Robust, ECG-based algorithm for Sleep Disordered Breathing detection in large population-based cohorts using an automatic, data-driven approach

Olsen, Mads; Mignot, Emmanuel; Jennum, Poul J.; Sørensen, Helge Bjarup Dissing

Published in:
Sleep

Link to article, DOI:
10.1093/sleep/zsz276

Publication date:
2020

Document Version
Peer reviewed version

Citation (APA):
TITLE PAGE

Title:
Robust, ECG-based algorithm for Sleep Disordered Breathing detection in large population-based cohorts

Authors
Mads Olsen a, b, c
Emmanuel Mignot b, †
Poul Jørgen Jennum, Professor c, †
Helge Bjarup Dissing Sorensen a, †
†Shared last author

Affiliations:
a Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark
b Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA
c Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, Glostrup, Denmark
Details of corresponding author:

Full name: Mads Olsen, M.Sc.

Address: Technical University of Denmark
Department of Electrical Engineering
Oersted Plads, Building 349, room 10
DK-2800 Kgs. Lyngby

E-mail: madsol@elektro.dtu.dk

Institution where work was performed:

Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA USA
Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark
Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, Glostrup, Denmark
ABSTRACT

Study objectives: Up to 5% of adults in Western countries have undiagnosed sleep disordered breathing (SDB). Studies have shown that electrocardiogram (ECG)-based algorithms can identify SDB and may provide alternative screening. Most studies however have limited generalizability as they have been conducted using the Apnea-ECG database, a small sample database that lacks complex SDB cases.

Methods: Here, we developed a fully automatic, data-driven algorithm that classifies apnea and hypopnea events based on the ECG using almost 10,000 polysomnographic sleep recordings from two large population-based samples, the Sleep Heart Health Study (SHHS) and the Multi-Ethnic Study of Atherosclerosis (MESA), which contain subjects with a broad range of sleep and cardiovascular diseases (CVD) to ensure heterogeneity.

Results: Performances on average were \( Se = 68.7\% \), \( Pr = 69.1\% \), \( F1 = 66.6\% \) per subject, and accuracy of correctly classifying AHI severity score was \( Acc = 84.9\% \). Target apnea-hypopnea index (AHI) and predicted AHI were highly correlated (\( R^2 = 0.828 \)) across subjects, indicating validity in predicting SDB severity. Our algorithm proved to be statistically robust between databases, between different Periodic Leg Movement Index (PLMI) severity groups, and for subjects with previous CVD incidents. Further, our algorithm achieved state-of-the-art performance of \( Se = 87.8\% \), \( Sp = 91.1\% \), \( Acc = 89.9\% \) using independent comparisons and \( Se = 90.7\% \), \( Sp = 95.7\% \), \( Acc = 93.8\% \) using a transfer learning comparison on the Apnea-ECG database.

Conclusion: Our robust and automatic algorithm constitutes a minimally intrusive and inexpensive screening system for the detection of SDB event using the ECG to alleviate the current problems and costs associated with diagnosing SDB cases and to provide a system capable of identifying undiagnosed SDB cases.

Keywords: electrocardiogram, sleep disordered breathing, apnea, recurrent neural network
STATEMENT OF SIGNIFICANCE

Most electrocardiogram-based sleep disordered breathing (SDB) event detection algorithms in the literature have limited generalizability as they have been conducted on small sample databases that lack complex SDB cases. We developed and tested a data-driven SDB event detection algorithm on almost 10,000 unique polysomnographic recordings from large population-based studies comprising subjects with a broad range of cardiovascular diseases (CVD) and sleep disorders to ensure heterogeneity. Our algorithm showed robust performance between databases and across subjects with periodic leg movement and previously reported CVD. Our hope is that our algorithm will constitute a minimally intrusive and inexpensive screening system to alleviate the current problems and costs associated with diagnosing SDB cases, and to provide a system capable of identifying undiagnosed SDB cases.
INTRODUCTION

Sleep disordered breathing (SDB) comprises several sleep disorders characterized by abnormal respiratory patterns or insufficient ventilation during sleep. A diagnosis of SDB typically relies on the identification of apnea (A), i.e. a total cessation of breathing for 10 sec, and hypopnea (H), a reduction of breathing associated with either hypoxia or an arousal. These events are then added together and divided by total sleep time to constitute the Apnea-Hypopnea Index (AHI), the gold standard measure of SDB in the field. Apnea events are defined as central (CSA) if cessation of breathing is associated with a lack of respiratory effort (typically of central nervous system origin), as obstructive (OSA) if they occur due to a physical obstruction of the upper airway, and as mixed (MSA) if they show a mix of both CSA and OSA behavior. Obstructive events are associated with obesity and are by far the most common cause of SDB. A population-based epidemiologic meta study concluded that untreated SDB of obstructive nature is independently associated with increased likelihood of hypertension, cardiovascular disease, stroke, daytime sleepiness, motor vehicle accidents, and diminished quality of life, with even mild SDB being associated with significant morbidity.

Measuring the AHI involves one or several nights of recording of multiple physiologic cardio-respiratory parameters with either polysomnography (PSG) or respiratory polygraphy (RP), performed in a sleep clinic or as a home-sleep-test (HST). Interpretation requires manual or assisted semiautomatic analysis of recordings by trained medical personnel to derive AHI and diagnosis. This gold standard gives rise to practical issues that are left unaddressed, and the AHI is frequently criticized as too reductionistic. First, a single night recording in a lab or carrying HST equipment does not represent usual at home sleep pattern, notably typical body position, a factor that can greatly affect AHI, thus relying on a single night of sleep may lead to inconclusive and false diagnostic outcome. Second, there is significant internight variability and per event inter scorer reliability is poor. Finally, manual assessment of sleep recordings is a cumbersome and expensive task. The American Academy of Sleep Medicine (AASM) Inter-scorer Reliability Program determined that the inter-scorer agreement per epoch for CSA, OSA, MSA, and H was only 52.4%, 77.1%, 39.8%, and 65.4%, respectively. It is estimated that up to 5% of the adults in Western countries are likely to have undiagnosed OSA syndrome. This clearly indicate a huge amount of people goes under the radar with the current gold standard.

In 2000, Physionet made the small Apnea-ECG dataset publicly available to facilitate development of electro-cardiogram (ECG)-based SDB detectors. This cohort has been the foundation of hundreds of publications on testing alternative SDB detectors. Best performing systems has been done by visual scorings, achieving accuracy (Acc): 92.6%, whereas the best performing automatic algorithms have achieved Acc = 90.6% and Acc = 90.9%. These performances indicate that SDB detection using ECG is possible. However, the Apnea-ECG dataset has limited generalizability, due to its small sample size of only 70 recordings, and most importantly, the dataset lacks complex apneic cases, e.g. subjects with cardiovascular diseases (CVD), central apnea events, parasomnia, periodic leg movements (PLM) or other comorbidities that would truly capture diversity encountered in population and patient based samples, and which are crucial for the development of an effective screening tool. Illustrating this point, a study from 2018 tested the generalizability of the Apnea-ECG dataset and reported poor sensitivity (Se): 14.6% and specificity (Sp): 95.0% and Se = 33.1% and Sp = 88.7% for SDB detection on two separate cohorts.
containing subjects with a broader spectrum of apneic events and other sleep disorders when applying a model trained on the Apnea-ECG dataset with test performance of $Se = 88.3\%$ and $Sp = 88.3\%$\textsuperscript{12}.

Beside use of the Apnea-ECG dataset, few attempts have been made to develop robust models for SDB detection with diverse samples. A study from 2014 presented an SDB classifier using 121 Holter ECG recordings from a group of subjects that were suspected to have SDB. The study reported an independent test performance of $Se = 87.3\%$ and $Sp = 88.0\%$ on 69 of the 70 recordings on the Apnea-ECG dataset\textsuperscript{13}. Unfortunately, all subjects from the group with comorbidities were used to train the model, hence performance in a never seen test set with diverse samples is unknown. Another study from 2010 achieved $Acc = 83.\%$ on all 70 recordings of the Apnea-ECG dataset, and accurately classified 86\% of 862 population-based samples into groups of normal or moderate-to-severe AHI, respectively. However, this was after excluding 28\% of the original 1193 recordings, and using a model which was trained on only 63 ECG recordings from subjects without arrhythmia and free from PLM\textsuperscript{14}. All in all, the true value of ECG based detection systems to detect SDB in the real world remains unknown.

In this study, we describe the development and validation of a fully automated data-driven model that classifies SDB events using the ECG. The model was developed using two large samples: The Sleep Heart Health Study (SHHS) and the Multi-Ethnic Study of Atherosclerosis (MESA), samples that contain subjects with a broad range of sleep and cardiovascular comorbidities. Further, independent replication of the results was done in the Apnea-ECG database to demonstrate generalizability of the model, and to allow comparison to models from other studies.
METHODS

The best performing automatic systems that predict SDB are those using feature engineering, i.e. classifiers trained on features manually extracted from the ECG derived signals: RR interval trend (RR) and ECG-derived respiration (EDR), which contain heart rhythm and heart rate, and respiratory information, respectively\textsuperscript{10,11}. Data-driven classification methods have the benefit of automatizing feature extraction, and thereby eliminating a manual decision step which are often cumbersome and biased when data is abundant. Automatic, data-driven methods have recently shown to outperform feature engineering in many medical applications and have shown to work with only minimal preprocessing\textsuperscript{15–18}. An automatic, data-driven algorithm was selected for the task of classifying SDB in large population-based samples to utilize the generic feature learning capabilities of such systems and to account for limitation of manual feature engineering.

A. Data used in the study

The model was trained, evaluated and tested on a dataset created using two large population based, multi-center, multi-ethnical samples from the National Sleep Research Resource (NSRR) comprising subjects with a broad range of CVD and sleep disorders to ensure heterogeneity\textsuperscript{19,20}. Separate tests were performed in an independent, publicly available dataset, which had not been seen by the model before. Table 1 summarizes the samples used along with demographics and sleep related characteristics.

1) Sleep Heart Health Study (SHHS): The SHHS is a multi-center prospective study that was designed to investigate SDB as risk factor for the development of CVD\textsuperscript{19–22}. It is a population-based, multi-ethnical, gender-balanced cohort, comprised of subjects of at least 40 years of age, with only individuals in treatment for sleep apnea excluded from the study. 6441 subjects were enrolled during a first visit, and 3295 were reenrolled for a follow up PSG recording visit approximately five years later. In total, 8444 PSG recordings were available through the NSRR when the two visits were combined\textsuperscript{19,20} with overlapping SDB related events scored once according to AASM criteria that scores hypopnea events when associated with either hypoxia or arousals\textsuperscript{23}. Furthermore, 22.6\% of the recordings came from subjects that had reported one or more incidents of CVD happening prior to the recording as presented in Table 1. 82 recordings had less than 1 hour of consecutive sleep and were excluded. From the remaining 8362 PSG recordings single lead ECG (lead II) digitized at 125 Hz was extracted for further processing. The recordings were randomly split into a training, evaluation, and test set with partitioning sizes: 80\%, 10\%, and 10\%, respectively (Table 1).

2) Multi Ethnic Study of Atherosclerosis (MESA): MESA sleep is a prospective cohort study designed to investigate how variations in sleep and sleep disorders vary across gender and ethnical groups\textsuperscript{19,20}. The cohort is population based and includes 2237 men and women aged 45–84 years recruited from 6-centers in USA\textsuperscript{24}, and similarly to SHHS only individuals in treatment for sleep apnea were excluded from the study. All subjects underwent a full overnight unattended PSG and sleep related events were scored once according to the same AASM criteria\textsuperscript{23}, including AHI and PLM index (PLMI) related events as presented in Table 1. Of the 2237 subjects 2056 had a corresponding annotation file, and 714 files were excluded because medical personnel who scored the recording had one of the following notes: “Scoring of Apnea/Hypopnea was unreliable” or “Scoring of respiratory events was unreliable”, or due to the
minimum of 1-hour of consecutive sleep criteria also used for SHHS. From the remaining 1342 PSG recordings single lead ECG (lead II) digitized at 256 Hz with 12-bit resolution was extracted for further processing. The recordings were randomly split into a training, evaluation, and test set with partition sizes: 80%, 10%, and 10%, respectively (Table 1).

3) Apnea-ECG dataset: The study design for the Apnea-ECG Physionet dataset has been published elsewhere\textsuperscript{7,25}. The Apnea-ECG sample contains 70 ECG recordings from subjects with age ranging between 27 and 63 years and having varying SDB severity. The dataset is originally divided into a learning set and a test set of equal size and with balanced SDB cases, however as all recordings are used for testing in our study these were merged. The dataset does not contain episodes of CSA, and does not include subjects with CVD, parasomnia, limb movements or other sleep comorbidities. Experts have scored the recordings in one-minute segments as either normal or apneic (OSA or H) using overlapping respiratory channels that were not made available. Each recording contains a single ECG signal (lead II) digitized at 100 Hz with 12-bit resolution.

B. ECG signal pre-processing and segmentation

The data must be standardized to account for the differences in sampling frequency and recording duration between databases and recordings. A coarse overview of data standardization is presented in Figure 1. The ECG signal was extracted from each recording and interpolated to 256 Hz using cubic spline interpolation. Heart rate and rhythm was captured by extracting the RR from the ECG at 4 Hz using a robust, fully automatic RR tachogram extraction algorithm\textsuperscript{26} that accounts for movement artifacts and ectopic beats from the ECG. The R peaks were detected and localized and the differences between adjacent R peaks were calculated to produce the RR tachogram, which directly presents information on duration and variation between heartbeats. Only heartbeats considered as normal, i.e. not categorized as either artefacts or ectopic beats, were then cubic spline interpolated and resampled to form a uniformly spaced time series at a sampling rate of 4 Hz, which is referred to as RR. Additionally, the EDR which encapsulates respiratory information, was extracted by calculating the phase space reconstruction area at each R peak\textsuperscript{11}.

The RR and EDR signals, referred to as channels, were time aligned in 5 min segments with 40% overlap and normalized separately using a soft min-max normalization scheme where the 5th and 95th percentiles are used instead of the extrema to minimize outlier influence.

18 segments containing the label “Respiratory artifact” were excluded from the study, leaving a total of 1.059.131, 132.159, and 132.859 5 min segments available for training, evaluation, and test sets, respectively.

C. Data-driven model for automatic detection of SDB.

In this study, the task of detecting SDB events using ECG was defined as a sequence-to-sequence learning task, where the model takes as input stacked sequences of RR and EDR sampled at 4 Hz, \( X = [x_1, \ldots, x_{4S}] \), and outputs a sequence of binary predictions for every second, \( y = [y_1, \ldots, y_S] \), categorizing an absence or a presence of SDB events.
The model parameters were optimized by minimizing the cost function given by the monotonically
decreasing cross entropy. Cross entropy is especially suitable for classification problems since it gives
penalty that increases exponentially the further away the output probability is from the target class.

1) Architecture: The choice of classifier is paramount in an automatic data-driven learning task. Sequence-
to-sequence modelling rely heavily on context awareness, and the classifier should include modules that
incorporate temporal information. Long-short term memory (LSTM) and gated recurrent units (GRU) have
shown to outperform regular recurrent units comprehensive time series classification tasks, e.g. speech
signal modelling\textsuperscript{27}. Inspired by this, we chose to build a neural network with GRU modules, since these
require less computational power than LSTM and have shown similar performance. Regular GRU modules
processes the input signals forwardly, i.e. from left to right, we selected bidirectional GRU modules that
allow both forward and reverse processing to increase context awareness. The topology of the of the
classifier follow a block design, where each block has a feature extraction module, i.e. GRU layers with
rectified linear units (ReLU) activation, a max-pool layer to reduce the special dimension, ultimately for
the predicted output to match the target resolution of 1 s, and finally, batch normalization\textsuperscript{28} and dropout\textsuperscript{29}
layers that are included to reduce overfitting. The last block is the classification block, which consists of a
regular dense-connected layer with ReLU activation, and a sigmoid layer that assigns the probability of
each output to be an SDB event or not. The exact configuration of the classifier was selected by conducting
several experiments over predetermined ranges for each block: Number of Bidirectional GRU’s in
succession: \{1, 2, 3\}, increasing number allow more complex features to be learned; Number of filters of
the recurrent layers and the densely-connected layer: \{64, 128, 256, 512\}, again increasing the number
increases complexity of the model. Fewer parameters were prioritized for experiments yielding similar
performance to optimize processing time. The final model has 1,125,378 parameters and is presented in
Table 2.

2) Implementation: The model was trained from scratch using the adaptive moment estimation (ADAM)\textsuperscript{30}
optimization with default hyperparameters. The model was initialized using a robust initialization
scheme\textsuperscript{31} with a batch size of 128, a weight decay of 10\textsuperscript{−4}, and a learning rate of 10\textsuperscript{−3}, which in turn was
divided by 10 if evaluation accuracy plateaued for more than 10 epochs, i.e. one iteration through the
training set. The model was trained until evaluation accuracy plateaued for more than 20 epochs. RR and
EDR extraction from the ECG was done in MATLAB as described earlier\textsuperscript{26}. All other code was written using
Python 3.6.7 and the model was developed with Keras 2.2.4 using Tensorflow 1.12.0 backend. The code
for the detector, including model weights is available on GitHub at: https://github.com/stanford-
stages/ECG-SDB-event-detector.

D. Targets

Our screening tool was designed be able to detect all SDB subtypes; but not to distinguish between them,
therefore it was chosen to design a binary target that was SDB positive when any of the SDB events were
present, i.e. CSA, OSA, MSA, and H, and negative in any other cases. The SDB subtypes are visualized in
Figure 2 by time locking the normalized RR feature and the standard deviation of the normalized EDR
feature to the beginning of each SDB subtype event for each database. Each blue curve displays the
median of the all normalized SDB subtype events 60 s before and after the beginning of the event, and
the black curve display the median of all SDB events for each database. Increasing values of RR means that
the distance between each heart beat increases i.e. bradycardia, and vice versa, and increasing values of EDR indicates deeper breathing, whereas decreasing signifies shallow breathing. Figure 2 shows that all SDB subtypes introduce bradycardia and shallow breathing followed by tachycardia and reestablishment of normal breathing from visual inspection of RR and EDR, respectively.

SDB is one of many phenomena influencing cardiac activity during sleep, and confounding factors are expected in large population-based cohorts such as MESA and SHHS. In addition to SDB events, MESA provided PLM event scorings and SHHS had CVD information for each recording (Table 1), which allowed further testing of the robustness of the model in the presence of these comorbidities. Figure 3 shows in addition to the median of all SDB events from Figure 2 the median of all PLM events from the MESA database and shows that PLM introduce abrupt tachycardia and deep breathing. It should be clear from visual inspection of Figure 3 that the algorithm should be able to PLM from SDB.

Each 5 min segment was provided with a target vector containing binary SDB scorings with 1 s resolution. Only the central 60% of each 5 min segment was evaluated to ensure each prediction had good context awareness. Using 40% overlap ensures that the entire original recording can be reconstructed and evaluated except from the outer most minutes, as shown in Figure 4.

E. Performance evaluation

SDB classification constitutes an imbalanced learning problem, since most of the PSG recordings do not contain any SDB events. Therefore, it only makes sense to consider event-based performance metrics, e.g. Se, precision (Pr), and F1 score (F1) as shown in Equation 1.

\[ F_1 = 2 \frac{Pr \cdot Se}{Pr + Se} \] (1)

The model was trained and optimized on a second-by-second basis, however in practice it is not expected to be necessary for the model, which is based on an indirect SDB detection modality, i.e. ECG, to learn the exact on and offset of SDB events, rather it must identify presence or absence of SDB events. Performance was thus reported on an event-by-event basis. The output from the model is the probability of an SDB event to be present at each second. SDB events are formed by merging consecutive model predictions, whose probability surpasses a given threshold, which is determined from maximizing F1 (See Results). Events shorter than 10 s were removed, inspired by the AASM criteria of SDB, and successive events located closer than 10 seconds of each other were merged into one single event. A predicted event was reported as True Positive (TP) if it overlapped with an annotated SDB event by 1 s or more. A predicted event with no overlap to a target event was reported as False Positive (FP), and a target event with no overlap to a predicted event was reported as False Negative (FN).

F. Baseline comparison

The Apnea-ECG dataset was used to compare performances of the data-driven model to solutions from other studies. To allow comparisons, model predictions were translated from 1 s to 1 min resolution predictions, by labeling a target as positive if any SDB event, i.e. CSA, OSA, MSA, or H, was present for at
least 10 consecutive seconds within the corresponding 1 min window to match the scorings of the Apnea-
ECG dataset. The performance on Apnea-ECG dataset is reported using two different comparison
schemes: independent comparison and transfer learning comparison. Independent comparison was
performed by applying the data-driven model, which was trained on the SHHS and MESA cohorts, to all
70 recordings of the Apnea-ECG dataset. Transfer learning comparison was performed by re-training the
above data-driven model on the 35 recordings from the released Apnea-ECG dataset, and applying it on
the withheld 35 recordings from the Apnea-ECG dataset.
RESULTS

The performance of the second-by-second data-driven model in terms of $Se$, $Pr$, and $F1$ on the training, evaluation, and test datasets is displayed in Figure 5 (left and middle). The threshold that optimized $F1$ was determined to 0.332 as presented in Figure 5 (middle). From visual inspection, the model appears only to have overfitted slightly to the training dataset, whereas the model shows similar performance on the evaluation and test datasets. The event-based performance was visualized in Figure 5 (right) by evaluating different criteria for the intersection over union, i.e. the required overlap between a predicted and an annotated event. This figure displays that the performance takes major drops when strict criteria are applied.

The test performance of the $F1$ optimized model (Figure 5) for the combined and individual test sets from each cohort is presented in Table 3. For the entire test set the model performed with $Se = 68.7\%$, $Pr = 69.1\%$, and $F1 = 66.6\%$ on a subject basis, and with no significant performance difference between cohorts. For all events the model performed with $Se = 70.9\%$, $Pr = 73.4\%$, and $F1 = 72.1\%$, and SDB event subtype analysis revealed that the model detected 94.6$, 94.2$%, 94.7$, and 65.8$% of CSA, OSA, MSA, and H events, respectively.

The target AHI (AHItar) and the predicted AHI (AHIpre) were calculated from the ratio of the number of SDB events to the total sleep time. Figure 6 (left) shows a scatter plot of the AHItar and AHIpre along with a regression fitting line (solid line) that shows a correlation of $R^2=0.829$.

Figure 6 (right) shows a Bland-Altman plot that was used to analyze the agreement between AHItar and AHIpre, and shows the difference between the two AHI values against the mean of the two AHI values. From the Bland-Altman analysis a mean difference between AHItar and AHIpre of +0.9 events/h was found, meaning that the DL model underestimated AHI, with a 95% confidential interval ranging from -14.9 to 16.8 events/h.

A confusion matrix of the target and predicted test AHI groups is presented in Table 4. An overall accuracy of $Acc = 84.9\%$ was achieved for categorizing AHI groups, and all but one false classifications fall into the neighboring category. There is one extreme outlier: shhs1-202029 with AHItar=42.3 and AHIpre=2.93.

Further analysis revealed that this person had pacemaker as the only person in the test set.

A) Replication of performance on independent dataset

The performance on the Apnea-ECG dataset is presented in Table 5 using the two different comparison schemes: Independent comparison (lower Table 5) and transfer learning comparison (upper Table 5). For both comparison schemes the model presented in this work achieve state-of-the-art performance and event surpasses the performance of visual scorings.

B) Model robustness

To analyze the model’s capability to distinguish SDB from PLM, all recordings from the MESA cohort were categorized into 9 subgroups based on AHI and PLMI severity as presented in Supplementary Table 1. A multi-comparison analysis was performed which included calculation of 36 pairwise t-tests for the $F1$ performance of the 9 subgroups; $p$ values are presented in Supplementary Table 1. Figure 7 presents
comparison intervals for each group using a Bonferroni corrected significance level of $\alpha = \frac{10^{-2}}{36}$. It shows that the AHI severity groups perform significantly different; where performance is better for higher AHI severity groups. Additionally, Figure 7 shows that performance does not differ significantly between PLMI severity groups, which means the model is robust to PLM events.

To assess the robustness of the model to subjects with reported CVD, recordings from the test set of the SHHS cohort were partitioned based on subjective reporting of any CVD incident before the PSG recording. A t-test between the CVD group and the non-CVD group was conducted and presented in Supplementary Table 1. The t-test yielded a $p$ value = 0.15, with a significance level of $\alpha = 0.01$, which shows that the two groups are not significantly different and therefore the model is robust to subjects with previous reported CVD incidents.
DISCUSSION

In this study, a fully automatic data-driven model is presented to classify SDB events in sleep using the ECG only. On the MESA and the SHHS cohorts the model performed on average with $Se = 68.7\%$, $Pr = 69.1\%$, and $F1 = 66.6\%$ per subject, and with $Se = 70.9\%$, $Pr = 73.4\%$, and $F1 = 72.1\%$ for all events (see Table 3). Additionally, the model correctly classified 84.9\% of subjects into their corresponding AHI severity groups, which in this study was given by of AHI<5, 5<AHI<30, and 30<AHI (see Table 4). On the Apnea-ECG dataset the model achieved state-of-the-art performance with $Se = 87.8\%$, $Sp = 91.1\%$, $Acc = 89.9\%$ using independent comparison and $Se = 90.7\%$, $Sp = 95.7\%$, $Acc = 93.8\%$ using a transfer learning comparison.

The performance of the model is based on using the criterion that a predicted event is reported as TP if it overlaps with an annotated SDB event by 1 s or more. Figure 5 (right) shows that performance is reduced drastically when stricter criteria are applied. The 1 s overlap criterion is considered to be justifiable for two reasons; Firstly, the criterion is actually less generous than applying the minute-by-minute scorings that many have been using from the Apnea-ECG database$^7$, since the minute-by-minute scorings makes it possible for a predicted and an annotated event not to overlap at all; Secondly the purpose of the screening tool is to identify the presence or absence of SDB events, and not the exact onset and duration.

Maier et al. (2014)$^{13}$ proposed a threshold-based classifier trained on 121 ECG recordings from subjects with comorbidities, including subjects with diabetes and myocardial infarction, and the only exclusion criteria were to remove cases of persistent supraventricular arrhythmias. The study showed good independent test performance of $Se = 87.3\%$, $Sp = 88.0\%$ on the Apnea-ECG database. Unfortunately, no test performance on subjects with comorbidities was reported for this study, which made comparisons between groups with comorbidities to this study unrealistic. Hayano et al. (2011)$^{14}$ trained a model that accurately classified 86\% of 862 population-based test samples into groups of normal or moderate-to-severe AHI, and with a correlation of $R^2=0.84$ between $AHI_{tar}$ and $AHI_{pre}$. They show a slightly better AHI severity group accuracy and correlation than the one reported in this study ($Acc = 84.9\%$ and $R^2=0.83$, respectively). However, they excluded 28\% of the original 1193 recordings$^{14}$, and the model was only trained on 63 ECG recordings from subjects without any arrhythmia or PLM. This suggests that the model they present will not generalize well without substantial data cleaning. Hayano et al. (2014) confirmed this in their paper by reporting a specificity drop of 94\% to 61\% when their model was applied to subjects with PLMI > 10, and by their relatively low accuracy of 83\% on the Apnea-ECG database. Clearly, our model generalizes better since it shows to be robust across PLMI severity groups (Figure 7) and achieves a significantly better accuracy of 89.9\% when applied to the Apnea-ECG dataset (Table 5).

Figure 7 and Supplementary Table 1 shows that performance is significantly higher for more severe AHI groups. This is a consequence of how the model parameters were optimized, which was done on a second-by-second basis. The model does not distinguish between AHI severity groups, and consequently recordings with higher AHI that have more AH events will dictate the performance of the model. Equal performance for the different AHI severity groups could be achieved if it was chosen to optimize the separation of AHI severity groups instead.
Statistical analysis showed that the model is robust across the MESA and SHHS cohorts (p value = 0.97, Table 3), across PLMI severity groups (Figure 7 and Supplementary Table 1), and between the CVD group and the non-CVD group (p value = 0.15). For the latter case, the CVD group is based on a subject reporting any CVD incident prior to the PSG recording, therefore abnormalities that could affect ECG detection of SDB such as arrhythmias may or may not be present in these PSG recordings. Furthermore, the robust RR tachogram extraction algorithm that was used is intended to account for movement artifacts and ectopic beats by only interpolating the non-ectopic and non-artifact R peaks from the ECG. The use of the RR tachogram extraction algorithm may explain model robustness to the CVD group. Heart rhythm abnormalities or arrhythmias were not annotated for the PSG files, therefore further analysis could not be conducted.

The final screening device may have multiple applications and can be tailored to the desired end-user by changing the threshold of the model (Figure 5 left and middle). Choosing a high sensitivity on the expense of a low precision is beneficial in a clinical setting as initial screening, since this ensures that all subjects with SDB are identified, and those that are identified as SDB negative can be dismissed with certainty without any further treatment. Conversely, choosing a high precision on the expense of a low sensitivity is a preferred setting to identify some of the estimated 5% of adults in Western countries that are likely to have undiagnosed OSA syndrome. This is to ensure that the cases that are identify as SDB positive are in fact positive, such that healthy people will not get worried without reasonable doubt even before they start a clinical intervention. We are confident that this setting will identify the more severe SDB cases in the public given the high sensitivity of our algorithm to the more severe SDB events: CSA=94.6%, OSA=94.2%, and MSA=94.7%, and given the high correlation of 84% for AHI severity cases. In practice, the screening device must include a validation test of the signal quality and the heart rhythm reliability before the actual screening is carried out. Our hope is that our algorithm will constitute a minimally intrusive and inexpensive screening system to alleviate the current problems and costs associated with diagnosing SDB cases.

A) Limitations

SDB event subtype analysis showed that the model correctly detected 94.6%, 94.2%, 94.7%, and 65.8% of CSA, OSA, MSA, and H events, respectively (See Table 3). The large drop in performance for the detection of H events constitute a limitation of our model, which can be explained by the fact that H events affect breathing less than actual apnea events and are only scored if they are accompanied by either an oxygen desaturation or an arousal. Consequently, as shown in Figure 2, H events show less impact on our features when compared to actual apnea events. This fact and given that this study does not include modalities to identify oxygen desaturations or arousals explain why H events have lower performance than actual apnea events. Since H constitute over 85% of the total amount of SDB events, this has a large impact on the overall performance.

The RR and EDR features are extracted from the ECG following a robust preprocessing algorithm. This step could be omitted if the model was applied directly on the raw ECG signal, and this may even increase performance, since not all info from the ECG is retained in the RR and EDR signals.
We developed and tested a data-driven SDB event detection algorithm on almost 10,000 unique polysomnographic recordings from large population-based studies comprising subjects with a broad range of CVD and sleep disorders to ensure heterogeneity. Our algorithm proved to be statistically robust between databases, between different PLMI severity groups, and for subjects with previous reported CVD incidents, and achieved state-of-the-art performance on an independent dataset, albeit this was the Apnea-ECG database which has limited generalizability. Our algorithm did show weakness to the only subject with pacemaker from the test set, whom also presented to be the only extreme outlier (Table 4). Further analysis on subjects with pacemaker was not carried out due to the limited sample size, however, this should be accounted for in future iterations of the screening device, considering the rapid adoption of pacemaker technology, either by training the model on pacemaker-focused datasets or by using additional modalities. Furthermore, if the goal is to detect subjects that remain undiagnosed in the public, the ECG may not be the ideal tool as it typically needs technical intervention to be set up. The MESA and SHHS cohorts do contain subjects with a broad range of other comorbidities, however those comorbidities that rely on a subjective severity rating of a disorder, e.g. insomnia, pain, and depression, were not analyzed further due to the lack of an objective diagnosis. Ultimately, there is need for an independent, large standardized database that contain subjects with a variety of objectively annotated comorbidities to provide a complete analysis of the true, unbiased performance of our algorithm.
ACKNOWLEDGMENTS

Mads Olsen and Emmanuel Mignot are partially funded by the Klarman Family Foundation.
Mads Olsen reports grants from Augustinus Fonden, Knud Højgaards Fond, Marie & M.B. Richters Fond, Danmark Amerika Fonden, Reinholdt W. Jorck og Hustus Fond, Otto Mønsted Fonden, CACHET, and from the Technical University of Denmark Emmanuel Mignot occasionally consults and has received contracts from Jazz Pharmaceuticals, is and has been a Principal Investigator on clinical trials using sodium oxybate and Solriamfetol, Jazz Pharmaceutical products, for the treatment of Type 1 Narcolepsy. He has received research and clinical trial support at Stanford from Jazz Pharmaceuticals, Apple, Google, Merck, and Balance; and is a consultant for Rhythm (a sleep consumer product company) and Alairion (a sleep apnea pharmacology company) and is on the speakers’ bureau for Vox Media. Poul J. Jennum has nothing to disclose. Helge B. D. Sorensen has nothing to disclose. Conflict of interest: None of the above have any scientific relationships to the current study. There exist no financial or other relationships that might lead to a perceived conflict of interest.
452 NON-FINANCIAL DISCLOSURE
453 Mads Olsen has nothing to disclose. Emmanuel Mignot has nothing to disclose. Poul J. Jennum has
454 nothing to disclose. Helge B. D. Sørensen has nothing to disclose.
455
REFERENCES


Figure 1 – Coarse standardization overview. The ECG input from each database is firstly resampled to 256 Hz, then the RR and EDR signals are extracted at 4 Hz using a robust, fully automatic RR tachogram extraction algorithm, and finally the RR and EDR signals are segmented into 5 min windows and soft-normalized. The output, $X$, are the standardized and segmented signals. SHHS: Sleep Heart Health Study; MESA: Multi-Ethnic Study of Atherosclerosis cohort; ECG: Electrocardiogram; RR: RR interval trend; EDR: ECG-derived respiration.
Figure 2 – Normalized RR feature and standard deviation of the normalized EDR feature to the beginning of each SDB subtype event from the MESA and SHHS cohort. Each blue curve displays the median of the all normalized SDB subtype events 60 s before and after the beginning of the event, and the black curve display the median of all SDB events for each database. Increasing values of RR means that the distance between each heart beat increases i.e. bradycardia, and vice versa, and increasing values of EDR indicates deeper breathing, whereas decreasing signifies shallow breathing. MSA is not shown since too few were scored to show a meaningful curve; however, MSA are included in the median. ECG: Electrocardiogram, RR: RR interval trend, EDR: ECG-derived respiration, SDB: Sleep disordered breathing, MESA: Multi Ethnic Study of Atherosclerosis, SHHS: Sleep Heart and Health Study, CSA: Central sleep apnea, OSA: Obstructive sleep apnea, MSA: Mixed sleep apnea, H: Hypopnea.
Figure 3 – Normalized RR feature and standard deviation of the normalized EDR feature to the beginning of each SDB subtype event from the MESA database. Each blue curve displays the median of the all normalized SDB subtype events 60 s before and after the beginning of the event, and the black curve display the median of all SDB events for each database. Increasing values of RR means that the distance between each heart beat increases i.e. bradycardia, and vice versa, and increasing values of EDR indicates deeper breathing, whereas decreasing signifies shallow breathing. ECG: Electrocardiogram, RR: RR interval trend, EDR: ECG-derived respiration, SDB: Sleep disordered breathing, PLM: Periodic leg movement, MESA: Multi Ethnic Study of Atherosclerosis, CSA: Central sleep apnea, OSA: Obstructive sleep apnea, H: Hypopnea.

Figure 4 – Schematic overview of segment evaluation. Central 60% of each 5 min segment is considered for evaluation and 40% overlap between neighboring segments ensures that the original recording can be reconstructed. Scorings are shown in blocks with 1-min resolution for the sake of simplification. Red blocks are not considered in the model, green blocks are SDB event positive, white blocks are SDB event negative, i.e. N: Normal block.
Figure 5 – Second-by-second and event-based performance of the model. Left: Precision and sensitivity for the training, evaluation, and test dataset. Middle: $F_1$ and Threshold for the training, evaluation, and test dataset. The threshold that optimized $F_1$ was determined to 0.332, which is indicated with a red dot. Right: $F_1$ and intersection over union of predicted and target events for MESA and SHHS.

Figure 6 – Scatter plots of target and predicted AHI. Left: Scatter plot of the correlation between $AHI_{tar}$ and $AHI_{pre}$. A correlation of $R^2=0.828$ was found. The grey lines divide the data into the AHI groups presented in Table 1. Right: Bland-Altman plot, mean difference between $AHI_{tar}$ and $AHI_{pre}$ of +0.9 events/h was found with a 95% confidential interval ranging from -14.9 to 16.8 events/h. AHI: Apnea-hypopnea index. $AHI_{tar}$: target AHI; $AHI_{pre}$: predicted AHI.
Figure 7 - comparison analysis of AHI and PLMI subgroups. Two group means are significantly different if their intervals are disjoint; they are not significantly different if their intervals overlap. Significance values are adjusted for inflation using Bonferroni correction to $\alpha = 2.78 \cdot 10^{-4}$. AHI: Apnea-hypopnea index; PLMI: Periodic leg movement index.
Table 1 – Data cohorts (9869 PSG’s), subgroups, demographics and SDB related characteristics.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Age (years), ( \mu \pm \sigma )</th>
<th>BMI (kg/m(^2)), ( \mu \pm \sigma )</th>
<th>Sex (% male)</th>
<th>Apnea ECG classifier</th>
<th>AHI</th>
<th>PLMI</th>
<th>CVD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Train</td>
<td>Eval</td>
<td>Test</td>
</tr>
<tr>
<td>SHHS</td>
<td>63.5±11.5</td>
<td>28.5±2.1</td>
<td>52.6</td>
<td></td>
<td>6754</td>
<td>845</td>
<td>845</td>
</tr>
<tr>
<td>MESA</td>
<td>69.1±8.9</td>
<td>28.7±5.2</td>
<td>45.5</td>
<td></td>
<td>1083</td>
<td>136</td>
<td>136</td>
</tr>
<tr>
<td>Apnea-ECG</td>
<td>46.5±10.1</td>
<td>28.0±6.6</td>
<td>81.4</td>
<td></td>
<td>-</td>
<td>-</td>
<td>70</td>
</tr>
<tr>
<td>Total recordings</td>
<td>-</td>
<td>-</td>
<td>7837</td>
<td></td>
<td>981</td>
<td>1051</td>
<td>2276</td>
</tr>
<tr>
<td>Total PSG used</td>
<td>-</td>
<td>-</td>
<td>7762</td>
<td></td>
<td>974</td>
<td>1038</td>
<td>2254</td>
</tr>
</tbody>
</table>

\( \mu \): mean; \( \sigma \): standard deviation; kg: kilogram; m: meter; Eval: evaluation; AH\(i\): Apnea-Hypopnea Index; PL\(MI\): Periodic Leg Movement Index; CVD: Cardiovascular Disease; SHHS: Sleep Heart Health Study; MESA: Multi-Ethnic Study of Atherosclerosis cohort. Note that PL\(MI\) were not recorded in the SHHS sample.

Table 2 – General architecture of the data driven classification model

<table>
<thead>
<tr>
<th>Layer type</th>
<th>Kernel size</th>
<th>Stride</th>
<th>Output dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input layer</td>
<td>-</td>
<td>-</td>
<td>(B, T, 2)</td>
</tr>
<tr>
<td>Bidirectional GRU</td>
<td>C</td>
<td>1</td>
<td>(B, T / 2(^1), 2-C)</td>
</tr>
<tr>
<td>Bidirectional GRU</td>
<td>C</td>
<td>1</td>
<td>(B, T / 2(^1), 2-C)</td>
</tr>
<tr>
<td>Maxpool</td>
<td>-</td>
<td>2</td>
<td>(B, T / 2(^1), 2-C)</td>
</tr>
<tr>
<td>ReLU</td>
<td>-</td>
<td>1</td>
<td>(B, T / 2(^1), 2-C)</td>
</tr>
<tr>
<td>Dropout (50%)</td>
<td>-</td>
<td>-</td>
<td>(B, T / 2(^1), 2-C)</td>
</tr>
<tr>
<td>Dense</td>
<td>4-C</td>
<td>1</td>
<td>(B, T / 2(^3), 4-C)</td>
</tr>
<tr>
<td>ReLU</td>
<td>-</td>
<td>1</td>
<td>(B, T / 2(^3), 4-C)</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>-</td>
<td>1</td>
<td>(B, T / 4, 1)</td>
</tr>
</tbody>
</table>

The model processes an input of shape (B, T, 2) corresponding to B time series of 2 features with T timesteps, and outputs the shape (B, T/2\(^1\), 1) corresponding to T/2\(^1\) predictions for B timeseries. The model was implemented with the following parameters: B=128, T=300-fs, C=128, k={1, 2} such that the input shape was (128, 1200, 2) and the output shape was (128, 300, 1). GRU: Gated Recurrent Unit; ReLU: rectified linear unit; \( p \): percent; fs: sample frequency.
Table 3 – Overall test performance of model.

<table>
<thead>
<tr>
<th></th>
<th>Test performance</th>
<th>Test performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subject-based</td>
<td>Event-based (mixed)</td>
</tr>
<tr>
<td></td>
<td>µ ± σ</td>
<td>µ ± σ</td>
</tr>
<tr>
<td></td>
<td>mixed</td>
<td>SHHS</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>68.7±18.4</td>
<td>68.2±18.3</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>69.1±18.9</td>
<td>69.2±19.0</td>
</tr>
<tr>
<td>F1</td>
<td>66.6±16.1</td>
<td>66.3±16.0</td>
</tr>
<tr>
<td>Number of recordings</td>
<td>968</td>
<td>833</td>
</tr>
<tr>
<td>Number of Events</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Left column shows subject-based test performance for the combined (mixed) and separate cohorts and right column shows test performance for all events for the combined cohorts (mixed). The F1 performance difference was tested between cohorts using a t-test; p value is presented. µ: mean; σ: standard deviation.

Table 4 – Confusion matrix of AHI severity

<table>
<thead>
<tr>
<th>Predicted</th>
<th>AHI&lt;5</th>
<th>5&lt;AHI&lt;30</th>
<th>AHI&gt;30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AHI&lt;5</td>
<td>16</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>5&lt;AHI&lt;30</td>
<td>7</td>
<td>645</td>
<td>33</td>
</tr>
<tr>
<td>AHI&gt;30</td>
<td>1</td>
<td>71</td>
<td>161</td>
</tr>
</tbody>
</table>

AHI: Apnea-Hypopnea Index
### Table 5 – Model performance on Apnea-ECG dataset.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Training data</th>
<th>Apnea-ECG Test</th>
<th>Apnea-ECG All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>35 ECG recordings</td>
<td></td>
<td>70 ECG recordings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Se (%)</td>
<td>Sp (%)</td>
<td>Acc (%)</td>
</tr>
<tr>
<td>Raymond (2000)</td>
<td>Visual</td>
<td>35 (Apnea-ECG)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>McNames (2000)</td>
<td>Visual</td>
<td>35 (Apnea-ECG)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>De Chazal (2003)</td>
<td>Visual</td>
<td>35 (Apnea-ECG)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>De Chazal (2003)</td>
<td>Auto-FE</td>
<td>35 (Apnea-ECG)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Janbakhshi (2018)</td>
<td>Auto-FE</td>
<td>35 (Apnea-ECG)</td>
<td>89.6</td>
<td>91.8</td>
</tr>
<tr>
<td>This work</td>
<td>Auto-DD transfer</td>
<td>Pretrain: 8736 (SHHS, MESA)</td>
<td>90.7</td>
<td>95.7</td>
</tr>
<tr>
<td>Maier (2014)</td>
<td>Auto-FE</td>
<td>121 Holter ECG</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hayano (2011)</td>
<td>Auto-FE</td>
<td>63 ECGs</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>This work</td>
<td>Auto-DD</td>
<td>8736 (SHHS, MESA)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Upper table shows transfer learning comparison; Lower table shows independent comparison. A t-test between the two best performing solutions was conducted for both comparison schemes and p values are presented next to the best solution.

*Only 69 out of 70 recordings were used.

*Calculated from Se and Sp by using a 38.1% and 61.9% distribution between SDB and non-SDB events in the Apnea-ECG database, respectively.

**SUPPLEMENTARY MATERIAL**

Supplementary Table 1 – F1 performance and statistical analysis.

<table>
<thead>
<tr>
<th></th>
<th>AHI&lt;5</th>
<th>5&lt;AH&lt;30</th>
<th>AHI&gt;30</th>
<th>CVD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLMI&lt;5</td>
<td>PLMI&lt;5</td>
<td>PLMI&lt;5</td>
<td>PLMI&lt;5</td>
</tr>
<tr>
<td></td>
<td>5&lt;PLMI&lt;30</td>
<td>5&lt;PLMI&lt;30</td>
<td>5&lt;PLMI&lt;30</td>
<td>5&lt;PLMI&lt;30</td>
</tr>
<tr>
<td></td>
<td>PLMI&gt;30</td>
<td>PLMI&gt;30</td>
<td>PLMI&gt;30</td>
<td>PLMI&gt;30</td>
</tr>
<tr>
<td>F1 µ ± σ</td>
<td>49.0 ± 17.3</td>
<td>50.5 ± 15.7</td>
<td>46.9 ± 15.9</td>
<td>83.8 ± 11.0</td>
</tr>
<tr>
<td>Number of records</td>
<td>86</td>
<td>50</td>
<td>27</td>
<td>205</td>
</tr>
</tbody>
</table>

**AHI<5**

<table>
<thead>
<tr>
<th>p value</th>
<th>PLMI&lt;5</th>
<th>5&lt;PLMI&lt;30</th>
<th>PLMI&gt;30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.52 ± 10^-24</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2.92 ± 10^-19</td>
<td>-</td>
</tr>
</tbody>
</table>

**5<AH<30**

<table>
<thead>
<tr>
<th>p value</th>
<th>PLMI&lt;5</th>
<th>5&lt;PLMI&lt;30</th>
<th>PLMI&gt;30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**AHI>30**

<table>
<thead>
<tr>
<th>p value</th>
<th>PLMI&lt;5</th>
<th>5&lt;PLMI&lt;30</th>
<th>PLMI&gt;30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**CVD**

<table>
<thead>
<tr>
<th>p value</th>
<th>PLMI&lt;5</th>
<th>5&lt;PLMI&lt;30</th>
<th>PLMI&gt;30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**F1** performance and statistical analysis of PLMI and AHI subgroups from the entire MESA cohort, and CVD subgroups from the SHHS cohort test set. Pairwise t-tests between all subgroups in each cohort and p values are presented.

AHI: Apnea-hypopnea index; PLMI: Periodic leg movement index; F1: F1 score; CVD: Cardiovascular Disease; SHHS: Sleep Heart Health Study; MESA: Multi-Ethnic Study of Atherosclerosis cohort.