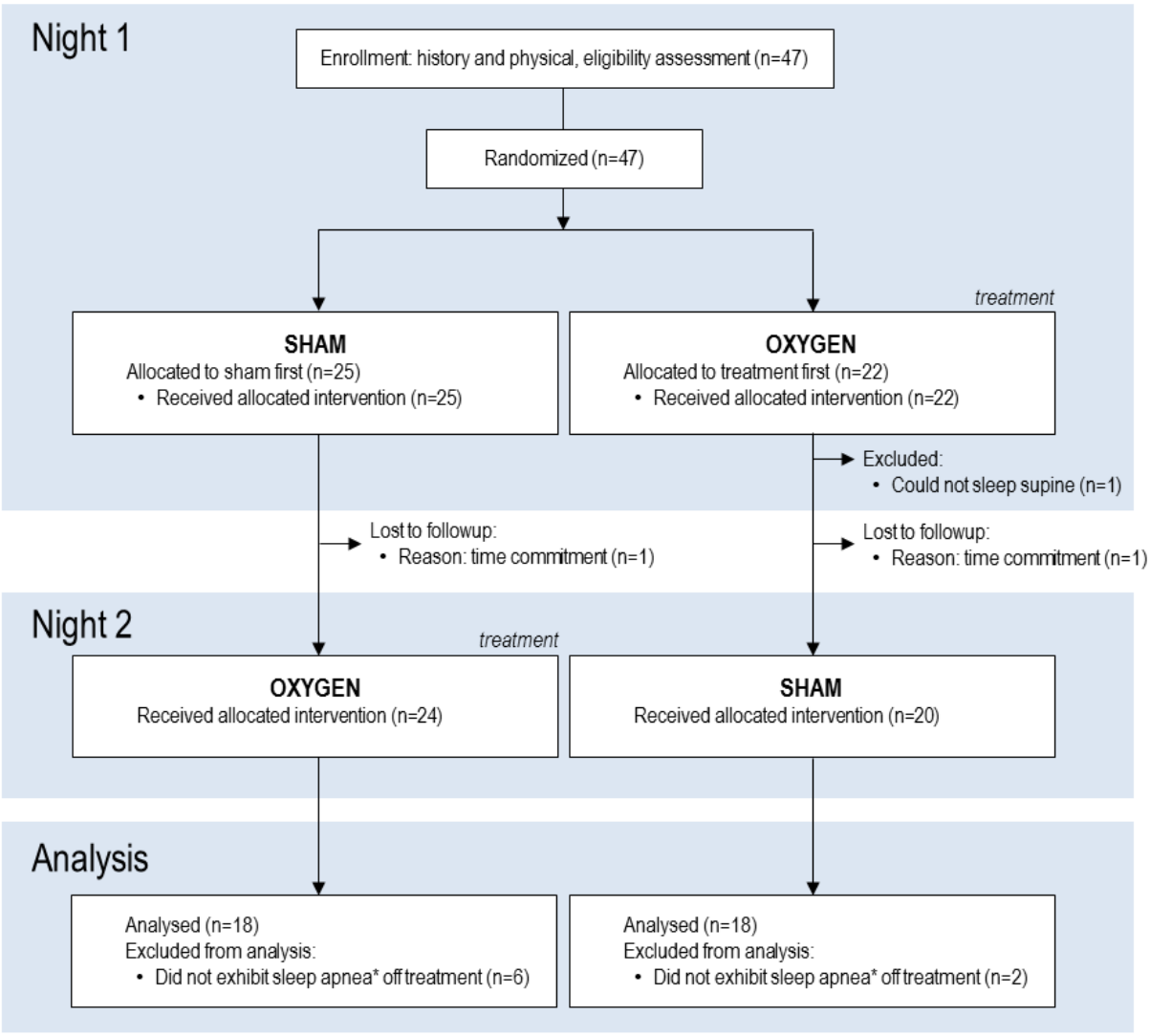

Figure S1. Study flow diagram. 47 patients with diagnosed OSA were randomized to either the sham first or treatment first arms. Randomization was performed using a computer random number generator in blocks of 2. As patients were excluded, new patients filled their slots to ensure equal group sizes for analysed data. Overall, 44 patients completed the study, but 8 patients did not have OSA on their sham study night (*criterion: non-REM AHI>20 events/hour) and therefore could not contribute data for analysis. By design, analysis was *per protocol* rather than *intention to treat*; sham night polysomnograms provided baseline data to measure phenotypic traits for categorizing patients into subgroups as well as for assessing the change in OSA severity (apnea-hypopnea index, AHI) with treatment. Of note, the goal was not to assess the effect of oxygen on OSA in unselected patients *per se*; rather it was to assess the relative reduction in AHI between phenotypic subgroups.

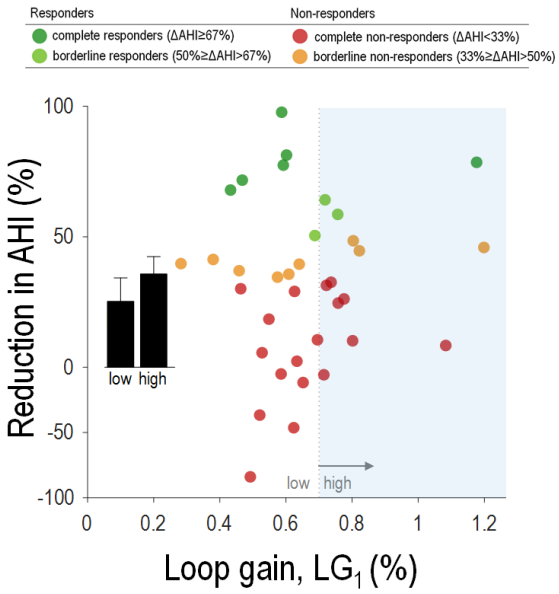

Figure S2. Contrary to our primary hypothesis, patients with high versus low loop gain based on LG_1 (pre-specified cutoff = 0.7, shading illustrates "high") did not show a significantly greater response to supplemental oxygen (reduction in apnea-hypopnea index AHI on treatment versus sham). Bars illustrate the reduction in AHI with treatment in the high vs. low subgroups. LG_1 is the magnitude of the chemoreflex ventilatory drive response to a 1 cycle/min swing in ventilation.

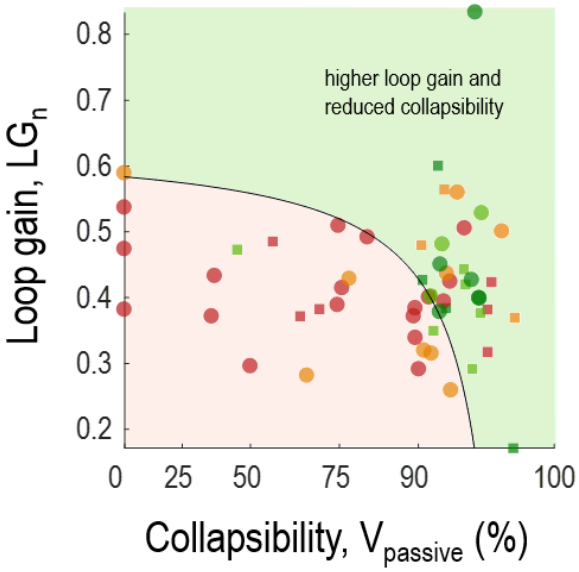

Figure S3. Two-trait simplified model confirming that loop gain and collapsibility ($V_{passive}$) can be combined to predict responses to oxygen therapy. Dots are individual patients (circles are patients from current study, squares are patients from Edwards et al [S12]); colors are consistent with figures in the main manuscript. Shading illustrates the regions of "predicted responders" (green) and "predicted non-responders" (red). See Table S2 for the equation for the logistic regression line.

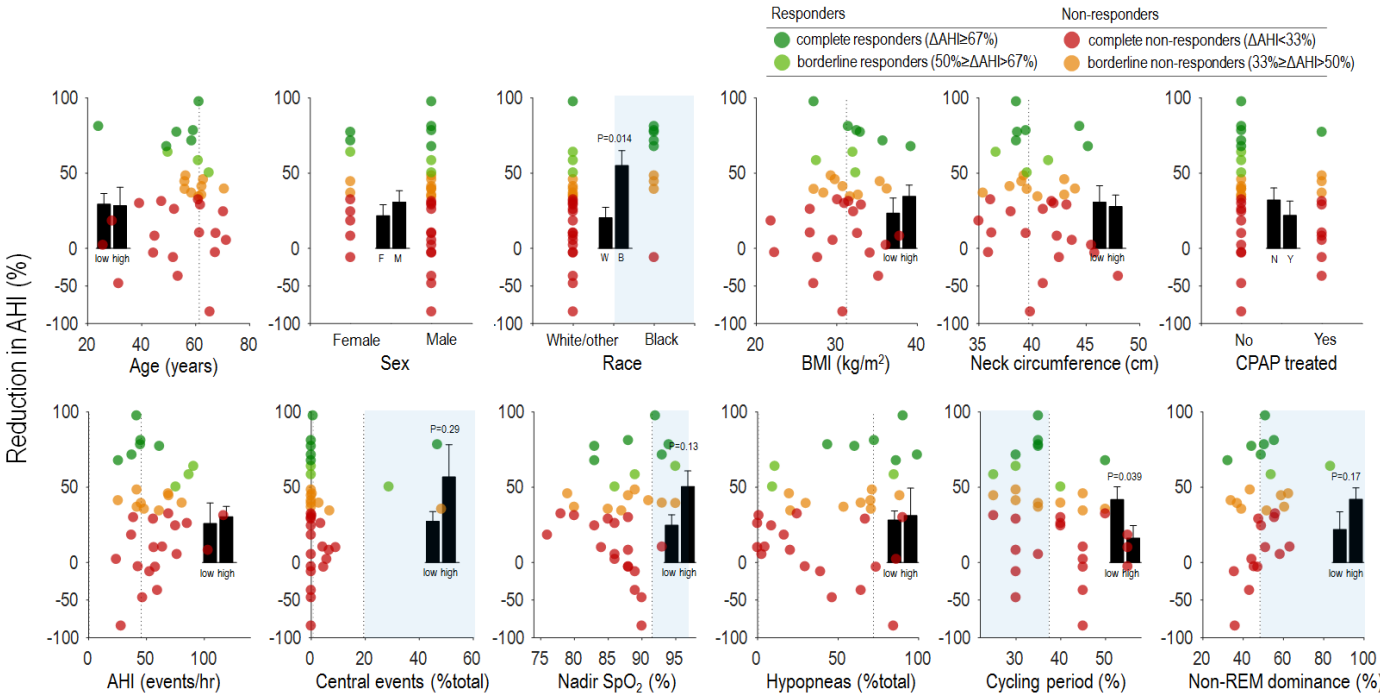

Figure S4. Clinical and other polysomnographic factors and the response to supplemental oxygen. Dashed vertical lines illustrate the optimal cutoffs. Bars illustrate the reduction in apnea-hypopnea index (AHI) with treatment in the subgroups. P-values>0.3 are not shown. There were no very strong predictors of the response to supplemental oxygen. Notably, however, black race significantly predicted a stronger response to treatment, which has not been reported previously. In addition, a faster cycling period (most common time from the end of one respiratory event to the end of the next, i.e. *mode*) was also a significant predictor. Non-significant trends were observed for a greater proportion of central events, a higher nadir oxygen saturation (SpO_2), and a greater non-REM dominance of OSA ($AHI_{non\text{-}REM}/[AHI_{non\text{-}REM}+AHI_{REM}]$; 0 = REM exclusive OSA, 100% = non-REM exclusive OSA, 50% = same OSA severity in non-REM and REM). BMI = body mass index, CPAP = continuous positive airway pressure, REM = rapid eye movement sleep. Note that p-values presented are not adjusted for multiple comparisons and variables were not proposed *a priori* as putative predictors.

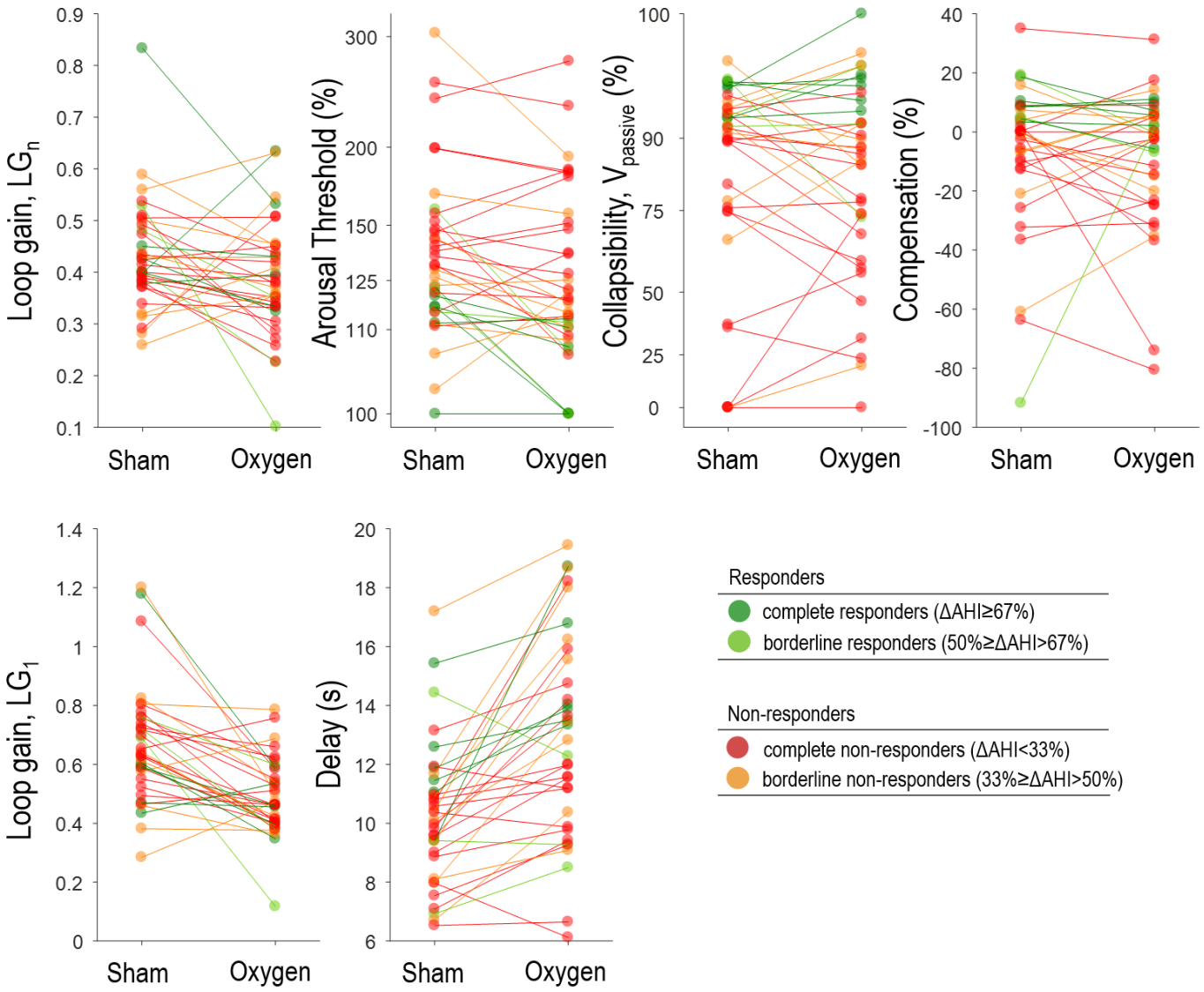

Figure S5. Effect of oxygen therapy on the physiological traits. Summary data are shown in Table 3. *Top:* The four traits causing sleep apnea are shown on sham and on oxygen therapy. Loop gain (LG_n, *instability*) was reduced, consequent to a reduction in feedback sensitivity (LG_1, *Bottom*), and was counteracted somewhat by an increase in estimated delay (*Bottom*). Arousal threshold was also slightly reduced with oxygen, possibly a direct physiological effect of oxygen, but could potentially be consequent to the improvement in sleep apnea severity. There was no evidence of a change in collapsibility or compensation with intervention. 35 patients contributed to these data; 1 individual had insufficient data on oxygen therapy for analysis.