Non-invasive machine learning estimation of effort differentiates sleep-disordered breathing pathology

Hanif, Umaer Rashid; Schneider, Logan D.; Trap, Lotte; Leary, Eileen B.; Moore, Hyatt; Guilleminault, Christian; Jennum, Poul; Sørensen, Helge Bjarup Dissing; Mignot, Emmanuel J. M.

Published in:
Physiological Measurement

Link to article, DOI:
10.1088/1361-6579/ab0559

Publication date:
2019

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Title: Non-invasive machine learning estimation of effort differentiates sleep-disordered breathing pathology

Umaer Hanif$^{1,2,3}$
Logan D. Schneider$^1$
Lotte Trap$^{1,2,3}$
Eileen B. Leary$^1$
Hyatt Moore, IV$^1$
Christian Guilleminault$^1$
Poul Jennum$^2$ *
Helge B. D. Sorensen$^3$ *
Emmanuel J. M. Mignot$^1$ *

$^1$Stanford University Center for Sleep Sciences and Medicine, Psychiatry and Behavioral Medicine, Stanford University, California, U.S.A. (work was performed at this site)

$^2$Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, University of Copenhagen, Glostrup, Denmark

$^3$Biomedical Engineering, Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby, Denmark

* Shared last authors

Name and address of corresponding author:

Emmanuel J. M. Mignot
Contributorship statement:

Concept and design, U.H., L.D.S., H.B.D.S, P.J., and E.J.M.M. Data acquisition, U.H., C.G., E.J.M.M. Analysis, all authors. Interpretation, draft and review, and final approval, all authors. U.H., L.D.S., and E.J.M.M. had full access to the study data and take responsibility for the integrity of the data and accuracy of the analyses.

Sources of support:

Supported by grants from the Feldman Family Foundation, the Klarman Family Foundation, the Lundbeck Foundation, Technical University of Denmark, and Danish Center for Sleep Medicine. The funding institutes played no role in the design and conduct of the study; no role in the collection, management, analysis, or interpretation of the data; and no role in the preparation, review, or approval of the manuscript.

Text word count: 3100
Abstract word count: 250
Figure count: 7
Table count: 4
Reference count: 24

Supplementary Material: 1 Word document containing 14 supplementary figures (Figure S1-S14) and 2 supplementary tables (Table S1-S2)
Abstract

Objective: Obstructive sleep-disordered breathing (SDB) events, unlike central events, are associated with increased respiratory effort. Esophageal pressure (P_{es}) monitoring is the gold standard for measuring respiratory effort, but it is typically poorly tolerated because of its invasive nature. The objective was to investigate whether machine learning can be applied to routinely collected non-invasive, polysomnography (PSG) measures to accurately model peak negative P_{es}.

Approach: 1119 patients from the Stanford Sleep Clinic with PSGs containing P_{es} served as the sample. The selected non-invasive PSG signals included nasal pressure, oral airflow, thoracoabdominal effort, and snoring. A long short-term memory (LSTM) neural network was implemented to achieve a context-based mapping between the non-invasive features and the P_{es} values. A hold-out dataset served as a prospective validation of the algorithm without needing to undertake a costly new study with the impractically invasive P_{es}.

Main results: The median difference between the measured and predicted P_{es} was 0.61 cmH_{2}O with an interquartile range (IQR) of 2.99 cmH_{2}O and 5^{th} and 95^{th} percentiles of -5.85 cmH_{2}O and 5.47 cmH_{2}O, respectively. The model performed well when compared to actual esophageal pressure signal (ρ_{median}=0.581, p=0.01; IQR = 0.298; ρ_{5%} = 0.106; ρ_{95%} = 0.843).

Significance: A significant difference in predicted P_{es} was shown between normal breathing and all obstructive SDB events; whereas, central apneas did not significantly differ from normal breathing. The developed system may be used as a tool for quantifying respiratory effort from the existing clinical practice of PSG without the need for P_{es}, improving characterization of SDB events as obstructive or not.
INTRODUCTION

Sleep-disordered breathing (SDB) is a multidimensional disorder characterized by abnormal breathing during sleep. The consequences of SDB are daytime sleepiness caused by disturbed sleep, as well as increased risk of cardiovascular disease and metabolic disturbances\(^1,2\).

Obstructive sleep apnea (OSA) is by far the most common form of SDB. As a result of increased upper airway resistance, respiratory effort during events increases in an effort to maintain airflow\(^3\). Comparatively, in central sleep apnea (CSA), SDB events are characterized by diminished central nervous system breathing stimulation\(^4\).

The severity of SDB is currently measured using the apnea hypopnea index (AHI), a count of SDB events per hour of sleep. In this measure, SDB events are separated into those with complete (apnea) or partial (hypopnea) reduction of airflow. Apneas are further differentiated into obstructive or central based on the presence or absence of respiratory effort, respectively. However, differentiating hypopneas based on inspiratory effort requires the absence of obstructive characteristics (snoring, airflow limitation, thoracoabdominal dyssynchrony) or the use of an esophageal manometer (\(P_{es}\)), to ascertain the level of effort.\(^5\) Furthermore, an even more subtle form of obstructive SDB, termed upper airway resistance syndrome (UARS), characterized by the presence of respiratory effort-related arousals (RERAs), requires the demonstration of a “\(P_{es\ crescendo}\)” of effort to definitively demonstrate the presence of UARS.\(^6,7\) **Figures 1-3** demonstrate how respiratory signals act during normal breathing, obstructive hypopneas, and obstructive apneas, respectively, including how the “\(P_{es\ crescendo}\)” is witnessed when the subject’s breathing is obstructed.
Figure 1. A 90-second segment of all respiratory signals measured during normal breathing.
**Figure 2.** A 90-second segment of all respiratory signals measured during an obstructive hypopnea. The dashed lines indicate where the event is observed in each signal, whereas the dotted line illustrates the presence of thoracoabdominal dyssynchrony.
Evidence is mounting to suggest that multiple mechanisms contribute to the pathophysiology of SDB, including: anatomical predisposition (e.g., craniofacial structure, fat deposition, etc.); airway dilator reflexes/recruitment; arousal threshold; and centrally-mediated, chemotactic breathing stability (loop gain). With patients exhibiting various combinations of these factors\textsuperscript{8–11}, in addition to mixed central and obstructive pathologies that can impede treatment\textsuperscript{12,13}, proper determination of respiratory efforts is critical to properly identify these phenotypes and individualize therapies.
Respiratory inductance plethysmography (RIP) is often used as an indirect measure of breathing effort derived from thoracoabdominal excursion signals. While the non-invasive nature of the monitor has generated widespread acceptance, there are a number of limitations to the interpretation of the signals, due in part to variations in placement, anatomy, and signal processing. Other attempts to measure effort non-invasively (e.g., chest-wall electromyography) have demonstrated reasonable accuracy in controlled research settings, but in clinical practice are limited by signal quality and variability. Comparatively, esophageal pressure (P\textsubscript{es}) manometry is the gold standard for directly measuring respiratory effort, but tolerability of this invasive procedure limits its utility.

In this study, we adopt a novel approach in the search for a less invasive method to measure respiratory effort. Past studies have attempted to validate non-invasive state-of-the-art techniques against P\textsubscript{es} monitoring with little success. However, the promise of cutting-edge machine-learning techniques’ ability to assess for the presence of sleep-disordered breathing events on even a single, non-invasive signal – electrocardiogram (ECG) – highlights the potential for robust event detection and differentiation. Therefore, the aim of this study was to develop a novel, fully automatic system for predicting peak negative P\textsubscript{es} pressure as a measure for respiratory effort by deriving a composite signal from other routinely-collected, non-invasive, PSG signals, using the world’s largest database of P\textsubscript{es}-integrated polysomnograms.

METHODS

Subjects

1119 subjects with P\textsubscript{es} measurements were selected from the Stanford Sleep Cohort (SSC), a patient-based registry of several thousand PSGs collected at the Stanford Sleep Clinic since 2004.
PSGs were performed due to a suspicion for SDB or UARS, and were staged and scored according to standard criteria. This study protocol was approved by the Stanford IRB (#19207).

**PSG Signal Processing and Feature Extraction**

For an in-depth explanation of PSG signal processing and feature extraction, please see Supplementary Material (*Preprocessing, Feature Extraction, and Normalization* sections). In brief, signals used included nasal airflow (via nasal cannula pressure transducer), oral airflow (via oronasal thermistor), thoracic and abdominal effort (via RIP belts), and snoring (via piezoelectric vibration sensor). Esophageal manometry (P<sub>es</sub>) was performed using a water-filled catheter. Sampling frequency ranges were: 16-256 Hz for nasal pressure and oral airflow, 16-64 Hz for thoracic and abdominal effort, 128-512 Hz for snoring, and 32-64 Hz for P<sub>es</sub>; in order to ensure greater comparability across all PSG recordings and remove redundant information, all signals - except snoring - were down-sampled to 16 Hz (see Supplementary Material). These signals were further corrected for baseline drift, 60 Hz, power-line notch-filtered when necessary, and subjected to flatline detection and removal. Features for each signal were extracted in sliding windows of 10 seconds with an overlap of 50% between consecutive windows (**Table 1**).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Phase-locking value of RIP</td>
<td>Characterizes the phase difference between the thoracic and abdominal effort excursions</td>
</tr>
<tr>
<td>2. CVE of oral airflow</td>
<td>Characterizes the morphology of the respiratory signal by measuring the similarity between the signals and perfectly sinusoidal oscillations</td>
</tr>
<tr>
<td>3. CVE of nasal pressure</td>
<td></td>
</tr>
<tr>
<td>4. CVE of thoracic effort</td>
<td></td>
</tr>
<tr>
<td>5. CVE of abdominal effort</td>
<td></td>
</tr>
<tr>
<td>6. Envelope variance of oral airflow</td>
<td>Characterizes how much the envelope of the respiratory signal changes.</td>
</tr>
<tr>
<td>7. Envelope variance of nasal pressure</td>
<td></td>
</tr>
<tr>
<td>8. Envelope variance of thoracic effort</td>
<td></td>
</tr>
<tr>
<td>9. Envelope variance of abdominal effort</td>
<td></td>
</tr>
</tbody>
</table>
10. Arc length of oral airflow  
11. Arc length of nasal pressure  
12. Arc length of thoracic effort  
13. Arc length of abdominal effort  

Characterizes the degree of breathing, since the arc length of the respiratory signal will be longer as the signal excursions become greater.

14. Area of oral airflow  
15. Area of nasal pressure  
16. Area of thoracic effort  
17. Area of abdominal effort  
18. Area of snoring  

Characterizes the total energy within the respiratory signal.

19. Variance of oral airflow  
20. Variance of nasal pressure  
21. Variance of thoracic effort  
22. Variance of abdominal effort  
23. Variance of snoring  

Characterizes how much the respiratory signal changes, which is particularly useful at the beginning and end of an event because the signal changes rapidly.

24. 2\text{nd} derivative of oral airflow  
25. 2\text{nd} derivative of nasal pressure  
26. 2\text{nd} derivative of thoracic effort  
27. 2\text{nd} derivative of abdominal effort  

Characterizes the curvature of the respiratory signal.

28. Absolute power of oral airflow  
29. Absolute power of nasal pressure  
30. Absolute power of thoracic effort  
31. Absolute power of abdominal effort  
32. Relative power of oral airflow  
33. Relative power of nasal pressure  
34. Relative power of thoracic effort  
35. Relative power of abdominal effort  

Characterizes the power in the breathing frequency band (0.125 - 0.5 Hz). The power was calculated using wavelet decomposition with a 1.5 biorthogonal mother wavelet.

| Table 1. Features extracted and their source signals. | RIP – respiratory inductance plethysmography, CVE – Coefficient of variation of the envelope. |

For an in-depth explanation of $P_{es}$ signal extraction, please see Supplementary Material ($P_{es}$ signal extraction section). In summary, because the absolute amplitude of the $P_{es}$ signal varied from subject to subject, it was normalized to the median of the lower envelope of the $P_{es}$ signal in periods over the night where the subject was awake and breathing normally, as such periods of resting wakefulness reflect unobstructed breathing. From this, each subject’s reference negative $P_{es}$ was
obtained from the median of the values from the wake segments. Finally, the change in peak negative $P_{es}$ relative to the reference value was calculated using the same sliding window as in the feature extraction. For each window, the median value of the lower envelope of the $P_{es}$ signal was calculated and the reference value was subtracted from the result (Figure 4).

**Figure 4.** A 10-second segment of a $P_{es}$ signal with its lower envelope. The relative $P_{es}$ value is calculated as the difference between the median of the lower envelope and a reference value for the negative esophageal pressure. The reference value is obtained as the median of the lower envelope in the segments of the $P_{es}$ recording where the subject is awake, as typical subjects experience unobstructed breathing when awake.
A long short-term memory (LSTM) neural network\textsuperscript{22} was implemented for mapping the non-invasive PSG features described above to the $P_{es}$ signal. LSTM networks provide context-based mapping by applying outputs at previous time steps to predict the output at the current time step. For an in-depth explanation of LSTMs, please see Supplementary Material (\textit{Long Short-Term Memory Network} section). In brief, LSTM cells can integrate information from prior time steps through incorporating various amounts of information from the prior cell (via a forget gate), adding various amounts of information to the prior cell’s information (via an input gate), and outputting various amounts of information (via an output gate)\textsuperscript{22}. The LSTM network constructed for $P_{es}$ mapping had access to a maximum of 100 preceding time steps (500 seconds) per batch of data analyzed, each containing the 35 extracted PSG features.

\textit{Model Development}

The entire dataset was split into three parts: a training set (70%), a test set (15%), and a hold-out, validation set (15%) (\textbf{Table 2}). Optimal choice of LSTM size (25, 50, or 100 cells) was made by performing 500,000 iterations of model formulation and testing over the 85\% training/testing subset, with the final LSTM size chosen based on lowest error in the test set, after a plateau in learning was observed.

<table>
<thead>
<tr>
<th>Data subset</th>
<th>Number of subjects</th>
<th>Sample size</th>
<th>Number of events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$ (%)</td>
<td>(Windows $\times$ Features)</td>
<td>OA: 31,323</td>
</tr>
<tr>
<td>Training set</td>
<td>783 (70%)</td>
<td>4,212,289 $\times$ 35</td>
<td>OH: 101,716</td>
</tr>
</tbody>
</table>
### Table 2.

Partitioning of the dataset into a training set (70%), a test set (15%), and a hold-out, validation set (15%). OA – obstructive apnea, OH – obstructive hypopnea, RERA – respiratory-effort-related arousal, CA – central apnea.

Windows of 10-second time steps with 50% overlap (500 seconds total for each batch of 100 windows) were generated from the PSG, each containing the 35 extracted features of interest. These 100-contiguous-time-step x 35-feature batches were then fed into the LSTM network in random order for model training and optimization through comparison to windows in the test dataset (Supplementary Material, *Long Short-Term Memory Network* section).

**Model Validation**
Following model training and optimization on training (70%) and test (15%) datasets, model performance was assessed on the never-before-seen, hold-out/validation dataset (15%), which contained 914,324 windows of 10-second, 50%-overlapping time steps. This hold-out dataset served as a prospective validation of the cross-validated algorithm without needing to undertake a costly new study with the impractically invasive $P_{es}$. The metrics used to assess model performance in the validation data set were the differences between predicted and measured $P_{es}$ pressure as well as correlation coefficients.

*Software and hardware specifications*

The algorithms related to all preprocessing and feature extraction were developed in Matlab 2017a, while the LSTM neural network was implemented in Python 2.7 combined with the Tensorflow library. Training of the network was carried out using a NVIDIA Tesla K80 GPU at Stanford University’s computing cluster Sherlock. All the code is available on Github\textsuperscript{23}.

**RESULTS**

This sample of 1119 subjects was predominantly male (71%), young ($36.4 \pm 16.3$ years), and slightly overweight ($BMI \ 25.6 \pm 10.9 \ kg/m^2$). Because $P_{es}$ was principally used to determine if a subject had RERAs, most subjects in this cohort had mild sleep disordered breathing. Incorporating RERAs into an overall respiratory disturbance index (RDI), the population demonstrated a moderate degree ($21.0 \pm 18.5$ events/h) of sleep-disordered breathing, thereby allowing for a sufficient number of events for model development.
Table 3 outlines the overall performance of the developed system as measured by the difference and correlation coefficient, between the predicted and measured peak negative $P_{es}$ using the hold-out 15% (168-subject) validation dataset. The 50-cell model was deemed to have the lowest error rate in the test set (see Supplementary Material, Long Short-Term Memory Network section) and was thus used for further analysis. The median difference was 0.61 cmH$_2$O with an interquartile range (IQR) of 2.99 cmH$_2$O. The median correlation coefficient was 0.581 with an IQR of 0.298.

<table>
<thead>
<tr>
<th>MEASURE</th>
<th>5%</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE $P_{es}$ - PREDICTED $P_{es}$ (cmH$_2$O)</td>
<td>-5.851</td>
<td>-0.892</td>
<td>0.609</td>
<td>2.096</td>
<td>5.472</td>
</tr>
<tr>
<td>CORRELATION COEFFICIENT</td>
<td>0.106</td>
<td>0.412</td>
<td>0.581</td>
<td>0.710</td>
<td>0.843</td>
</tr>
</tbody>
</table>

Table 3. Comparison between true and predicted $P_{es}$ values in the validation cohort (N=168).

5$^{\text{th}}$ percentile (5%), 25$^{\text{th}}$ percentile (Q1), median (Q2), 75$^{\text{th}}$ percentile (Q3), and 95$^{\text{th}}$ percentile (95%) of the differences and correlation coefficients between measured and predicted peak negative $P_{es}$.

The general model performance in predicting the peak negative $P_{es}$ for all 168 subjects in the validation set is illustrated in Figure 5. It appears that the model has shortcomings in two specific situations: it rarely predicts positive values, i.e. when the peak negative $P_{es}$ is less negative than the wake reference; and it tends to underestimate more extreme, negative values. A clearer picture of the relationship between measured and predicted $P_{es}$ values emerges upon inspecting shorter time segments (Figure 6). The prediction curve effectively follows the trend of the target curve closely, but the exact values are not predicted as accurately during either very low respiratory effort (less negative/more positive $P_{es}$) or very high respiratory effort (more negative $P_{es}$).
Figure 5. Distribution of the measured and predicted peak negative $P_{es}$ values for all 168 subjects in the validation set. The measured and predicted peak negative $P_{es}$ values depicted in the histogram are derived from 914,324 10-second windows each.
**Figure 6.** The measured and predicted peak negative $P_{es}$ in 1000-second segments for different subjects in the validation set, corresponding to 200 windows in which predictions have been performed.

**Figure 7 a)** depicts the difference between measured and predicted peak negative $P_{es}$ divided into SDB events, and **Figure 7 b)** shows correlation coefficients between measured and predicted $P_{es}$ in the different events. The best performance is achieved in RERA events and normal sleep (i.e. without events), as the difference between measured and predicted $P_{es}$ are lowest in these categories, the correlations between measured and predicted $P_{es}$ are highest, while the IQRs are also small. The difference between measured and predicted $P_{es}$ is slightly lower in OH events than
in OA events, while the IQR in OH is also notably smaller. The poorest performance is undoubtedly obtained in CA events in which the largest difference and lowest correlation between measured and predicted $P_{es}$ is observed. During events scored as central apneas, the model generally tends to over-predict (i.e., produce a more negative result).

![Boxplots showing differences and correlation coefficients](image)

**Figure 7.** (a) Differences between measured and predicted peak negative $P_{es}$, as observed in different types of SDB events, calculated for all 168 subjects of the validation set. (b) Correlation coefficients between measured and predicted peak negative $P_{es}$, as observed in different types of SDB events, calculated for all 168 subjects in the validation set. The values displayed in the boxplots are the median values. Outliers are not displayed to assist visualization.

Table 4 summarizes p-values obtained using paired t-tests at a significance level of 5% to test the hypothesis that the mean peak negative $P_{es}$ – for both measured and predicted values – is significantly different in the different subtypes of SDB events compared to each other. Based on measured peak negative $P_{es}$, normal sleep is significantly different from OA, OH, and RERA events. No significant difference can be shown between normal sleep and CA events. The same conclusions are reached based on predicted peak negative $P_{es}$. The predicted $P_{es}$ values (non-bolded) are even shown to be significantly different between OAs and all other events, including CAs; whereas even the true $P_{es}$ (bold p-values) did not substantially differentiate between OAs and RERAs. Note that the p-values are generally lower for predicted than measured peak negative $P_{es}$ in most cases, i.e. there is generally greater statistical evidence for a separation between events based on predicted peak negative $P_{es}$.

<table>
<thead>
<tr>
<th>Event Comparison</th>
<th>Normal</th>
<th>OA</th>
<th>OH</th>
<th>RERA</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>-</td>
<td>3.37*10^-6</td>
<td>0.0180</td>
<td>0.004</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.89*10^-10</td>
<td>1.58*10^-8</td>
<td>9.83*10^-4</td>
<td>0.140</td>
</tr>
<tr>
<td>OA</td>
<td>-</td>
<td>-</td>
<td>0.001</td>
<td>0.135</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.56*10^-5</td>
<td>0.008</td>
<td>0.022</td>
</tr>
<tr>
<td>OH</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.180</td>
<td>0.217</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.097</td>
<td>0.783</td>
</tr>
<tr>
<td>RERA</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>0.183</td>
</tr>
</tbody>
</table>
DISCUSSION

In this study, we demonstrate that machine-learning modeling using clinically collected SDB-related signals from typical polysomnograms is able to derive a non-invasive sleep-breathing effort signal that closely parallels ($\rho_{\text{median}}=0.581$, $p=0.01$) the actual values of the impractically invasive $P_{es}$. While median differences between the measured and predicted values was only 0.61 cmH$_2$O, the predictive algorithm tended to underestimate extremely negative $P_{es}$ values, as well as positive $P_{es}$ values (Figure 5). Nonetheless, the non-invasively predicted effort values could discern significant differences between normal and obstructive breathing; whereas, centrally-mediated SDB events were not differentiated from normal breathing effort (Table 4), but remained differentiable from obstructive apneas. This ultimately suggests that the derived signal has the potential to provide a measure of sleep breathing effort that could be explored in additional datasets.

Table 4: Comparison of measured and model-predicted $P_{es}$ values between different SDB events in the validation dataset (N=168). p-values derived from two-tailed, paired t-tests assessing the hypothesis that the mean peak negative $P_{es}$ is significantly different between various technician-labeled sleep-disordered breathing events (OA, OH, RERA, CA). OA – obstructive apnea, OH – obstructive hypopnea, RERA – respiratory-effort-related arousal, CA – central apnea. The p-values in bold are p-values for measured $P_{es}$ values, while the other p-values values are p-values for predicted $P_{es}$ values.
to determine optimal thresholds for differentiating between obstructive and central events, and for improved identification of RERAs.

Our model development process determined that a 50-cell LSTM network performed best with a plateau in learning (as tracked by error minimization in the test data subset of 168 individuals) achieved with approximately 300,000 iterations. As mentioned, this model demonstrated moderately strong correlation ($\rho_{\text{median}}=0.581$, $p=0.01$) with measured $P_{es}$ signal in the additional, hold-out/validation subset of 168 individuals (Table 3). The signal correlation provides reassurance that predicted $P_{es}$ value provides a meaningful measure of effort despite an underestimation of positive and extremely negative actual $P_{es}$ values.

The fact that extreme negative values are not well predicted likely relates to two primary issues: (1) there is an inherent imbalance between the number of obstructive breathing events and breathing with normal effort and (2) extremely negative values may be artefactual. Nonetheless, extremely negative $P_{es}$ values are a relatively infrequent occurrence in this dataset, such that the difference between actual and predicted $P_{es}$ values was not substantial (median and IQR: 0.61 cmH$_2$O, -0.89-2.10 cmH$_2$O), likely reflecting the consequences of the normal I:E ratio causing a slight over-representation of time spent with less-negative $P_{es}$ values.

When examining specific SDB subtypes, the pattern of $P_{es}$ values was as expected: obstructive apneas had more negative values than obstructive hypopnea, which had more negative values than RERAs, with normal breathing having the least negative values. Comparatively, performance was quite poor with regard to central apneas. During events labeled as central apneas, the predicted value was often more negative than the actual $P_{es}$ value. While this finding was unexpected, a number of possible explanations exist, the most likely of which is the small sample of central events. In support of this explanation is the finding that, like the actual $P_{es}$ values, despite a
tendency to over-predict (i.e., more negative), predicted $P_{es}$ values during central apneas did not significantly differ from baseline, “normal” breathing, OH, and RERA $P_{es}$ values; however, this differentiation was not possible using the real $P_{es}$ value either (Table 4). That being said, the predictive model appropriately failed to differentiate central apneas from normal breathing based on $P_{es}$ value predictions and had similar difficulty differentiating mild obstructive breathing and central apneas – mirroring the problems inherent in the true $P_{es}$ value – which may reflect multiple issues (including incorrect event labeling, insufficient sample sizes of different event types).

Comparatively, in all obstructive SDB event subtypes, there was a significantly different $P_{es}$ predicted between SDB events and normal breathing (Table 4). This makes sense, because this population was inherently selected for the determination of obstructive SDB, and, therefore, a cohort enriched for centrally-mediated SDB would certainly improve the model’s robustness for detecting these types of breathing events. Nonetheless, this finding is still clinically informative, as the predicted $P_{es}$ values which are not significantly different from the predicted baseline $P_{es}$ values, may be helpful in at least discriminating centrally-mediated from obstructive SDB, all subtypes of which significantly differed from “normal” breathing. Using this information in future studies, a sensitivity analysis could then be used to determine the optimal cutoff for the predicted $P_{es}$ value to differentiate between obstructive and central hypopneas.

This study has some limitations. First, the lopsided nature of the data is an implicit aspect of any effort to explore intermittent phenomena over long periods of observation. This was compounded by the fact that these patients were generally selected due to a suspicion for having less-severe, obstructive sleep-disordered breathing (e.g., UARS). Similarly, as $P_{es}$ baseline bio calibration periods were not collected as part of routine clinical practice, the more positive peak negative $P_{es}$ values are linked to subjects who have questionable wake segments. For instance, some subjects
only have one wake segment and additionally this segment is very noisy, which may have caused the reference value to be very negative. Consequently, when the relative $P_{es}$ was calculated, it could have large, positive values for most windows. This is likely due to the fact that the reference value is based on a minimum of one wake segment of at least two epochs - i.e. one-minute duration - per subject, which may have been insufficient for establishing a baseline. In future model development, ensuring a wake breathing period from the pre-polysomnogram bio calibration period, would likely improve the model’s baseline determination.

Despite these limitations the major strength of this study is the unparalleled size of the cohort undergoing $P_{es}$ monitoring, which allowed for a broad array of SDB phenotypes and the most robust samples of SDB events available. As employing a $P_{es}$ in further cohorts with more severe demonstrations of obstructive or central SDB would require a large consumption of resources, the aforementioned limitations can inform future study designs to prospectively analyze the value of this algorithm in identifying and differentiating among various SDB events.

To date, several non-invasive measures have been investigated, with varying degrees of validation against $P_{es}$. The best alternative by far is the RIP belts, which is currently the AASM-recommended method of choice when it comes to measuring respiratory effort in a routine clinical PSG. It has been shown that the RIP can detect increased respiratory effort to a certain extent; however, it remains to be seen whether the degree of effort can be quantified by RIP as it can by the $P_{es}$. Another shortcoming of RIP is that it measures volume expansion, which does not reflect muscle force of the respiratory muscles in the same manner that $P_{es}$ does. Toward this end this study adopted a completely new approach to solve the issues of measuring respiratory effort. Instead of aiming for an alternative technique that may provide sufficient accuracy in measuring respiratory effort compared to $P_{es}$, a system was developed that can predict the peak negative pressure swings
from the gold-standard $P_{es}$ signal relative to normal wake breathing. This can be used as a measure of respiratory effort based on currently collected, gold-standard, non-invasive PSG signals. To our knowledge, this study constitutes the first attempt to create such a system. The novelty in this system is amplified by the uniqueness of the provided database, which contains more than a thousand $P_{es}$-incorporated PSG recordings. This uniquely robust sample used for training and validation purposes ensures that the proposed system can generalize well to a variety of different subjects with varying degrees and types of sleep-disordered breathing, potentially providing a means of differentiation between obstructive and central events, as well as the detection of RERAs, in a clinical setting.
REFERENCES


16. Berry RB, Ryals S, Girdhar A, Wagner MH. Use of Chest Wall Electromyography to


