Ph.D. Dissertation of Medicine

Predicting Apnea-Hypopnea Index using Heart Rate Variability in Obstructive Sleep Apnea

폐쇄성 수면 무호흡에서 심박 변이도를 이용한 무호흡-저호흡 지수 예측

August 2023

Graduate School of Medicine
Seoul National University
Otorhinolaryngology

Pona Park

Predicting Apnea-Hypopnea Index using Heart Rate Variability in Obstructive Sleep Apnea

Jeong-Whun Kim

Submitting a Ph.D. Dissertation of
Medicine

April 2023

Graduate School of Medicine
Seoul National University
Otorhinolaryngology

Pona Park

Confirming the Ph.D. Dissertation written by
Pona Park
July 2023

Chair _____(Seal)
Vice Chair _____(Seal)
Examiner _____(Seal)
Examiner _____(Seal)
Examiner _____(Seal)

Abstract

Predicting Apnea-Hypopnea Index using Heart Rate Variability in Obstructive Sleep Apnea

Pona Park
Otorhinolaryngology, Department of Medicine
The Graduate School
Seoul National University

Background: The majority of patients with obstructive sleep apnea (OSA) do not receive timely diagnosis and treatment because of the complexity of a diagnostic test. We aimed to predict OSA severity based on heart rate variability (HRV), body mass index (BMI), and demographic characteristics.

Methods: We investigated the linear correlation between apnea-hypopnea index (AHI), and age, sex, BMI and HRV parameters in the pilot study. Models of binary classification for predicting OSA severity by estimation of AHI were constructed using 14 or 17 features including 11 or 14 HRV variables, age, sex, and BMI. Binary

classification was conducted separately using AHI thresholds of 5, 15, and 30. Sixty percent of the participants were randomly allocated to training and validation sets while the other forty percent were designated as the test set. Classifying models were developed and validated with 10-fold cross-validation using logistic regression, random forest, support vector machine, and multilayer perceptron algorithms.

Results: A total of 792 (651 men and 141 women) subjects were included in the main study. The sensitivity of the best performing binary classifier was 73.6%, 70.7%, and 78.4% when the AHI threshold criterion was 5, 10, and 15, respectively. The prediction performances of the best classifiers at AHI of 5, 15, and 30 were as follows: accuracy, 72.2%, 70.0%, and 70.3%; specificity, 64.6%, 69.2%, and 67.9%; area under the receiver operating characteristic curve, 77.2%, 73.5%, and 80.1%, respectively. In the main study, the logistic regression model using the AHI criterion of 30 showed the best classifying performance overall among all models.

Conclusion: OSA severity was fairly predicted by using HRV, BMI, and demographic characteristics. Prescreening and continuous treatment monitoring of OSA may be possible simply by measuring

HRV.

Keyword: Obstructive Sleep Apnea, Heart Rate Variability, Polysomnography, Machine Learning

Student Number: 2017-32830

Table of Contents

List of Figures and Tables

Introduction

Obstructive sleep apnea (OSA) is a disease characterized by repeated obstruction of the upper airway during sleep, leading to intermittent oxygen desaturation and frequent arousal. Patients often experience excessive daytime sleepiness, poor cognitive function, and impaired quality of life.[1] Patients who remain untreated for long periods of time have an increased risk of multiple health-related outcomes, including cardiovascular disease, stroke, and death from all causes.[2] OSA is a significant and growing public health problem, and is estimated to affect 13-33% and 16-19% of men and women, respectively.[3] However, the majority of patients with OSA do not receive timely diagnosis and treatment.[4]

The gold standard for the diagnosis of OSA is polysomnography (PSG). The examinee must spend an entire night in the hospital with a sleep specialist monitoring them while they sleep. Then, the PSG data is scored manually by the specialist according to the American Academy of Sleep Medicine guidelines.[5] Hence, PSG is not suitable for massive screening and/or repeated measurements of OSA. Considering night-to-night variability,[6,7] PSG is limited for diagnosis as well as continuous monitoring for effectiveness during treatment.

Various less complicated sleep diagnostic devices have been developed, with advantages of low prices and simple installation. These devices are portable and can be used at home without medical support. Systems developed in the context of OSA diagnosis use nasal airflow and/or oxygen saturation as parameters.[8,9] However, these measurement tools have some intrinsic limitations since mouth breathing or incorrect positioning can cause signal loss.

Measurable heart rate variability (HRV) from single lead electrocardiogram signals reflects cardiac autonomic activity and can be used to evaluate quantitative changes between normal heartbeats.[10] The association between OSA and increased sympathetic nervous system activity of HRV has been well documented since the clinical significance of HRV was introduced firstly in 1965.[11,12] Recently, further studies on the correlation between HRV factors and PSG indices have been conducted.[13,14] Currently, various devices are commercially available that provide heart rate monitoring, such as smartwatches, consumer sleep wearables, and adhesive electrocardiographic monitor patches. Such devices create the possibility of evaluating sleep-disordered breathing status of patients by measuring HRV using a heart rate monitor.

The present study aimed to predict apnea–hypopnea index (AHI) using age, sex, body mass index (BMI), and HRV parameters. We also aimed to verify the feasibility of using HRV for OSA prescreening and treatment monitoring.

Part I: A pilot study

Methods (Part I)

Study participants

We retrospectively reviewed patients who visited a department of otorhinolaryngology in Seoul National University Bundang Hospital (SNUBH) due to snoring or sleep apnea from January 2013 to December 2017. We included adult patients, age ≥ 18 years, who were treated with mandibular advancement devices or underwent OSA surgery after full-night PSG in the study. Then, we excluded subjects with any of the followings: 1) significant arrhythmias; 2) sleep disorders such as insomnia or narcolepsy; 3) inveterate use of sedatives and hypnotics; 4) specific conditions related to HRV changes (i.e., myocardial disease, diabetic neuropathy, or heart transplantation); 5) low quality data (artefacts > 20% of total sleep time); 6) total sleep time < 5 hours; and 7) awake for > 30 minutes from midnight-5 AM. A flow chart of the subject selection process is summarized in Figure 1. All participants underwent attended, in-laboratory, full-night PSG (Embla® N7000, Neurolite Advanced Medical Solutions, Belp, Switzerland). The anthropometric data of subjects were reviewed. The ethics committee of SNUBH (IRB No.

B-1907-555-108) approved the use of these data. The need for written informed consent was waived by the Institutional Review Board.

HRV analysis

HRV was measured using exported electrocardiogram data and commercially available PSG software (Embla® RemLogic™ 3.0 HRV analyzer; Neurolite Advanced Medical Solutions). This methodology was described in our previous study.[15] Calculations of the time- and frequency-domain parameters were performed according to standard methods for HRV measurements.[10] The electrocardiogram signals were interpolated and resampled at 5.0 Hz. Normal-to-normal (NN) heartbeat intervals > 2,400 ms and < 400 ms were omitted. We only analyzed the electrocardiogram signal from midnight–5 AM to maintain consistency for comparisons. Therefore, the AHI was measured again from midnight to 5 AM. We used seven time-domain and seven frequency-domain measures in the present pilot study. The time-domain measures included following parameters: (1) average normal-to-normal (NN) interval; (2) standard deviation of NN interval (SDNN); (3) standard deviation of

the 5-minute averages of NN intervals (SDANN); (4) square root of the mean of the squared differences of adjacent NN intervals (RMSSD); (5) number of pairs of adjacent NN intervals more than 50 ms (NN50 count); (6) the rate of NN50 in the total number of NN intervals (pNN50); and (7) HRV triangular index, the number of all NN intervals divided by the maximum of the density distribution. The parameters of the frequency-domain measures were as follows: (1) total power (variance of all NN intervals); (2) very low frequency (VLF; power in 0.003-0.04 Hz range); (3) low frequency (LF; power in 0.04-0.15 Hz range); (4) high frequency (HF; power in 0.15-0.4 Hz range); (5) LF/HF ratio; (6) normalized LF [LF nu; LF / (LF + HF) × 100]; and (7) normalized HF [HF nu; HF / (LF + HF) × 100]. Frequency-domain measures were yielded as the average of the values calculated every 5 minutes.

Prediction of AHI

We explored the linear relationship between AHI, and age, sex, BMI (body mass index) and the HRV parameters. Multiple linear regression in a free machine learning software (Weka; University of Waikato, Hamilton, New Zealand)[16] was used in the analysis.

Conformity between predicted AHI and true AHI was evaluated. Predicting performances of the analysis were expressed as sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC).

Development of binary classifiers of OSA

A model of OSA classification was constructed using 17 features and actual class according to AHI threshold (Figure 2A). Output was the predicted class using estimated AHI scores. The 17 features included age, sex, BMI, and the 14 HRV measures. We built the classifying model using logistic regression, random forest, support vector machine, and multilayer perceptron algorithms in Weka. Sixty percent of participants were randomly selected for the training and validation sets while the other forty percent were designated as the test set. The difference between classes in the training set was adjusted using the Synthetic Minority Oversampling Technique in Weka. Validation was conducted by applying 10-fold cross-validation. The enrolled subjects were split randomly into 10 subgroups. One subgroup was preserved for validation of the prediction model. The other nine subgroups were served for the training set. The cross-validation process was repetitively

conducted 10 times (10-fold) with all 10 subgroups. This produced 10 evaluation results, which were then averaged. The learning algorithm was then applied a final (11th) time in the whole data set to gain the final model.[16] Ultimately, the test set was included in the final training model (Figure 2B). Output was expressed as a confusion matrix and AUC.

The performance outcomes of the machine learning classifiers were mainly evaluated by sensitivity, specificity, and accuracy. The AUC was also presented to assess the performance of the classifiers. All results were extracted from the test set.

Results (Part I)

General patient characteristics

A total of 189 participants (165 men and 24 women) with a mean age of 49.3 ± 11.2 years were included in the pilot study. The mean BMI and AHI score was 26.0 ± 3.4 kg/m^2 and 28.7 ± 18.5, respectively. Participants were classified according to AHI score as follows: 9, normal (AHI < 5); 44, mild OSA (5 ≤ AHI < 15); 61, moderate OSA (15 ≤ AHI < 30); and 75, severe OSA (AHI ≥ 30). General and polysomnographic characteristics are summarized in Table 1.

Predicting performance of multiple linear regression analysis

After multiple linear regression analysis, estimated AHI was equal to 0.279Age +2.0999BMI −0.0385Avg NN interval −0.0945SDANN +0.0698RMSSD −0.0018NN50 +0.7285pNN50 +0.8397Triangular index +0.0004VLF −0.0003LF −4.2576LF/HF +0.4963LF nu −0.3409HF nu −33.27. Among 17 variables, 4 variables including sex, SDNN, Total power, and HF were excluded automatically by Weka.

A conformity map between predicted OSA severity and true OSA severity was drawn in Figure 3. Predicting performances of multiple linear regression analysis were summarized in Table 2. The analysis showed best predicting performance when AHI criterion was 30. Sensitivity was 78.7%, specificity was 74.6%, accuracy was 76.2%, and AUC was 84.8%.

Performance of classifying models

The classifying performance results derived from four algorithms with the AHI threshold for binary classification defined as 5, 15, and 30, are summarized in Table 3. When the AHI criterion was 5, logistic regression showed the best performance. The sensitivity and specificity of the model were 79.7% and 100%, respectively. The accuracy and AUC of the model were 80.3% and 89.9%, respectively. When the AHI criterion was 15, the best classifying model was also multilayer perceptron. The sensitivity and specificity of the model were 72.6% and 52.0%, respectively. The accuracy and AUC of the model were 65.8% and 70.7%, respectively. When the AHI threshold criterion reached 30, a model using random forest showed the best classifying results as follows: sensitivity, 63.3%; specificity, 80.4%; accuracy, 73.7%; and AUC, 77.5%. The logistic regression model

using the AHI criterion of 5 showed the best classifying performance among all models.

Part II: A main study

Methods (Part II)

Study participants

We retrospectively reviewed outpatients and inpatients who visited SNUBH sleep center owing to snoring or sleep apnea between January 2013 and December 2017. Adult patients, aged ≥ 18 years were included in this study. We excluded patients who underwent split-night PSG. After reviewing AHI scores, we randomly allocated patients to non-OSA and mild, moderate, and severe OSA groups proportionally to reduce selection bias. Participants were then excluded according to the same exclusion criteria as in the pilot study. A flow chart of the participant selection process is summarized in Figure 4. All subjects underwent in-laboratory, full-night PSG (Embla® N7000, Neurolite Advanced Medical Solutions, Belp, Switzerland). Age, sex, and BMI of the subjects were collected. The present study protocol was reviewed and approved by the ethics committee of SNUBH (IRB No. B-2111-723-110). The requirement for written informed consent was waived by the ethics committee owing to the retrospective nature of the study.

HRV parameters and development of binary classifiers of OSA

Fourteen HRV parameters were measured in the same manner as in the pilot study. Only ECG signals from midnight to 5 AM were analyzed. Then, AHI was measured anew from midnight to 5 AM. First, binary classification according to OSA severity was performed using 17 features including age, sex, BMI, and 7 time- and 7 frequency-domain variables in a consistent fashion as the pilot study. Next, the same process was performed using a total of 14 variables, excluding two time-domain variables and one frequency-domain variable. The excluded parameters were SDANN, HRV triangular index, and VLF. Weka displayed the output as a confusion matrix and AUC. The machine learning classifiers' performance outcomes were measured by sensitivity, specificity, accuracy, and AUC. All results came from the test set.

The final 11 HRV measures according to OSA severity are presented as the mean ± standard deviation, and were compared by one-way analysis of variance using SPSS software (version 22.0, IBM Corp., Armonk, NY, USA). Turkey, Bonferroni or Dunnett T3 test were used for post-hoc comparisons. P values < 0.05 were

considered statistically significant.

Results (Part II)

General patient characteristics

The main study involved 792 participants (651 men and 141 women) who had an average age of 55.1 ± 12.6 years. The average BMI and AHI score was 25.9 ± 3.5 kg/m^2 and 22.9 ± 19.2, respectively. Based on AHI score, participants were grouped as follows: 124, normal (AHI < 5); 223, mild OSA (5 ≤ AHI < 15); 221, moderate OSA (15 ≤ AHI < 30); and 224, severe OSA (AHI ≥ 30). Table 4 shows the general and polysomnographic characteristics.

Performance of classifying models using 17 features

Table 5 shows a summary of the results of four algorithms for binary classification using different AHI thresholds: 5, 15, and 30. The best performance for the AHI threshold of 5 was achieved by logistic regression. The model had a sensitivity of 72.5% and a specificity of 62.5%. The accuracy and AUC of the model were 71.0% and 75.6%, respectively. For the AHI threshold of 15, logistic regression analysis was the best model as well. The model had a sensitivity of 68.4% and a specificity of 67.8%. The accuracy and AUC of the model

were 68.1% and 74.2%, respectively. For the AHI threshold of 30, logistic regression was also the best model. The model had a sensitivity of 68.9% and a specificity of 72.0%. The accuracy and AUC of the model were 71.3% and 78.3%, respectively. Among all models, the logistic regression model with the AHI threshold of 30 had the best classifying performance.

Performance of classifying models using 14 features and Final 11 HRV parameters

The classifying performance results obtained from four algorithms were presented in Table 6. Logistic regression had the best performance when the AHI criterion was 5. The model had a sensitivity and specificity of 73.6% and 64.6%, respectively. The model also had an accuracy and AUC of 72.2% and 77.2%, respectively. Logistic regression was also the best classifying model when the AHI criterion was 15. The model had a sensitivity and specificity of 70.7% and 69.2%, respectively. The model also had an accuracy and AUC of 70.0% and 73.5%, respectively. Logistic regression had the best classifying results when the AHI threshold criterion was 30 as well: sensitivity, 78.4%; specificity, 67.9%;

accuracy, 70.3%; and AUC, 80.1%. The logistic regression model with the AHI criterion of 30 using 14 features had the best classifying performance among all models in the main study.

The means and standard deviations of the final 11 HRV parameters according to OSA severity are presented in Table 7. The HRV parameters presented a significant difference according to OSA severity except for RMSSD, NN50 count, and pNN50 ($P < 0.05$). Post-hoc analyses were performed on 8 HRV variables showing significant differences in Table 7. The total power, LF, LF/HF ratio, and LF nu values increased with OSA severity. The average NN interval, HF, and HF nu showed a decreasing tendency as the severity of OSA increased.

Discussion

In the pilot and main study, OSA severity was fairly predicted using age, sex, BMI, and time- and frequency-domain HRV. In the binary classifiers using 14 features of the main study, sensitivity of the best performing model was over 70% in all AHI criteria. In particular, it showed better performance with 78.4% sensitivity in discriminating severe OSA. Furthermore, the best binary classifier of the pilot study showed performance with sensitivity of 79.7% and specificity of 100% in distinguishing OSA. As the purpose of the present study was to evaluate the use of HRV in OSA prescreening rather than to develop an accurate diagnostic tool as an alternative to PSG, we determined that HRV can be an intuitive and informative tool for large-scale preliminary screening to determine whether or not a patient should undergo additional PSG. Because it is binary classification, identifying the quantitative difference in terms of treatment effect can be difficult if the patient remains at the same level of OSA severity. However, considering the simplicity of this concept, measuring HRV to determine OSA severity, HRV has the potential to become a very powerful assessment tool. The main study was conducted on a larger scale (including 792 subjects) than previous other studies that have investigated the relationship between HRV parameters and PSG

indices.[13,17,18] Furthermore, considering that OSA severity differs by race,[19] our study has significance in that it was conducted in an Asian population (in particular, Koreans) with physical characteristics that differ those of Westerners.

Our findings indicate that the best performing algorithm was logistic regression among the four machine learning algorithms. Since we aimed to confirm the feasibility of using HRV in OSA screening and treatment monitoring, the machine learning algorithm itself was not that important.

Several studies have attempted to differentiate OSA based on HRV that changes owing to sympathetic nervous system activity in hypoxic conditions. One study found that by using the difference between the daytime and nighttime standard deviations of the NN interval index, they could detect OSA with a sensitivity of 89.7% and 83% and a specificity of 98.1% and 96.5% in 91 and 52 subjects, respectively.[17] However, this study was limited in that PSG and Holter monitoring, which is used to measure HRV, were not simultaneously performed. Since heart rate is affected by various external factors, measuring the heart rate at the different times may create problems with reproducibility. Another study attempted to calculate equations for HRV variables and PSG parameters such as

AHI score, micro-arousal index, and oxygen desaturation index through multiple regression analyses in 25 patients with OSA.[13] However, the statistical significance was only evaluated using one-way analysis of variance analysis, with no evaluation of prediction accuracy. In contrast, another study attempted to develop a model for OSA detection using machine learning methods based on HRV, pulse oxygen saturation, and BMI in 148 patients with OSA and 33 non-OSA participants.[18] A proposed model showed high prediction accuracy. However, the accuracy was not reliable in that oxygen desaturation index was used as a predictor of AHI score, which is nearly the same thing as providing the correct answer in machine learning. Some studies have validated commercial fitness trackers, such as Fitbit, the Oura Ring and Whoop, which utilize HRV, by comparing them with PSG indices.[20-22] These studies compared sleep variables, such as sleep onset latency, total sleep time, sleep efficiency, and sleep stages. However, no studies have been conducted yet on the prediction of AHI score or assessment of OSA classifiers using such utilities. Moreover, further studies on the correlations between HRV parameters and PSG diagnostic indices such as AHI score have been conducted recently.[14] However, few studies have been conducted on a model of binary classification for

OSA severity by predicting AHI score using simultaneously measured HRV variables without the use of oxygen saturation levels.

Various methods for screening OSA have been introduced. The STOP-Bang questionnaire is one of the well-known sleep questionnaires to predict OSA. One study found that when total score was greater than 3, the probability for severe sleep apnea continued to increase from 31.3% (score 4) to 81.9% (score 8), while the probability for everything else (non-OSA, mild and moderate OSA) decreased.[23] The questionnaire is probably the simplest way to screen for OSA without any cost or time. However, it can be difficult to obtain information other than that the probability of severe OSA increases as the score increases. Respiratory sounds during sleep may be used for prediction of OSA severity. In our previous study, a simple algorithm for prescreening of OSA on the basis of respiratory sounds recorded during PSG was developed.[24] Accuracies of classification at AHI criteria of 5, 15, and 30 were 82.7%, 84.4%, and 85.3%, respectively. However, this modality needs to overcome technical issues such as noise cancellation. There may be restrictions on its application, for example, when a patient has a bed partner or a companion animal. In comparison, each method for screening OSA has its own strength and weakness.

When chronic intermittent hypoxia and sleep interruption are caused by OSA, the accumulation of oxidative stress and inflammatory responses, as well as increased sympathetic nervous system activity can occur.[25] HRV reflects cardiac autonomic activity, which is a balance between sympathetic and parasympathetic nervous system activity. Regarding the frequency-domain parameters of HRV, LF activity reportedly reflects both sympathetic and parasympathetic nervous system activity, while HF has been correlated to parasympathetic nervous system activity.[26] Therefore, the LF/HF ratio is regarded as the balance between sympathetic and parasympathetic nervous system activity, and has been linked with AHI scores in patients with OSA.[11] LF nu and HF nu are considered markers of sympathetic and parasympathetic nervous system activity, respectively.[26] Therefore, theoretically, the increased risk for OSA can be estimated using HRV analysis. In our study, the total power, LF, LF/HF ratio, and LF nu values increased with OSA severity. Conversely, when the OSA severity increased, the average NN interval, HF, and HF nu presented a decreasing tendency. From these results, a linear relationship can be inferred.

The development of predicting models for OSA severity by estimation of AHI are important in several respects. First, if

predictive models were applied to various electrocardiogram measuring devices such as smartwatches, wearable sleep devices, and adhesive patches, more patients would get the chance to prescreen their OSA at home on a daily basis. Indeed, considering the aforementioned limitations of PSG, many OSA patients are underdiagnosed despite the high prevalence. Individuals with recurrent snoring, daytime drowsiness, or poor quality of sleep should see a doctor firstly and undergo PSG. Ambulatory at-home monitoring devices have similar limitations in that a symptomatic patient must visit a hospital with the appropriate equipment at first. Second, given the night-to-night variability in OSA severity, this kind of prediction algorithm may assist clinicians with a more accurate diagnosis beyond the weaknesses of PSG and home testing devices by replacing repetitive tests without additional economic burden. Third, a prediction algorithm for OSA severity may be useful for verifying the therapeutic effects of OSA treatments such as mandible advancement devices or positive airway pressure in real time. Moreover, serial monitoring of the therapeutic results of OSA surgery are feasible using HRV. Conversely, such treatment effect analysis may serve as the basis for determining how to treat OSA in the future. Ultimately, early diagnosis and decisions regarding the

most effective treatment modality reduce the social and economic burden of OSA since OSA is a risk factor for multi-organ diseases, such as neurovascular, cardiovascular, and metabolic diseases.

In the pilot study and the main study, each feature for classifying model was contributed without weight in constructing the classification model. In the main study, classifying modeling was performed with 14 HRV variables at first, and after 3 HRV variables were reduced by referring to the study of Nakayama et al.[27], it was confirmed that overall classification performance for OSA severity improved. In the future, it may be possible to build a classification model that shows better classification performance by giving different weights to each feature. In addition, the process of optimizing the classification model, such as selecting the HRV variables necessary for model construction or putting different weights on each variable, can also use machine learning or even artificial intelligence.

The most suitable situation for using this study is to use the electrocardiogram data from the 24-hour Holter monitoring patch used for the diagnosis of heart disease to screen for OSA. In the future, it is expected to develop a user-friendly OSA screening device based on the HRV variables used in this study.

This study had some limitations.

First, we included a relatively small number of subjects without OSA because the participants were patients who visited a tertiary hospital for snoring or sleep apnea. Therefore, we tried to eliminate any numerical imbalance between classes by randomly selecting patients with mild, moderate, and severe OSA to proportion in the main study. In addition, we used the Synthetic Minority Oversampling Technique, a general method often used to solve this type of problem in the machine learning process.

Second, there was possibility that arrythmias which could not be ruled out with a single lead might be included during electrocardiogram analysis although significant arrhythmia was excluded from the study subject selection process. Therefore, NN intervals that were greater than 2,400 ms and less than 400 ms were excluded during electrocardiogram analysis to avoid arrhythmias as much as possible.

Third, our binary classifier may not cover all sleep stages because we only analyzed data between midnight and 5 AM. Heart rate varies by sleep stage, although we did not consider sleep stages in this study. Since rapid eye movement sleep generally occurs 90 minutes after sleep onset, it would be better to measure HRV after the onset

of deep sleep. If the sleep stage were detectable, HRV analysis could begin after one sleep cycle. However, since HRV can vary by time period because of circadian rhythms, studying measurements by particular time period may have an advantage from a certain perspective.

Fourth, considering the characteristics of the machine learning process, classifying its performance may have been affected by the distribution of patients in the training set. Indeed, there are many more people without OSA in the general population; therefore, discrepancies between research results and the real world can occur, although this can be corrected later through real-world validation.

Fifth, the external validation was not performed. Since the present study was conducted in a single institution, there may be potential bias. It can be supplemented through future multi-institutional studies.

Finally, there are hurdles to overcome in order to apply our classifying model to multiple devices. Since there is difference in the electrode for each device, validation is required to predict AHI. In the future, studies which comparing HRV of PSG and HRV of a sleep device or a smart watch are needed by wearing the multiple devices and implementing PSG simultaneously for application of our

classification model.

Conclusion

OSA severity by estimation of AHI was fairly predicted using HRV parameters, BMI, and demographic characteristics. In the final classifying model, the binary classification sensitivity for OSA was over 70% in each AHI criterion. In addition, the best binary classification model in the main study showed performance with 78% sensitivity in discriminating severe OSA. This study may have high value regarding the feasibility of HRV for OSA prescreening and continuous monitoring of treatment effects. Further real-world validation studies may lead to improvements in the classifying performance of HRV for OSA severity.

References

1. Engleman HM, Douglas NJ. Sleep. 4: Sleepiness, cognitive function, and quality of life in obstructive sleep apnoea/hypopnoea syndrome. *Thorax* 2004;59(7):618-22.

2. Xie C, Zhu R, Tian Y, Wang K. Association of obstructive sleep apnoea with the risk of vascular outcomes and all-cause mortality: a meta-analysis. *BMJ Open* 2017;7(12):e013983.

3. Senaratna CV, Perret JL, Lodge CJ, Lowe AJ, Campbell BE, Matheson MC, et al. Prevalence of obstructive sleep apnea in the general population: a systematic review. *Sleep Med Rev* 2017;34:70-81.

4. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MS, Morrell MJ, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med* 2019;7(8):687-98.

5. Kapur VK, Auckley DH, Chowdhuri S, Kuhlmann DC, Mehra R, Ramar K, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. *J Clin Sleep Med* 2017;13(3):479-504.

6. Newell J, Mairesse O, Verbanck P, Neu D. Is a one-night stay in the lab really enough to conclude? First-night effect and night-to-night variability in polysomnographic recordings among different clinical population samples. *Psychiatry Res* 2012;200(2-3):795-801.

7. White LH, Lyons OD, Yadollahi A, Ryan CM, Bradley TD. Night-to-night variability in obstructive sleep apnea severity: relationship to overnight rostral fluid shift. *J Clin Sleep Med* 2015;11(2):149-56.

8. Penzel T, Blau A, Garcia C, Schöbel C, Sebert M, Fietze I. Portable monitoring in sleep apnea. *Curr Respir Care Rep* 2012;1(2):139-45.

9. Corral-Peñafiel J, Pepin JL, Barbe F. Ambulatory monitoring in the diagnosis and management of obstructive sleep apnoea syndrome. *Eur Respir Rev* 2013;22(129):312-24.

10. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation and clinical use. *Circulation* 1996;93(5):1043-65.

11. Vanninen E, Tuunainen A, Kansanen M, Uusitupa M, Länsimies E. Cardiac sympathovagal balance during sleep apnea episodes. *Clin Physiol* 1996;16(3):209-16.

12. Somers VK, White DP, Amin R, Abraham WT, Costa F, Culebras A, et al. Sleep apnea and cardiovascular disease: an American heart association/American college of cardiology foundation scientific statement from the American heart association council for high blood pressure research professional education committee, council on clinical cardiology, stroke council, and council on cardiovascular nursing in collaboration with the national heart, lung, and blood institute national center on sleep disorders research (national institutes of health). *J Am Coll Cardiol* 2008;52(8):686-717.

13. Gong X, Huang L, Liu X, Li C, Mao X, Liu W, et al. Correlation analysis between polysomnography diagnostic indices and heart rate variability parameters among patients with obstructive sleep apnea hypopnea syndrome. *PLoS One* 2016;11(6):e0156628.

14. Santos RR, Silva TM, Silva LE, Eckeli AL, Salgado HC, Fazan R Jr. Correlation between heart rate variability and polysomnography-derived scores of severities for obstructive sleep apnea. *FASEB J* 2022;36 S1:36.

15. Kim JW, Kwon SO, Lee WH. Nocturnal heart rate variability may be useful for determining the efficacy of mandibular advancement devices for obstructive sleep apnea. *Sci Rep* 2020;10(1):1030.

16. Frank E, Hall MA, Witten IH. *The Weka Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. 4th ed. Burlington, NJ: Morgan Kaufmann; 2016.

17. Roche F, Gaspoz JM, Court-Fortune I, Minini P, Pichot V, Duverney D, et al. Screening of obstructive sleep apnea syndrome by heart rate variability analysis. *Circulation* 1999;100(13):1411-5.

18. Li Z, Li Y, Zhao G, Zhang X, Xu W, Han D. A model for obstructive sleep apnea detection using a multi-layer feed-forward neural network based on electrocardiogram, pulse oxygen saturation, and body mass index. *Sleep Breath* 2021;25(4):2065-72.

19. Pranathiageswaran S, Badr MS, Severson R, Rowley JA. The influence of race on the severity of sleep disordered breathing. *J Clin Sleep Med* 2013;9(4):303-9.

20. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing

sleep: systematic review and meta-analysis. *J Med Internet Res* 2019;21(11):e16273.

21. de Zambotti M, Rosas L, Colrain IM, Baker FC. The sleep of the ring: comparison of the ŌURA sleep tracker against polysomnography. *Behav Sleep Med* 2019;17(2):124-36.

22. Miller DJ, Lastella M, Scanlan AT, Bellenger C, Halson SL, Roach GD, et al. A validation study of the WHOOP strap against polysomnography to assess sleep. *J Sports Sci* 2020;38(22):2631-6.

23. Farney RJ, Walker BS, Farney RM, Snow GL, Walker JM. The STOP-Bang equivalent model and prediction of severity of obstructive sleep apnea: relation to polysomnographic measurements of the apnea/hypopnea index. *J Clin Sleep Med* 2011;7(5):459-65B.

24. Kim JW, Kim T, Shin J, Choe G, Lim HJ, Rhee CS, et al. Prediction of obstructive sleep apnea based on respiratory sounds recorded between sleep onset and sleep offset. *Clin Exp Otorhinolaryngol* 2019;12(1):72-8.

25. Narkiewicz K, Somers VK. Sympathetic nerve activity in obstructive sleep apnoea. *Acta Physiol Scand* 2003;177(3):385-90.

26. Lee WH, Kwon SO, Kim JW. Effectiveness of sleep surgery versus a mandibular advancement device for obstructive sleep apnea in terms of nocturnal cardiac autonomic activity. *Sleep Breath* 2020;24(4):1695-703.

27. Nakayama C, Fujiwara K, Sumi Y, Matsuo M, Kano M, Kadotani H. Obstructive sleep apnea screening by heart rate variability-based apnea/normal respiration discriminant model. *Physiol. Meas* 2019;40(12):125001.

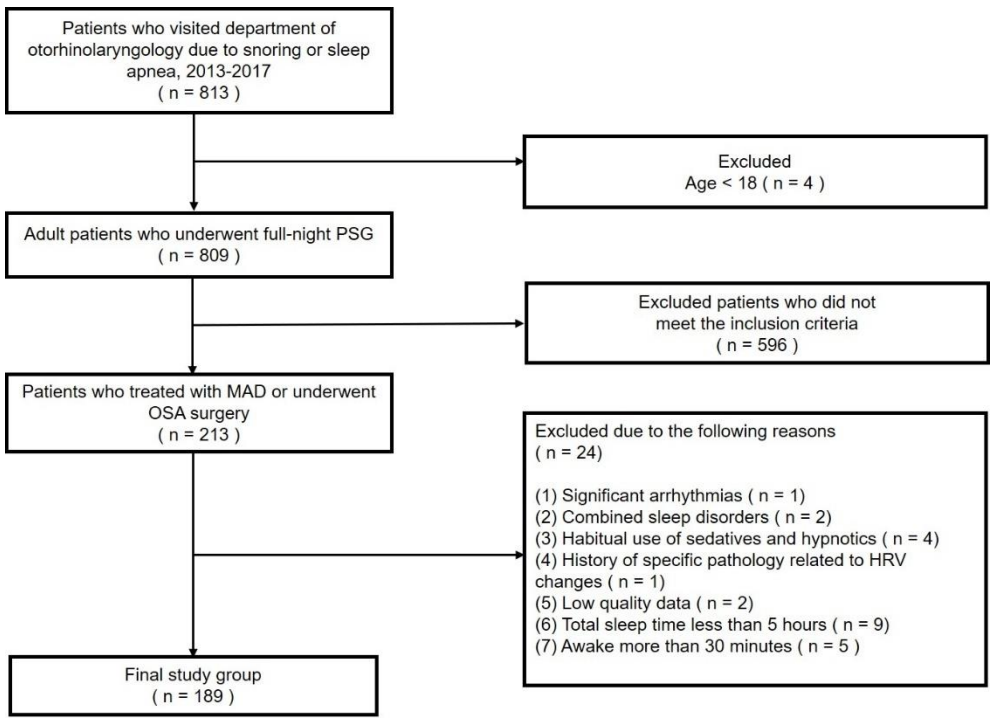

Figure 1. Flow chart of participant selection process in the pilot study. Data on outpatients who visited a department of otorhino-laryngology due to snoring or sleep apnea between 2013 and 2017 were collected. Patients who were under 18 years old were excluded. Subjects who treated with MAD or underwent OSA surgery were included. The final study participants were selected by excluding additional patients with factors likely to influence the classifying performance for OSA using heart rate variability parameters.

PSG = polysomnography, MAD= mandible advancement device, OSA = obstructive sleep apnea, HRV = heart rate variability.

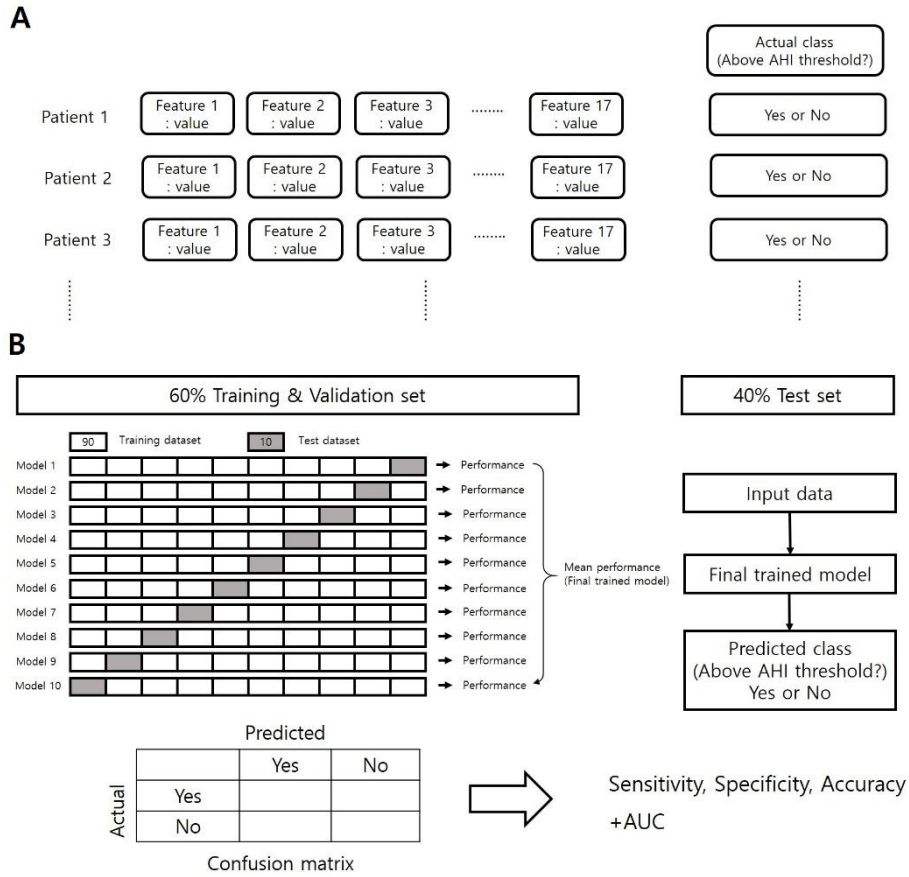

Figure 2. Study framework of machine learning process. (**A**) Seventeen features including age, sex, body mass index, and 14 measures of heart rate variability were used in the machine learning process for binary classification of obstructive sleep apnea by predicting the apnea-hypopnea index score. (**B**) Sixty percent of the data was used for training and a 10-fold cross-validation process. The remaining forty percent of data was used for the final training model. The classifying result was expressed as a confusion matrix. Sensitivity, specificity, and accuracy were calculated by the confusion matrix. AUC was calculated by the program.
AHI = apnea-hypopnea index, AUC = area under the receiver operating characteristic curve

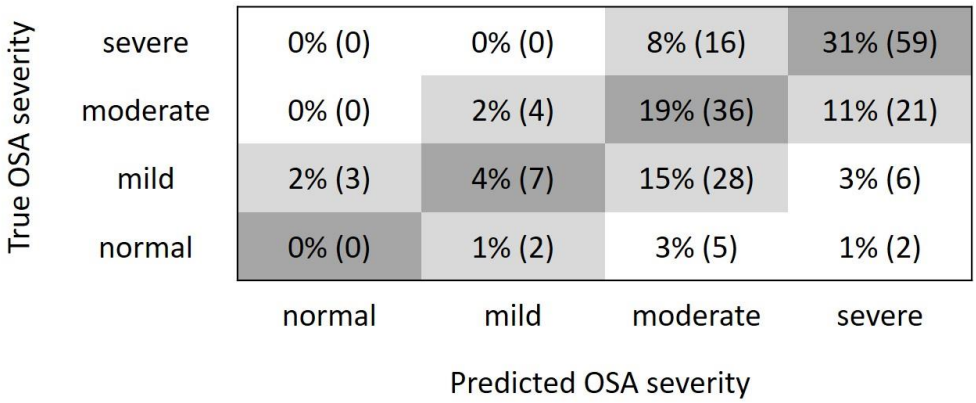

Figure 3. A conformity map of the predicted obstructive sleep apnea severity for three different apnea hypopnea index thresholds. Obstructive sleep apnea (OSA) severity was estimated through multiple linear regression analysis using heart rate variability parameters, age, and body mass index. The cell shows both the number of subjects and the percentage for each OSA severity level (which is also indicated by the color scale).

OSA = obstructive sleep apnea.

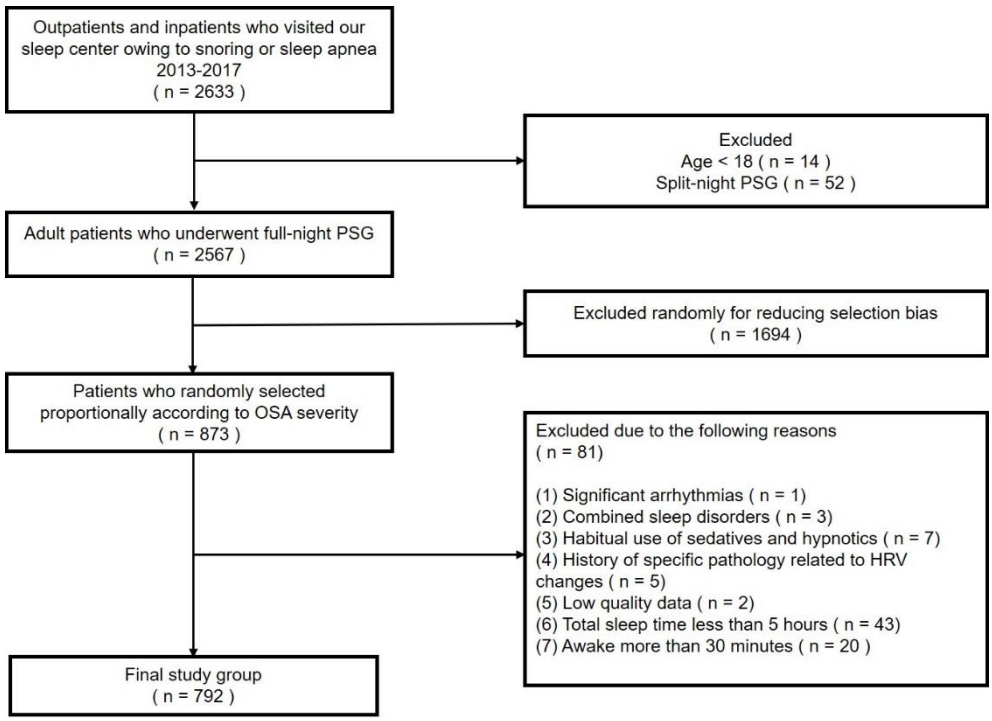

Figure 4. Flow chart of subject selection process in the main study. Data on outpatients and inpatients who visited our sleep center due to snoring or sleep apnea between 2013 and 2017 were collected. Patients who were under 18 years old and underwent split-night polysomnography were excluded. To reduce selection bias, patients were randomly allocated proportionally according to OSA severity. The final study group was selected by excluding additional patients with factors likely to influence the classifying performance for OSA using heart rate variability parameters.

HRV = heart rate variability, OSA = obstructive sleep apnea, PSG = polysomnography.

Table 1. General and polysomnographic characteristics of the pilot study

Variables	AHI < 5 (n = 9)	5 ≤ AHI < 15 (n = 44)	15 ≤ AHI < 30 (n = 61)	AHI ≥ 30 (n = 75)	*P* value
Age, yr	40.7±12.5	49.4±11.4	52.1±11.1	47.9±10.2	0.003
M / F	7:2	33:11	52:9	73:2	0.015
BMI, kg/m^2	24.1±3.2	24.8±3.2	25. 4±2.8	27.4±3.4	< 0.001
AHI, per hour	3.1±1.1	10.1±2.7	22.1±4.4	47.9±12.4	< 0.001

AHI = apnea hypopnea index, M / F = Male/Female ratio , BMI= body mass index.

Table 2. Predicting performance of multiple linear regression analysis in the pilot study

Criteria	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
AHI 5	93.7	98.3	0.0	72.9
AHI 15	76.2	97.1	22.6	79.8
AHI 30	76.2	78.7	74.6	84.8

AUC = area under receiver operating characteristic curve, AHI = apnea-hypopnea index

Table 3. Classification performances using 17 features in the pilot study

Criteria	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
AHI 5	Logistic regression	80.3	79.7	100.0	89.9
	Random forest	93.4	96.0	0.0	67.2
	SVM	97.4	100.0	0.0	50.0
	Multilayer perceptron	82.9	85.1	0.0	66.2
AHI 15	Logistic regression	67.1	70.6	60.0	71.7
	Random forest	57.9	66.7	40.0	59.9
	SVM	67.1	100.0	0.0	50.0
	Multilayer perceptron	65.8	72.6	52.0	70.7
AHI 30	Logistic regression	64.5	53.3	71.7	65.2
	Random forest	73.7	63.3	80.4	77.5
	SVM	60.5	0.0	100.0	50.0
	Multilayer perceptron	67.1	56.7	73.9	67.0

AUC, area under receiver operating characteristic curve; AHI, apnea hypopnea index; SVM, support vector machine.

Table 4. General and polysomnographic characteristics of the main study

Variables	AHI < 5 (n = 124)	5 ≤ AHI < 15 (n = 223)	15 ≤ AHI < 30 (n = 221)	AHI ≥ 30 (n = 224)	*P* value
Age, yr	48.9 ± 14.0	56.1 ± 11.9	57.8 ± 11.8	54.7 ± 12.0	< 0.001
M / F	86:38	168:55	184:37	213:11	< 0.001
BMI, kg/m^2	24.0 ± 3.1	25.3 ± 3.5	25.9 ± 2.9	27.6 ± 3.4	< 0.001
AHI, per hour	2.6 ± 1.4	9.8 ± 2.7	21.4 ± 4.0	48.6 ± 14.8	< 0.001

AHI = apnea hypopnea index, M / F = Male/Female ratio , BMI= body mass index.

Table 5. Classification performances using 17 features in the main study

Criteria	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
AHI 5	Logistic regression	71.0	72.5	62.5	75.6
	Random forest	78.2	86.2	33.3	70.5
	SVM	15.1	0.0	100.0	50.0
	Multilayer perceptron	75.4	80.7	45.8	61.2
AHI 15	Logistic regression	68.1	68.4	67.8	74.2
	Random forest	66.2	67.2	65.0	70.2
	SVM	54.9	100.0	0.0	50.0
	Multilayer perceptron	63.7	62.1	65.7	67.5
AHI 30	Logistic regression	71.3	68.9	72.0	78.3
	Random forest	73.5	64.9	76.1	76.9
	SVM	23.3	100.0	0.0	50.0
	Multilayer perceptron	65.6	74.3	63.0	76.1

AUC, area under receiver operating characteristic curve; AHI, apnea hypopnea index; SVM, support vector machine.

Table 6. Classification performances using 14 features in the main study

Criteria	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
AHI 5	Logistic regression	72.2	73.6	64.6	77.2
	Random forest	80.1	87.0	41.7	74.0
	SVM	15.1	0.0	100.0	50.0
	Multilayer perceptron	76.3	81.8	45.8	62.9
AHI 15	Logistic regression	70.0	70.7	69.2	73.5
	Random forest	62.8	67.2	57.3	69.5
	SVM	54.9	100.0	0.0	50.0
	Multilayer perceptron	65.6	77.6	51.0	71.9
AHI 30	Logistic regression	70.3	78.4	67.9	80.1
	Random forest	74.1	66.2	76.5	77.2
	SVM	23.3	100.0	0.0	50.0
	Multilayer perceptron	60.3	81.1	53.9	75.8

AUC = area under receiver operating characteristic curve, AHI = apnea–hypopnea index, SVM = support vector machin

Table 7. Final 11 heart rate variability characteristics according to obstructive sleep apnea severity

Variables	AHI < 5 (Group 1)	5 ≤ AHI < 15 (Group 2)	15 ≤ AHI < 30 (Group 3)	AHI ≥ 30 (Group 4)	*P* value
Average NN interval, ms	993.6 ± 109.5	984.0 ± 116.9	989.7 ± 129.2	937.8 ± 123.3	< 0.001[a]
SDNN, ms	98.7 ± 41.4	92.5 ± 37.5	92.0 ± 30.8	101.5 ± 52.1	0.043[b]
RMSSD, ms	69.4 ± 64.7	63.1 ± 56.2	59.4 ± 42.4	61.6 ± 47.8	0.386
NN50 count	3,345.1 ± 3,096.6	2,780.4 ± 2,698.7	2,875.8 ± 2,940.3	3,047.5 ± 2,869.2	0.328
pNN50, %	20.1 ± 19.5	15.8 ± 16.2	15.8 ± 16.4	15.8 ± 14.4	0.069
Total power, ms^2	42,784.8 ± 20,451.1	43,610.0 ± 20,253.2	50,459.8 ± 48,297.3	59,346.9 ± 36,301.6	< 0.001[c]
LF, ms^2	13,334.0 ± 7,085.8	13,835.2 ± 7,808.8	14,494.5 ± 7,632.1	18,414.9 ± 11,904.5	< 0.001[d]
HF, ms^2	7,780.8 ± 3,501.2	7,080.6 ± 3,615.8	7,081.2 ± 3,823.0	6,412.5 ± 3,089.1	0.006[e]
LF/HF ratio	2.0 ± 1.2	2.4 ± 1.8	2.8 ± 1.8	3.5 ± 3.4	< 0.001[f]

LF nu	61.2 ± 13.0	63.4 ± 14.5	64.4 ± 14.9	70.3 ± 11.9	< 0.001[g]
HF nu	38.5 ± 13.0	35.5 ± 14.0	33.8 ± 13.8	28.1 ± 11.3	< 0.001[h]

AHI = apnea-hypopnea index, NN = normal to normal, SDNN = standard deviation of NN intervals, RMSSD = square root of the mean of the squared differences of adjacent NN intervals, NN50 count = number of pairs of adjacent NN intervals more than 50 ms, pNN50 = rate of NN50 in the total number of NN intervals, LF = low frequency, HF = high frequency, LF nu = LF power in normalized units, HF nu = HF power in normalized units.

Post-hoc analyses. [a,d,g]Group 4 vs. Group 1, 2 and 3 (all $P < 0.001$); [b]Group 2 vs. Group 4 ($P = 0.002$); [c]Group 4 vs. Group 1, 2 and 3 ($P < 0.001$, $P < 0.001$ and $P = 0.004$); [e]Group 1 vs. Group 4 ($P < 0.001$); [f]Group 4 vs. Group 1, 2 and 3 (all $P < 0.001$), Group 1 vs. Group 3 ($P = 0.005$); [h]Group 4 vs. Group 1, 2 and 3 (all $P < 0.001$), Group 1 vs. Group 3 ($P = 0.01$).

Abstract in Korean

폐쇄성 수면무호흡에서 심박 변이도를 이용한 무호흡-저호흡 지수 예측

박보나
의학과 , 이비인후과학 전공
서울대학교 대학원

배경: 폐쇄성 수면 무호흡 환자의 대부분은 진단 검사의 복잡함 때문에 적시에 진단 및 치료를 받지 못한다. 본 연구에서는 심박 변이도 변수, 나이, 성별, 체질량지수를 기반으로 무호흡-저호흡 지수를 예측함으로써 폐쇄성 수면 무호흡을 선별할 수 있는 모델을 개발하고자 했다.

방법: 선행 연구에서 무호흡-저호흡 지수와 연령, 성별, 체질량 지수 및 심박 변이도 변수들 간의 선형적 상관관계를 조사했다. 여러 심박 변이도 변수들, 연령, 성별, 체질량지수를 이용하여 폐쇄성 수면 무호흡의 중증도를 예측하기 위한 이진 분류모델을 구축하였다. 이진 분류는 무호흡-저호흡 지수 5, 15, 30을 기준으로 각각 수행하였다. 연구 대상자의 60%는 훈련 및 검증군에 무작위로 할당되었고, 나머지 40%는 테스트군으로 지정되었다. 분류 모델은 로지스틱 회귀, 랜덤 포레스트, 서포트 벡터 머신 및 다층 퍼셉트론 알고리즘을 사용하여 10배 교차 검증으로

개발 및 검증되었다.

결과: 본 연구에는 총 792명(남성 651명, 여성 141명)의 대상자가 포함되었다. 무호흡-저호흡 지수 기준이 5, 15, 30일 때 최고 성능을 보이는 분류 모델의 민감도는 각각 73.6%, 70.7%, 78.4%였다. 그리고 무호흡-저호흡 지수 5, 15, 30 기준일 때 최고 예측 능력은 각각 정확도, 72.2%, 70.0%, 70.3%; 특이도, 64.6%, 69.2%, 67.9%; 수신기 작동 특성 곡선 아래의 면적, 77.2%, 73.5%, 80.1% 였다. 본 연구에서는 무호흡-저호흡 지수 30을 기준으로 사용한 로지스틱 회귀 모델이 모든 모델 중 전반적으로 최고의 분류 성능을 보였다.

결론: 폐쇄성 수면 무호흡 중증도는 심박 변이도, 나이, 성별, 체질량 지수를 기반으로 비교적 정확하게 예측되었다. 이는 단순히 심박 변이도를 측정하는 것만으로 폐쇄성 수면 무호흡의 사전 선별 및 지속적인 치료 경과 추적이 가능할 수도 있다는 것을 보여준다.

주요어: 폐쇄성 수면 무호흡, 심박 변이도, 수면 다원 검사, 기계 학습

학 번: 2017-32830