Automated Sleep Apnea Quantification Based on Respiratory Movement

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Published Version
doi:10.7150/ijms.9303

Accessed
January 6, 2018 5:31:42 AM EST

Citable Link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:12406702

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

(Article begins on next page)
Automated Sleep Apnea Quantification Based on Respiratory Movement

M.T. Bianchi1,2,*, T. Lipoma3, and C. Darling3, Y. Alameddine1, M.B. Westover1

1. Neurology Department, Sleep Division, Massachusetts General Hospital, Boston MA, USA
2. Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA
3. Rest Devices, Boston, MA, USA

*Corresponding author: Dr Matt T. Bianchi, Wang 7 Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA02114. mtbianchi@partners.org; Phone: 617-724-7426; Fax: 617-724-6513

© Ivyspring International Publisher. This is an open-access article distributed under the terms of the Creative Commons License (http://creativecommons.org/licenses/by-nc-nd/3.0/). Reproduction is permitted for personal, noncommercial use, provided that the article is in whole, unmodified, and properly cited.

Received: 2014.04.05; Accepted: 2014.05.23; Published: 2014.05.30

Abstract

Obstructive sleep apnea (OSA) is a prevalent and treatable disorder of neurological and medical importance that is traditionally diagnosed through multi-channel laboratory polysomnography (PSG). However, OSA testing is increasingly performed with portable home devices using limited physiological channels. We tested the hypothesis that single channel respiratory effort alone could support automated quantification of apnea and hypopnea events. We developed a respiratory event detection algorithm applied to thoracic strain-belt data from patients with variable degrees of sleep apnea. We optimized parameters on a training set (n=57) and then tested performance on a validation set (n=59). The optimized algorithm correlated significantly with manual scoring in the validation set (R² = 0.73 for training set, R² = 0.55 for validation set; p<0.05). For dichotomous classification, the AUC was >0.92 and >0.85 using apnea-hypopnea index cutoff values of 5 and 15, respectively. Our findings demonstrate that manually scored AHI values can be approximated from thoracic movements alone. This finding has potential applications for automating laboratory PSG analysis as well as improving the performance of limited channel home monitors.

Key words: algorithm; prediction; respiration, classification; sleep apnea

Introduction

Obstructive sleep apnea (OSA) is a common disorder with implications for health, well-being, and neuro-cognitive performance. Prevalence estimates are typically in the range of 5-25% of adults, depending on the demographics and definitions of respiratory events [1, 2], with higher prevalence occurring with certain co-morbidities [3]. Despite the high prevalence and treatable nature of the disorder, many individuals with OSA remain undiagnosed, in part due to the cost, inconvenience, and variable availability of sleep laboratory testing. Although laboratory polysomnogram (PSG) remains the gold standard for diagnosis, portable diagnostic devices are becoming increasingly utilized for home testing, which holds important promise for reducing the burden of undiagnosed OSA. Although portable devices are often assumed to be cost-effective alternatives to lab PSG testing, these devices have important limitations and formal cost-effectiveness studies have questioned the widely held cost-saving assumption [4-8]. The literature shows that the sensitivity and specificity of portable devices are only modest, such that the American Academy of Sleep Medicine practice parameters restricts their use to patients with strong pre-test probability (>80%) of moderate or severe OSA [9, 10], in other words, home testing should be used only to confirm the diagnosis in those with already high clinical suspicion. However, if improved performance of home monitoring devices is accomplished, their utility as an alternative to in-lab testing...
might in the future enjoy more broad applicability. This is particularly important given increasing cost-burden associated with modern health care delivery.

Current manual scoring guidelines use multiple streams of PSG-derived data to characterize apnea and hypopnea events during sleep, including measures of effort, airflow, and pulse oximetry. In fact, a review of the evidence supporting the use of portable monitors, no single-channel devices were considered sufficiently accurate for quantifying OSA [10]. However, several research reports suggest promising results for single-channel approaches to OSA detection. For example, analysis of oximetry, nasal pressure, and load cells have shown sensitivity and specificity values in the low to mid 80s percentage range [11-15]. If the fluctuations in the multiple data streams utilized in manual scoring are internally correlated, it should be possible to extract substantial information from subsets of data. We tested the hypothesis that respiratory effort alone correlated with the apnea-hypopnea index (AHI) obtained by manual scoring. Having an easily accessible surrogate for respiratory events that conventionally require multiple channels would offer important opportunities to improve upon portable OSA testing. Specifically, to the extent that a respiratory index such as the AHI remains the clinical standard for diagnosing OSA, improvements in extracting this information are important to advance diagnostic accuracy.

Methods

Patient population

Polysomnography (PSG) data from adult men and women who underwent clinical testing in our center were selected for analysis. The Institutional Review Board approved retrospective analysis of our clinical PSG database without additional consent. The PSGs were not selected based on clinical indication, although most were conducted for evaluation of sleep apnea. Selection was based on AHI and respiratory disturbance index (RDI) values to capture a broad spectrum of severity. We did not exclude recordings based on medication or medical history. This allowed a broad spectrum of clinical features to be present and preservation of the potential generality of the findings. The population was chosen to represent typical categories of sleep disordered breathing, including: none (AHI<5), mild (AHI 5-15), moderate (AHI 15-30), and severe (AHI>30). For the main analysis, we only considered PSGs in which the RDI was similar to the AHI. Specifically, for the group with "no OSA", all had RDI <6 and AHI<5. For patients with AHI 5-15, the RDI values were <25 and always within 15 points of the AHI. For patients with AHI values >15, the RDI was always within 20 points. We separately selected individuals who showed substantially higher RDI than AHI values. For one group (n=10), the AHI was <5 while the RDI was >15, and for the second group, the AHI was >15, and the RDI was >20 points above the AHI.

The PSGs were randomly divided into two groups, ensuring that the spectrum of severity was similar in each group. An exhaustive brute-force parameter search using n=57 diagnostic PSGs was performed on the training set, to maximize the correlation coefficient between the algorithm event index and the scored AHI. We used the parameters in the training set for validation on the remaining n=59 testing set. Independent of the algorithm optimization, we manually rated the signal quality of the abdominal and chest belt signals as good, intermediate, and high-artifact. We then repeated optimization after excluding the high artifact records (n=8 in the training set, and n=5 in the validation set). In this optimization set, we also restricted analysis to time spent in any scored sleep stage (i.e., we excluded scored wake time).

The respiratory event detection algorithm was designed to search respiration signal time series data for periods of decreased respiration amplitude from the abdominal and chest effort belts for a minimum duration of time (at least 10 seconds). The raw signals from the abdomen and the chest belts were filtered and normalized, and the peaks of each inhalation were identified. The algorithm then employed an envelope tracking function that follows the respiration movement peaks, in particular following periods when the respiration amplitude falls. This tracking function responds with a lag to amplitude decrements - that is, it is limited in terms of how fast it “falls”. In this way, the decline occurs at a slower rate than the normal respiration if there is a decrement in signal, such as might occur with apneas and hypopneas. Note that even obstructive apneas and hypopneas involve ongoing effort, but within the events the amplitude of the belts is often seen to be somewhat lower amplitude. The tracking function is adaptive, that is, for each detected breath, a new amplitude threshold is defined, based on the height of the previous peak, which accommodates potential drift in breath size or belt amplitude changes over a night. If a signal amplitude decrement is encountered for a minimum time duration, a respiratory event is marked. During the optimization routine, three parameters were investigated in a brute-force test of the parameter space: the rate of “fall” of the tracking function; the threshold of the drop to score an event, and the minimum duration of that decrement in amplitude to score an event.
In each subject, the algorithm used either the abdominal belt or the chest belt depending on the signal to noise ratio in windows of 30 seconds duration.

To determine whether algorithm accuracy was influenced by factors such as age, sex, BMI, leg movements, or sleep efficiency, we conducted non-parametric correlation analysis with the absolute and relative error between the algorithm counts and the gold standard technician counts.

Results

Table 1 shows the clinical characteristics of the cohort, which included a spectrum of apnea severity. The values were similar for the restricted data set in which high-artifact individuals were removed, and this excluded group was not different in terms of Table 1 metrics (not shown).

The values were similar for the restricted data set in which high-artifact individuals were removed, and this excluded group was not different in terms of Table 1 metrics (not shown).

Figure 1 shows the relationship between the respiratory event index according to the algorithm and the technician scored AHI values from overnight PSG data. The correlations are shown for the training set and testing set in Figures 1A and 1B; both were significant to p<0.0001, but the coefficient was smaller in the validation set. We then repeated parameter optimization after excluding records judged to have excessive signal artifact (see methods) and restricting analysis to PSG epochs scored as wake. The improved correlation coefficients are shown for the training set and testing set in Figures 1C and 1D.

Table 1: Characteristics of training and testing groups

<table>
<thead>
<tr>
<th></th>
<th>Train set (n=57)</th>
<th>Validation set (n=59)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>50 (18-83)</td>
<td>54 (18-84)</td>
</tr>
<tr>
<td>Sex (% male)</td>
<td>59.6</td>
<td>52.5</td>
</tr>
<tr>
<td>BMI</td>
<td>31 (26-38)</td>
<td>28.0 (25-32)</td>
</tr>
<tr>
<td>TST (min)</td>
<td>367.3 (328.9-420.4)</td>
<td>376.0 (318.5-416.0)</td>
</tr>
<tr>
<td>Sleep latency (min)</td>
<td>6.0 (2.0-10.6)</td>
<td>5.0 (1.0-11.0)</td>
</tr>
<tr>
<td>Efficiency (%)</td>
<td>87.1 (77.1-90.7)</td>
<td>85.4 (74.7-92.2)</td>
</tr>
<tr>
<td>N1 (min)</td>
<td>48.3 (31.6-85.3)</td>
<td>41.0 (25.0-52.7)</td>
</tr>
<tr>
<td>N1 (%)</td>
<td>13.9 (8.5-21.2)</td>
<td>11.9 (6.8-16.9)</td>
</tr>
<tr>
<td>N2 (min)</td>
<td>205.5 (169.5-239.5)</td>
<td>187.5 (152.5-217.0)</td>
</tr>
<tr>
<td>N2 (%)</td>
<td>54.7 (46.7-64.0)</td>
<td>52.2 (45.9-59.5)</td>
</tr>
<tr>
<td>N3 (min)</td>
<td>45.3 (18.3-73.6)</td>
<td>60.0 (28.5-89.0)</td>
</tr>
<tr>
<td>N3 (%)</td>
<td>12.2 (5.6-19.6)</td>
<td>16.5 (7.0-25.1)</td>
</tr>
<tr>
<td>REM (min)</td>
<td>53.8 (27.5-76.5)</td>
<td>56.5 (34.5-82.0)</td>
</tr>
<tr>
<td>REM (%)</td>
<td>14.6 (9.3-20.2)</td>
<td>15.4 (10.9-21.6)</td>
</tr>
<tr>
<td>REM Lat (min)</td>
<td>122.3 (72.4-225.0)</td>
<td>112.5 (74.0-174.0)</td>
</tr>
<tr>
<td>PLMS (#/hr)</td>
<td>17 (10.6-36.8)</td>
<td>11.8 (1.2-29.1)</td>
</tr>
<tr>
<td>AHI (#/hr)</td>
<td>11.7 (3.1-24.3)</td>
<td>12.3 (3.1-22.1)</td>
</tr>
<tr>
<td>RDI (#/hr)</td>
<td>15.4 (4.5-33.0)</td>
<td>17.9 (6.7-32.9)</td>
</tr>
<tr>
<td># central events</td>
<td>2.0 (0.0-12.8)</td>
<td>2.0 (1.0-7.0)</td>
</tr>
</tbody>
</table>

Values are median with 25-75% range (except sex, which is given as % only, and age, which is given as a min-max range).

Figure 1. Correlation of algorithm with technician scored PSG AHI values. Panel A shows the scatter plot of the algorithm index against the technician-scored index for subjects in the training set. The best fit line (solid line) and 95% confidence intervals (dotted lines) are given, with the resulting R² value. Panel B shows the scatter plot and best fit line with confidence intervals for the validation set. The scatter plots and best fit lines are also shown after excluding high-artifact individuals and restricting analysis to scored sleep during the PSG, for the training set (panel C) and testing set (panel D).
Bland-Altman analysis was performed for validation set performance based on the full data set (Figure 2A) and the restricted data set of scored sleep in good quality records (Figure 2B). This analysis revealed that the algorithm tended to over-estimate the respiratory event rate for low AHI cases (by PSG criteria) and tended to under-estimate the event rate as the PSG-scored AHI value increased.

We next conducted receiver operator characteristic (ROC) analysis on the validation set data, by varying the algorithm score cutoff value for apnea-or-not classification, to determine the sensitivity, specificity, and area under the curve (AUC) for categorizing each case according to binary clinical cutoff values of PSG AHI =5 (Figure 3A and 3B) or PSG AHI =15 (Figure 3C and 3D). In each case, the analysis was performed when all subjects were included versus excluding high-artifact records and scored wake time. The AUC values using a cutoff of PSG AHI =5 cutoff were similar for the whole set and for the restricted data set (0.93 versus 0.95, respectively) (Figure 3A and 3B). At a threshold respiratory event rate value of 7.5 the algorithm, the sensitivity was 88.4%, and the specificity was 81.3%. This yields a positive likelihood ratio (LR(+) of 4.7, and LR(-) of 0.14. Considering the restricted data set, for a cutoff of AHI =5, a threshold respiratory event rate value of 7.5 for the algorithm improved the sensitivity to 90.2% and the specificity to 84.6%. The LR(+) and LR(-) values were 5.9 and 0.12, respectively. The AUC values for a cutoff of AHI =15 indicated decreased accuracy in the validation set including all subjects or the restricted set (Figure 3C and 3D). Here, the sensitivity and specificity values were ≤80% in each case.

We also investigated algorithm accuracy (optimized as above) when applied to two groups of patients with large disparities between the RDI and AHI values, one in which the AHI values were ≤5 while the RDI values were >15, and another in which AHI values were greater than 15 and the RDI exceeded the AHI by at least 20. In the former group, the correlation coefficient was 0.42 (p<0.05) using parameters optimized for the restricted data set, and was non-significant using parameters optimized for the entire data set (data not shown). In the latter group, the correlation coefficient was 0.46 (p<0.001) for the restricted set, and 0.34 (p<0.006) for the full set (data not shown). Thus, performance is sensitive to the types of respiratory events, such that the presence of excess RERAs was associated with reduced performance. That the algorithm could score individuals closer to the AHI in some cases and closer to the RDI in other cases raises the possibility that RERAs are heterogeneous and may include a subset that resemble apneas and hypopneas in terms of respiratory effort.

Finally, we performed correlation analysis to determine whether the accuracy of respiratory event rate values depended on certain clinical and PSG characteristics. Table 2 shows the correlation coefficient values for the accuracy defined as the absolute difference between PSG AHI and the algorithm score. Scoring accuracy did not vary with age, sex, BMI, or sleep architecture parameters.

Table 2: Correlations with algorithm index accuracy

<table>
<thead>
<tr>
<th>Variable</th>
<th>r-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.32</td>
</tr>
<tr>
<td>Sex (% male)</td>
<td>0.23</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.18</td>
</tr>
<tr>
<td>TST (min)</td>
<td>-0.05</td>
</tr>
<tr>
<td>Efficiency (%)</td>
<td>0.05</td>
</tr>
<tr>
<td>N1 (min)</td>
<td>0.24</td>
</tr>
<tr>
<td>N1 (%)</td>
<td>0.26</td>
</tr>
<tr>
<td>N2 (min)</td>
<td>0.05</td>
</tr>
<tr>
<td>N2 (%)</td>
<td>0.05</td>
</tr>
<tr>
<td>N3 (min)</td>
<td>-0.28</td>
</tr>
<tr>
<td>N3 (%)</td>
<td>-0.27</td>
</tr>
<tr>
<td>REM (min)</td>
<td>0.15</td>
</tr>
<tr>
<td>REM (%)</td>
<td>0.15</td>
</tr>
<tr>
<td>REM Lat (min)</td>
<td>-0.23</td>
</tr>
<tr>
<td>PLMS (#/hr)</td>
<td>0.21</td>
</tr>
<tr>
<td># central events</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Pearson r-values for the restricted data set (excluding scored wake, and excluding high-artifact records). None were significant at p<0.05
Figure 3. Algorithm performance for sleep apnea severity categorization. The area under the curve (AUC) is shown for dichotomous sleep apnea categorization using a cutoff AHI = 5 when applied to the whole data set (panel A) versus the filtered set (panel B). Below each AUC curve is the corresponding confusion matrix of the optimal performance, including sensitivity, specificity, and predictive value. Categories determined by PSG AHI are in the columns, while the cutoff values used for the algorithm to yield these performance indices are in the rows. Similar plots are shown using a cutoff value for PSG AHI = 15 (panels C and D).

Discussion

The respiratory events that occur in sleep apnea are categorized into conventional classes, such as obstructive versus central, or apnea versus hypopnea versus RERA, based on visual interpretation of a combination of physiological signals. We tested the hypothesis that the AHI can be quantified using only the respiratory effort signal, and found that respiratory effort signal contains sufficient information to quantify sleep apnea in an automated manner. Algorithm performance correlated well with manually scored AHI values obtained during routine diagnostic laboratory PSG, and the algorithm performed well across a range of sleep apnea severity. These results have implications both for automated event counting to facilitate human PSG scoring, and for improving the performance of home sleep monitoring devices that capture respiratory movements. Improving automated apnea detection has implications for improved screening for undiagnosed OSA using home monitoring devices, which are currently limited in their evidence basis.

Clinical implications

Under modern definitions, signals of respiratory effort, air flow, and oxygen saturation are involved in
scoring apneas and hypopneas, and also characterizing respiratory events as obstructive versus mixed or central. These distinctions continue to play an important role in clinical sleep medicine at diagnosis as well as during titration and treatment monitoring. Limited channel portable monitoring is increasingly available for OSA diagnostics [9, 10]. These devices typically acquire respiration effort, airflow, and oxygen saturation. There is only one portable monitor approved by Medicare in the United States that does not involve a measure of airflow and respiration effort (the WatchPAT from Itamar). None of the approved devices utilizes a single channel of physiology [9].

The standard for accuracy of a portable apnea monitor proposed by the AASM [9] was modest: a positive likelihood ratio of 5, which can be obtained for example with a device showing 84% sensitivity and 84% specificity. An LR(+) of 5 can also be achieved with other combinations, for example sensitivity as low as 50% or 25% if specificity were 90% or 95%. However, the AASM guidelines required the sensitivity to be at least 82.5% to combine with specificity in this to yield the LR minimum of 5 (that is, the minimum specificity would be 83.5%).

Our study demonstrates that changes in respiratory effort encode sufficient information regarding the respiratory pauses of sleep disordered breathing to allow for automated detection of apneas and hypopneas. Distinguishing between RERAs (contributing only to the RDI) and apneas/hypopneas (contributing to the AHI and the RDI) is primarily a concern when the two values differed substantially, especially when the AHI is in the normal range. In this setting, our algorithm performance was sub-optimal. However, when both the RDI and AHI values were greater than 5, accuracy was preserved in detection of OSA (AHI>5). In other words, although the accuracy was poor from a correlation coefficient standpoint, the algorithm was accurate in dichotomous classification of OSA in that setting, without any false negatives.

**Evidence for the diagnostic utility of single-channel physiology algorithms**

Several groups have investigated single channel approaches to OSA detection. In a cohort of patients with major cardiovascular disease and/or risk factors, two thirds of whom had AHI >15 on PSG, pulse oximetry signals showed high correlation with the AHI (>0.90), including accurate detection of Cheyne-Stokes respiration [13]. However, not all studies of oximetry have been positive [16], including important differences between automated desaturation indices compared to expert-reviewed signals [17]. In addition, inter-device measurement variance is a notable source of uncertainty when considering validation studies using oximetry [18].

Single lead ECG has been reported for detection of obstructive events using machine learning applied to data after excluding artifacts [19], as well as for distinguishing central from obstructive patterns of sleep apnea events [20]. Nasal pressure alone showed high detection accuracy across a range of OSA severity observed in a clinical population at high risk for OSA [11]. However, other studies showed that accuracy of this approach was lower in patients with diabetes [21]. In addition to using single channel analysis of routinely acquired signals, other methods for recording sleep disordered breathing have been proposed, including audio [22, 23], radio frequency analysis [24, 25], mattress sensors [26], and load cells placed under the bed [15].

**Conclusion**

The AHI remains a commonly used clinical metric of OSA presence and severity. Portable monitor measurements of the AHI, or a surrogate for it, hold important promise for improving the diagnostic utility of limited channel testing. Respiratory effort signals are easily accessible markers of respiratory pauses that would correlate with standard metrics that consider airflow, oxygenation, and EEG changes. Although clinical scoring of sleep apnea during PSG testing combines oximetry and EEG arousal, it is evident from this study that much of the information contained in these multi-channel scoring techniques, particularly information relevant for determination of apnea severity, is carried by respiratory effort in isolation. This approach has implications for automated analysis of data acquired in the laboratory setting, as well as potentially improved the diagnostic yield of home devices that employ respiratory effort. Improved screening technology is especially important for neurological populations, in whom undiagnosed OSA is prevalent, such as stroke [27], cognitive impairment [28], degenerative disorders [29], and epilepsy [30]. Treatment of this reversible cause of sleep disturbance holds promise for improving neurological function.

**Acknowledgments**

Dr Bianchi has received support from the department of Neurology, Massachusetts General Hospital, the Young Clinician Award from the Center for Integration of Medicine and Innovative Technology, and the Harvard Catalyst KL2 Medical Research Investigator Fellowship. Dr. Westover received support from the American Brain Foundation.
Competing Interests

MTB, TL, and CD are inventors on pending patents related to a tee-shirt device that measures respiration. MTB has consulting agreement with Sunovion, and is on the clinical advisory board of Foramis, Inc. YA and MBW have no competing interests to declare.

References


http://www.medsci.org