International Journal of General Practice Nursing, 2025, Volume 3, Issue 3
https://ojs.bbwpublisher.com/index.php/IJGPN
ISSN Online: 2981-9423
ISSN Print: 2981-9415

Analysis of Risk Factors Associated with Obstructive Sleep Apnea Based on a Classification Tree Model

Peng Li[1,2,3], Zirui Xu[1,2,3], Jimao Rong[4], Yan Zhou[1,2,3]*

[1]The Laboratory of Respiratory Disease, The First Affiliated Hospital of Guilin Medical University, Guilin 541001, Guangxi, China
[2]The Key Laboratory of Basic Research on Respiratory Diseases, Health Commission of Guangxi Zhuang Autonomous Region, Guilin 541001, Guangxi, China
[3]The Key Laboratory of Respiratory Diseases, Education Department of Guangxi Zhuang Autonomous Region, The First Affiliated Hospital of Guilin Medical University, Guilin 541001, Guangxi, China
[4]Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Guilin Medical University, Guilin 541001, Guangxi, China

***Corresponding author:** *Yan Zhou, yanzhou9988@sina.com*

Copyright: © 2025 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: *Objectives:* To explore the effects of various factors on the risk of obstructive sleep apnea (OSA) based on a classification tree model, in order to provide a scientific basis for the prevention of OSA in high-risk groups. *Methods:* Patients from the outpatient department, inpatient department, and physical examination center of the First Affiliated Hospital of Guilin Medical University who were treated for OSA-related symptoms from 2010 to 2022 were continuously included as study subjects. All study subjects received overnight polysomnographic monitoring (PSG), and were divided into the OSA group and control group based on PSG monitoring results. The demographic characteristics, lifestyle, blood pressure index, and laboratory index of the two groups were compared and analyzed. An undersampling method was applied to the OSA group to generate a case group, and the case group and control group were used as research objects to construct a classification tree model to screen the risk factors of OSA, and a cross-validation method and ROC curve were used to evaluate the model. *Results:* There were 1053 subjects after undersampling, including 517 in the case group and 536 in the control group. Compared with the control group, the age, male prevalence rate, smoking rate, and alcohol consumption rate of the case group were increased, and the levels of UA, TG, CHOL, LDLc, and FPG were increased, while the levels of HDLc were decreased, with statistical significance ($P < 0.05$). A total of 7 explanatory variables affecting OSA were included in the classification tree model, which were obesity, smoking history, age, drinking history, hypertension, abnormal glucose metabolism, and gender, among which obesity was the most important influencing factor. The re-substitution estimators and cross-validation estimators of the model were 0.192 and 0.200, respectively, and the standard errors were both 0.012. The area under the receiver operating characteristic (ROC) curve (AUC) value was 0.880 (95%CI:0.860~0.901), indicating that the model had a good prediction effect. *Conclusions:* (1) The main influencing factors of OSA were obesity, smoking history, age, drinking history, hypertension, abnormal glucose metabolism, and gender. (2) Although men are an independent risk factor for OSA, in the context of no obesity and no smoking history, people should pay more attention to pre-menopausal women with hypertension with OSA-related symptoms and middle-aged and elderly women above the age of perimenopause without a history of alcohol consumption. (3) Among

the metabolic diseases associated with OSA, glucose metabolism abnormalities may be the most important, and this association is independent of the confounding effects of obesity and metabolic syndrome.

Keywords: OSA; Risk factors; Classification tree model

Online publication: September 8, 2025

1. Introduction

Obstructive sleep apnea is a clinical syndrome caused by repeated upper airway collapse during sleep, resulting in apnea and hypoventilation, and the most important pathophysiological trait is Chronic intermittent hypoxia (CIH), which can lead to multi-organ and multi-system damage, such as cardiovascular and cerebrovascular diseases, diabetes mellitus, and so on. Currently, the complications caused by OSA, including cardiovascular and cerebrovascular diseases, diabetes mellitus, and so on, have become a serious public health problem [1]. According to the data from the global phenotypic study of OSA prevalence risk conducted by Benjafield et al, the overall global prevalence rate of OSA reaches up to 12.8%–13.7% [2]. From now on, polysomnography (PSG) is still the gold standard for diagnosing OSA [3]. However, it is unable to meet the huge demand for OSA diagnostics due to its disadvantages, including the need for expensive instrumentation and a long time to test and analyze. The complexity and diversity of clinical symptoms and metabolic profiles of OSA patients have created an urgent need to explore the intrinsic links between various clinical symptoms and metabolic profiles. The classification tree model can automatically produce feature importance, divide the total study population into several relatively homogeneous subintervals according to the features, and display them in a tree diagram, with a clearer and more direct result of the output. Then, to decrease the economic burden associated with OSA, the application of classification tree models to the analysis of risk factors for OSA may provide a new, concise, and efficient way for the initial screening and diagnosis of high-risk groups for OSA in clinical practice [4].

2. Methods

2.1. Research subjects and experimental design

The authors conducted a retrospective case-control study, and continuously collected 2258 patients who visited the First Affiliated Hospital of Guilin Medical University (including the outpatient department, inpatient department, and physical examination center) from 2020 to 2022 because OSA OSA-related symptoms and had complete information. All patients received PSG, and according to the results, 1722 patients who met the diagnostic criteria of OSA were divided into the OSA group; 536 patients who did not meet the diagnostic criteria of OSA during the same period of consultation were divided into the control group. Because the sample size of the OSA group was significantly higher than control group, and in order to avoid classification models having a bias towards the majority class and ignoring the minority class, which could reduce the classification accuracy, the study data were undersampled to improve the predictive accuracy of the model [5–6]. Selecting about 30% of the samples randomly from the OSA group by the sample function of R (Version 4.1.3) to generate the case group, in order to achieve a 1:1 match in the number of cases between the case group and the control group. At the same time, 70% of the remaining samples from the OSA group after sampling were used as the post-sampling residual group. Compared the two independently split samples after sampling (the

case group and the post-sampling residual group), which demonstrated that the difference in the data generated from the sampling error in the undersampling process was not statistically significant. Then used the pre-sampling inter-group comparison (OSA group and control group), post-sampling inter-group comparison (case group and control group) to prove that the OSA group and the case group in the process of comparing with the control group, respectively, their assessment of the results of the differences between the indicators were consistent, hence proved that the sampling results could be a better representation of the total sample, namely, the case group could be representative of the OSA group. Finally, using the case group and the control group as the research subjects, the classification tree algorithm was used to establish the OSA primary screening model for data analysis. The included criteria for OSA patients were (1) age≥18 years, (2) male or female, and (3) diagnosis of OSA made by a pulmonologist. The excluded criteria for OSA patients were (1) being treated for OSA, (2) having other sleep apnea, (3) having obstructive pulmonary disease, or (4) taking sedative or hypnotic medicine.

The included criteria for controls were (1) age≥18 years, (2) male or female, (3) without a history of chronic disease, (3) without dysfunction of the heart, liver, or kidney. The excluded criteria for controls were (1) having sleep apnea, (2) having obstructive pulmonary disease, or (3) taking sedative or hypnotic medicine.

2.2. Survey and measurement methods

The OSA was diagnosed according to the American Academy of Sleep Medicine Clinical Practice Guideline [7]. OSA was diagnosed based on the record from standard polysomnography (PSG) in the hospital. The mild and moderate OSA was identified when the apnea-hyponea index (AHI) was between 5 and 15, whereas the severe ones were more than 30.

The study protocol was approved by the institutional review board at the Affiliated Hospital of Guilin Medical University and conformed to the Declaration of Helsinki. Written informed consent was obtained from each subject.

2.3. Statistical methods

The data were handled by statistical software such as SPSS 26.0 and R (Version 4.1.3), and measurement data were expressed as mean ± standard deviation (Mean ± SD) or median (lower quartile, upper quartile) [M (P25, P75)], and comparisons were made by *t* test and rank sum test; count data were expressed as constituent ratio or rate (%). The $x2$ test or the exact test of Fisher was used to compare. The difference was statistically significant, while $P < 0.05$, and the sample function of R (Version 4.1.3) was used to accomplish the undersampling process of the OSA group. A classification tree model was constructed, and the fitting effect and predictive efficacy of the model were evaluated by 10-fold cross-validation and receiver operating characteristic (ROC) curve.

3. Results

3.1. Basic situation

This study collected relevant data of 2258 patients completely, including 1533 males and 725 females. All patients were divided into the OSA group and the control group by polysomnography results; the OSA group consisted of 1722 patients, including 1242 males and 480 females; the control group consisted of 536 patients, including 291 males and 245 females. The results of univariate analysis showed that, compared with the control group, the levels of age, male prevalence rate, smoking rate, alcohol consumption rate in OSA group were

increased, and the levels of indexes included UA, TG, CHOL, LDLc and FPG were increased, while the levels of HDLc were decreased, with statistical significance ($P < 0.05$). But the comparisons of the indicators of BUN, Cr, AST, ALT, and γ-GT between the two groups were statistically insignificant ($P > 0.05$), as shown in **Table 1**.

Table 1. Comparison of study data between the OSA group and the control group

Variable	**OSA group (*n*=1722)**	**Control group (*n*=536)**	***Z*/χ^2-value**	***P*-value**
Gender	1242 (Male) 480 (Female)	291 (Male) 245 (Female)	59.643	<0.001
Age (years)	47.00 (41.00,53.00)	43.00 (35.25,51.00)	7.243	<0.001
BMI (kg/m²)	28.05 (25.75,29.05)	23.70 (21.80,25.66)	23.069	<0.001
Smoking history, n (%)	933 (54.2%)	67 (12.5%)	287.82	<0.001
Drinking history, n (%)	811 (47.1%)	47 (8.8%)	254.88	<0.001
BP (mmHg)				
Systolic	124 (114,135)	117 (107,129)	8.545	<0.001
Diastolic	78 (71,86)	75 (67,82)	6.394	<0.001
UA (μmol/L)	370.00 (305.30,434.00)	335.30 (266.10,390.73)	8.214	<0.001
BUN (mmol/L)	4.85 (4.15,5.62)	4.71 (3.96,5.65)	1.574	0.115
Cr (μmol/L)	77.50 (67.80,86.80)	74.75 (65.13,88.78)	1.213	0.225
AST (U/L)	19.10 (16.70,22.20)	19.10 (16.00,23.28)	0.468	0.64
ALT (U/L)	19.9 (14.9,26.9)	19.3 (13.6,26.9)	1.697	0.09
TG (mmol/L)	1.50 (1.07,2.18)	1.31 (0.88,1.85)	6.13	<0.001
CHOL (mmol/L)	4.76 (4.24,5.25)	4.26 (3.77,4.75)	12.391	<0.001
LDLc (mmol/L)	3.17 (2.67,3.63)	2.90 (2.51,3.38)	5.785	<0.001
HDLc (mmol/L)	1.23 (1.06,1.44)	1.37 (1.16,1.62)	-9.309	<0.001
FPG (mmol/L)	5.90 (5.50,6.30)	5.30 (5.10,5.80)	16.538	<0.001
γ-GT (U/L)	30.45 (21.20,48.50)	31.00 (23.00,45.55)	-1.343	0.179
AHI (enents/h)	35.75 (20.80,55.60)	2.60 (1.40,3.90)	35.001	<0.001
LSaO2 (%)	73 (61,80)	87 (84,91)	-28.633	<0.001

3.2. Comparison between two independent segmented samples after sampling

Because the sample size of the OSA group was significantly higher than control group, and in order to avoid classification models having a bias towards the majority class and ignoring the minority class, which could reduce the classification accuracy, the study data were undersampled to improve the predictive accuracy of the model. The sample function of R (Version 4.1.3) was used to randomly select an appropriate 30% of the samples from the OSA groups to generate case group, so that the numbers of cases in case group matched the number of cases in control group by about 1:1. Meanwhile, the remaining 70% samples from the OSA group after sampling were used as the post-sampling residual group. Two independent segmented samples (case group and post-sampling residual group) after sampling were compared between groups. The results showed that differences were statistically insignificant between the variables of the two independently segmented samples (*P*

> 0.05). So, the authors considered that the differences in the data generated by the sampling error in the process of under-sampling were statistically insignificant, and we could select the case group for which the sample size was close to the control group from them for the subsequent data analysis (**Table 2**).

Table 2. Comparison of research data between two independent segmented samples after sampling

Variable	Case group (*n*=517)	Post-sampling residual group (*n*=1205)	Z/χ^2-value	*P*-value
Gender	370 (Male) 147 (Female)	872 (Male) 333 (Female)	0.115	0.735
Age (years)	47 (42,53)	46 (41,53)	-1.621	0.105
Smoking history	277	656	0.108	0.742
Drinking history	240	571	0.135	0.713
BMI (kg/m^2)	28.02 (25.55,29.18)	28.07 (25.81,28.99)	0.207	0.836
BP (mmHg)				
Systolic	123 (114,135)	124 (114,135)	0.474	0.636
Diastolic	78.00 (71.00,86.00)	79.00 (71.00，85.50)	0.281	0.779
UA (μmol/L)	370.0 (297.6,434.0)	370 (307,434)	0.52	0.603
BUN (mmol/L)	4.93 (4.16,5.64)	4.83 (4.15,5.56)	-1.709	0.28
Cr (μmol/L)	77.90 (67.65,86.90)	77.30 (67.80,86.70)	-0.332	0.74
AST (U/L)	19.40 (16.75,22.00)	19.00 (16.70,22.30)	-0.266	0.79
ALT (U/L)	19.70 (14.60,26.85)	20.00 (15.00,26.95)	0.96	0.337
TG (mmol/L)	1.45 (1.02,2.16)	1.51 (1.10,2.18)	1.544	0.123
CHOL (mmol/L)	4.79 (4.30,5.25)	4.74 (4.22,5.25)	-1.408	0.295
LDLc (mmol/L)	3.17 (2.68,3.63)	3.17 (2.66,3.63)	-0.204	0.838
HDLc (mmol/L)	1.23 (1.06,1.46)	1.23 (1.06,1.43)	-0.786	0.432
FPG (mmol/L)	5.9 (5.6,6.3)	5.9 (5.5,6.2)	-0.569	0.569
γ-GT (U/L)	29.90 (20.40,49.80)	30.50 (21.35,47.65)	0.497	0.619
HCY (μmol/L)	11.70 (10.60,12.70)	11.60 (10.50,12.55)	-1.298	0.194
AHI (events/h)	37.40 (18.65,57.55)	35.40 (21.20,54.70)	0.318	0.750
LSaO2 (%)	73 (61,79)	73 (61,80)	0.774	0.439

3.3. Comparison between two groups after sampling

After approving that the differences in the data generated by the sampling error in the process of under-sampling were not statistically significant, the authors further accomplished the comparison between the two groups after sampling. The result of comparison between two groups after sampling showed that compared with control group, the levels of age, male prevalence rate, smoking rate and drinking rate of case group were increased, the levels of indexes such as UA, TG, CHOL, LDLc and FPG were increased, and the levels of HDLc were decreased, with statistical significance ($P < 0.05$). The comparison of BUN, Cr, AST, ALT, and γ-GT between the two groups was statistically insignificant ($P > 0.05$), as shown in **Table 3**.

Table 3. Comparison of study data between the case group and the control group

Variable	case group (*n*=517)	Control group (*n*=536)	Z/x^2-value	*P*-value
Gender	370 (Male) 147 (Female)	291 (Male) 245 (Female)	33.61	<0.001
Age (years)	47.00 (42.00,53.00)	43.00 (35.25,51.00)	6.611	<0.001
BMI (kg/m²)	28.01 (25.55,29.18)	23.70 (21.80,25.66)	18.352	<0.001
Smoking history	277	67	201.885	<0.001
Drinking history	240	47	188.18	<0.001
BP (mmHg)				
Systolic	123 (114,135)	117 (107,129)	6.567	<0.001
Diastolic	78 (71,86)	75 (67,82)	4.932	<0.001
UA (μmol/L)	370.0 (297.6,434.0)	335.30 (266.10,390.73)	6.242	<0.001
BUN (mmol/L)	4.93 (4.16,5.64)	4.71 (3.96,5.65)	1.806	0.071
Cr (μmol/L)	77.90 (67.65,86.90)	74.75 (65.13,88.78)	1.083	0.279
AST (U/L)	19.40 (16.75,22.00)	19.10 (16.00,23.28)	0.531	0.595
ALT (U/L)	19.70 (14.60,26.85)	19.30 (13.60,26.90)	0.828	0.408
TG (mmol/L)	1.45 (1.02,2.16)	1.31 (0.88,1.85)	3.93	<0.001
CHOL (mmol/L)	4.79 (4.30,5.25)	4.26 (3.77,4.75)	10.561	<0.001
LDLc (mmol/L)	3.17 (2.68,3.63)	2.90 (2.51,3.38)	4.802	<0.001
HDLc (mmol/L)	1.23 (1.06,1.46)	1.37 (1.16,1.62)	-6.994	<0.001
FPG (mmol/L)	5.9 (5.6,6.3)	5.3 (5.1,5.8)	13.065	<0.001
γ-GT (U/L)	29.90 (20.40,49.80)	31.00 (23.00,45.55)	-1.34	0.18

3.4. Variable handling

In order to construct a classification tree model conveniently, and to make the output result of the model more explicit and intuitive, the above statistically significant single factors, excluding age, were assigned to the variables with strict reference to the diagnostic criteria of the previous study indicators (**Table 4**).

Table 4. Variable assignment table

Variable	Assignment situation
Obesity	Non obesity =0，Obesity =1
Gender	Female=0，Male =1
Drinking history	No history of drinking=0, Drinking history=1
Smoking history	No history of smoking=0，Smoking history =1
Hypertension	No hypertension=0，Hypertension =1
Dyslipidemia	No dyslipidemia =0，Dyslipidemia =1
Hyperuricemia	No hyperuricemia =0，Hyperuricemia =1
Abnormal glucose metabolism	No abnormal glucose metabolism =0，Abnormal glucose metabolism =1

3.5. Comparison of categorical variables before and after sampling

After accomplishing the assignment of each categorical variable, a chi-square test was performed by using the assignment results of categorical variables between the pre- and post-sampling groups. The study showed that: compared with the control group, obesity rate, male prevalence rate, drinking rate, hypertension rate, dyslipidemia rate, hyperuricemia rate, and abnormal glucose metabolism rate were all increased, with statistical significance ($P < 0.05$), as shown in **Tables 5** and **6**.

Table 5. Comparison of categorical variables between the OSA group and the control group

Variable	OSA group (*n*=1722)	Control group (*n*=536)	x^2-value	*P*-value
Obesity, n (%)	896 (52.0%)	46 (8.5%)	317.397	<0.001
Sex, n (%)	Male 1242 (72.1%)	Male 291 (54.2%)	59.643	<0.001
Drinking history, n (%)	811 (47.0%)	47 (8.7%)	254.878	<0.001
Smoking history, n (%)	933 (54.1%)	67 (12.5%)	287.820	<0.001
Hypertension, n (%)	920 (53.4%)	149 (27.7%)	107.69	<0.001
Dyslipidemia, n (%)	634 (36.8%)	125 (23.3%)	33.369	<0.001
Hyperuricemia, n (%)	529 (30.7%)	79 (14.7%)	53.059	<0.001
Abnormal glucose metabolism, n (%)	613 (35.5%)	81 (15.1%)	80.584	<0.001

Table 6. Comparison of categorical variables between the case group and the control group

Variable	Case group (*n*=517)	Control group (*n*=536)	x^2-value	*P*-value
Obesity, n (%)	260 (50.2%)	46 (8.5%)	222.066	<0.001
Sex, n (%)	Male 370 (71.5%)	Male 291 (54.2%)	33.61	<0.001
Drinking history, n (%)	240 (46.4%)	47 (8.7%)	188.18	<0.001
Smoking history, n (%)	277 (53.5%)	67 (12.5%)	201.885	<0.001
Hypertension, n (%)	279 (53.9%)	149 (27.7%)	74.689	<0.001
Dyslipidemia, n (%)	193 (37.3%)	125 (23.3%)	24.504	<0.001
Hyperuricemia, n (%)	163 (31.5%)	79 (14.7%)	41.909	<0.001
Abnormal glucose metabolism, n (%)	178 (34.4%)	81 (15.1%)	52.950	<0.001

Compared the results of pre- and post-sampling between groups, which demonstrated that the results of pre- and post-sampling were consistent with the OSA group when the case group was on behalf of the OSA group in the data analysis. And it proved that after the undersampling treatment, the sampling result of the case group still responded well to the total sample size. Hence, the case group could be selected instead of the OSA group for subsequent data analysis.

3.6. Construction of the classification tree model

The case group and control group were used to be study subjects, including 517 in the case group and 536 in the control group. OSA was used as the dependent variable; those who suffered from OSA were assigned a value of 1, and not suffer from OSA were assigned a value of 0. Obesity (obesity = 1, no obesity = 0),

gender (male = 1, female = 0), age, drinking history (drinking history = 1, no history of drinking = 0), smoking history (smoking history = 1, no history of smoking= 0), hypertension (hypertension = 1, no hypertension = 0), dyslipidemia (dyslipidemia = 1, no dyslipidemia = 0), hyperuricemia (hyperuricemia = 1, no hyperuricemia = 0), abnormal glucose metabolism (abnormal glucose metabolism = 1, no abnormal glucose metabolism = 0) were used as independent variables, and obesity, gender, drinking history, smoking history, hypertension, dyslipidemia, hyperuricemia and abnormal glucose metabolism were defined as classified variables, and age was defined as continuous variable. By pre-setting the growth depth and pruning rules of the tree, the results of the classification tree model of risk factors related to obstructive sleep apnea comprised 5 layers and 23 nodes, including 12 terminal nodes. Finally, 7 explanatory variables which affected OSA were selected, including obesity, smoking history, age, drinking history, hypertension, abnormal glucose metabolism, and gender (**Figure 1**).

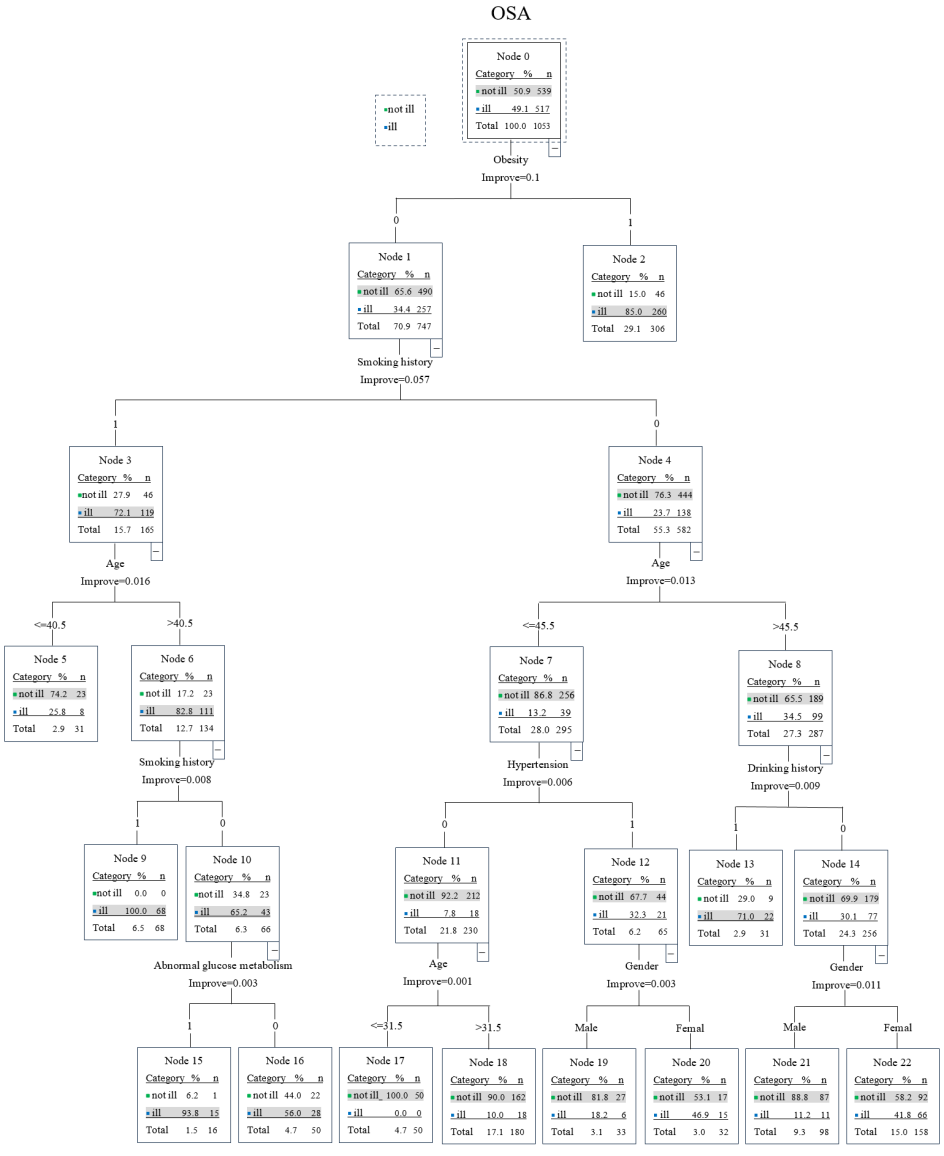

Figure 1. Classification tree model of factors influencing obstructive sleep apnea

3.7. The result of the classification tree model

The first level of the classification tree was divided into intervals according to obesity and no obesity; thus, obesity was the primary risk factor for OSA. And the prevalence rate of OSA in obese people (85%, 260/306) was significantly higher than no obese people (34.4%, 257/747). The second level of the model was split by smoking history and no history of smoking, and in no obese people, the prevalence rate of OSA in those who had a history of smoking (72.1%, 119/165) was significantly higher than those who had no history of smoking (23.7%, 138/582). OSA-related risk factor screened for the third level of the model was age. In non-obese obesity with a history of smoking, the prevalence rate of OSA in those aged >40.5 years (82.8%, 111/134) was significantly higher than those aged ≤40.5 years (25.8%, 8/31). But in those with no obesity and no history of smoking, the prevalence rate of OSA in those aged >45.5 years (34.5%, 99/287) was significantly higher than aged ≤45.5 years (13.2%,39/295). OSA-related risk factors screened for the fourth level of the model were drinking history and hypertension. Among those without obesity, had smoking history, and those aged >40.5 years, the prevalence rate of OSA in had drinking history (100%, 68/68) was significantly higher than those without drinking history (65.2%, 43/66). Among those without obesity, no smoking history, and aged ≤45.5 years, the prevalence rate of OSA in those with hypertension (32.3%, 21/65)was significantly higher than those without hypertension (7.8%, 18/230), whereas among those without obesity, without history of smoking, and age >45.5 years, the prevalence rate of OSA in those with a history of alcohol consumption (71%, 22/31) were significantly higher than those without a history of drinking (30.1%, 77/256). OSA-related risk factors screened for the fifth level of the model were abnormal glucose metabolism, age, and gender. In those without smoking history, aged >40.5 years, and without drinking history, the prevalence rate of OSA in abnormal glucose metabolism (93.8%, 15/16) was significantly higher than those who did not have abnormal glucose metabolism (56%, 28/50). In those no obesity, without smoking history, aged ≤45.5 years, and without hypertension, the prevalence rate of OSA in those aged >31.5 years (10%, 18/180) was significantly higher than those aged ≤31.5 years (0%, 0/50). In people without obesity, without smoking history, aged ≤45.5 years, and with hypertension, the prevalence rate of OSA in females (46.9%, 15/32) was significantly higher than in males (18.2%, 6/33). But in people without obesity, without smoking history, aged >45.5 years, and without drinking history, the prevalence rate of OSA in females (41.8%, 66/158) was significantly higher than in males (11.2%, 11/98).

3.8. Evaluation of the classification tree model

By conducting 10-fold cross-validation for the classification tree model, the authors obtained that the re-substitution estimators and cross-validation estimators of the model were 0.192 and 0.200, respectively, and the standard errors were both 0.012. The results indicated that the correct rate was 80.8% by using a classification tree model to predict OSA influencing factors and proved that the model fitting results were well. The ROC curve was plotted by using the multi-factor combined predictive probabilities from the model: the Youden index of the ROC curve of classification tree model was 0.614, sensitivity was 76.0%, specificity was 85.4%, the area under the curve (AUC) value was 0.880 (95% CI:0.860~0.901), and standard error was 0.010, ($P < 0.001$); it indicated that model had high accuracy and could effectively select OSA-related risk factors (**Figure 2**).

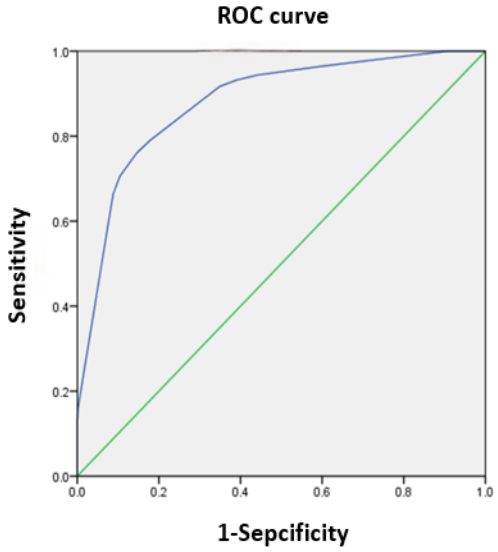

Figure 2. ROC curve of OSA predicted by the classification tree model

4. Discussions

OSA is a globally prevalent and not adequately diagnosed public health disease, which faces a huge and increasing demand for diagnosis in clinical practice. People need a reliable primary screening method to identify people at high risk of OSA. Meanwhile, the complexity and diversity of clinical symptoms and metabolic features of OSA patients are creating an urgent need for us to explore the intrinsic links between various types of clinical symptoms and metabolic profiles. At present, logistic regression or multifactorial linear regression models are usually used to screen the risk factors of OSA in China. Although the two types of models above are sufficient to analyze the main effects of the independent variables, they are difficult to use to analyze the hierarchical relationships among the independent variables, which reduces the analytical efficacy. The classification tree model, as a non-parametric model, can be used to divide the total research population into several relatively homogeneous subintervals according to the features by automatically generating the importance of features, and to display them in a tree diagram, which makes the output results clearer and more direct. Currently, the classification tree model is mainly used in market research studies, and some studies have also applied it to screen the risk factors of diseases. But no study has reported applying the model to screen for the risk factors of OSA. Therefore, the aim of this study is to screen the specific population affected by each variable through the classification tree model, thereby adopting individualized preventive and treatment measures for this kind of high-risk group of OSA, in order to prevent and treat OSA and the development of complications by early intervention and treatment of the disease.

The relationship between OSA and demographic characteristics: the result of this study demonstrated that obesity was detected as the primary explanatory variable for OSA in the classification tree model, which was consistent with the results of previous studies, confirmed by epidemiological data that obesity was the most important independent risk factor of OSA [8–9]. Meanwhile, the result of the classification tree model indicated that it was not possible to re-divide the obese population by using an influencing factor as a split point to generate new sub-nodes in the obese population, which suggested that under the influence of obesity as the primary risk factor, the strong correlation of pathogenic factors had already been shown. And the correlation of pathogenic of other

influencing factors of OSA was hidden, which further reminded the importance of obesity in the development of OSA. This study indicated that, compared with the control group, the age and prevalence rate of males in the OSA group and case group were increased, with statistical significance, which was consistent with the result confirmed by traditional epidemiological data. The classification tree model could also screen out age and gender as explanatory variables, but gender as an explanatory variable showed inconsistent results with the univariate analysis. The prevalence rate of OSA was significantly higher in women than in men for a given condition when screened by layer-by-layer subdividing in the population without obesity, which suggested that although the overall prevalence rate of OSA in males was higher than in females, males were more likely to develop the disease under the condition of obesity. The independence correlation of pathogenic obesity, as a primary risk factor of OSA, decreased the interaction between itself and other variables, including male, which could be verified in the first level of the model. In the meantime, the result of the third level of the model showed that, in those without obesity and with a history of smoking, the prevalence rate of OSA in those aged >40.5 was increased significantly, but in those without obesity and a history of smoking, the prevalence rate of OSA in those aged >45.5 was increased significantly. The result showed that as age increased, the risk of the prevalence rate of OSA also increased, whether or not there was a history of smoking. The result of the fifth level of model showed that, in those without obesity, without history of smoking, aged ≤45.5 and without hypertension, the prevalence rate of OSA in aged >31.5 was significantly higher than aged ≤31.5, which displayed that patients were likely onset by not shown in model, slightly lower degree explanation of hyperuricemia, dyslipidemia and other relative risk factors of OSA, and the age of onset was showing a younger-age trend. But in without obesity, without smoking history, aged >45.5 years old and without drinking history people, the prevalence rate of OSA in female was significantly higher than male, which suggested that female above menopause age were more likely to suffer from OSA than male after excluding influencing factors such as obesity and unhealthy lifestyle behaviors, and it indicated that OSA should not be limited to the traditionally recognized disease in male, we should pay more attention to middle-aged and elder women above the age of perimenopausal with OSA related symptoms.

The relationship between OSA and lifestyle behaviors: the result of this study demonstrated that, compared with the control group, the rate of smoking and drinking was increased in the OSA group and the drinking group, with statistical significance. The second level of the classification tree model was split by smoking history, and without smoking history, in the non-obesity population, the prevalence rate of OSA in those with a smoking history was significantly higher than those without smoking history. The third level of classification tree model was split by age, the result showed that the age as split point in without obesity and with smoking history people was 40.5 years, and in without obesity and without smoking history people was 45.5 years, which suggested that smoking, an unhealthy lifestyle behavior, not only contributed to increase significantly for the prevalence rate of OSA, but also led to a younger-age trend of disease. The result of the model in the fourth level showed that, without obesity, with smoking history, and aged >40.5 years, the prevalence rate of OSA in those with a drinking history was significantly higher than those without a drinking history. And in people without obesity, without smoking history, and aged >45.5 years, the prevalence rate of OSA in those with a drinking history was also significantly higher than those without a drinking history. The above two results indicated that a history of drinking contributed significantly to the increase in the prevalence rate of OSA, and even under the condition of without smoking history, the elderly OSA patients who had a drinking history still appeared to have a higher morbidity, which suggested that with the prolongation of drinking history, the risk of OSA was increased.

Relationship between OSA and hypertension: this study indicated that, compared with the control group, the

rate of hypertension in the OSA group and case group was elevated, with statistical significance, which is close to the result of epidemiological data [10]. The result of the model in the fourth level showed that, without obesity, without smoking history, and aged ≤45.5 years, the prevalence rate of OSA in hypertension was significantly higher than those without hypertension. And the next level sub-nodes of this node were split by gender, and under this background, the female population was more likely to develop OSA than the male population, which suggested that hypertension, a cardiovascular disease, might medicate the pathogenetic process of OSA in a large number of pre-menopausal females who hadn't unhealthy life style and behaviors such as obesity and smoking previous.

Relationship between OSA and metabolic syndrome: this study indicated that, compared with the control group, the rate of dyslipidemia, hyperuricemia, and abnormal glucose metabolism in the OSA group and case group was elevated, with statistical significance. However, the classification tree model only screens out the abnormal glucose metabolism as an influencing factor of OSA. Although neither type of model were not screened out dyslipidemia and hyperuricemia as risk factors of OSA, the result of non-parametric tests showed that the control group had a better level of control of TG, CHOL, LDLc, HDLc, and UA, when contrasted respectively with the OSA group and case group patients, with statistical significance. The appearance of this phenomenon might be related to the confounding effect of obesity and OSA, and obesity has been confirmed to be strongly associated with the metabolic syndrome by a vast amount of epidemiological data [11]. Metabolic syndrome is a group of clinical syndromes including hyperglycemia, dyslipidemia, hyperuricemia, etc., and with the obesity rate increasing in the world population, the incidence of metabolic syndrome also increases yearly [12]. Meanwhile, obesity was the primary risk factor of OSA; thus, the correlation of dyslipidemia and hyperuricemia for the occurrence of OSA might be masked by obesity. This statement was confirmed in the second-level classification model tree, which used obesity as a split point and failed to delineate the lower child nodes. Abnormal glucose metabolism was detected as a risk factor in two types of models, which suggested that the effect of abnormal glucose metabolism on OSA was likely the most important among the metabolic traits. Meanwhile, the results of the classification tree model were classified in the context of the population without obesity, which reflected that the pathogenic relevance of abnormal glucose metabolism for OSA was independent of the confounding effect of obesity and metabolic syndrome.

5. Conclusion

In conclusion, the main influencing factors of OSA were obesity, smoking history, age, drinking history, hypertension, abnormal glucose metabolism, and gender. Even though men were an independent risk factor of OSA, in the context of without obesity and without smoking history, we should pay more attention to pre-menopausal women with hypertension with OSA-related symptoms and middle-aged and elderly women above the age of perimenopause without a history of alcohol consumption. Abnormal glucose metabolism may be the most important among metabolic diseases associated with OSA, and this association is independent of the confounding effects of obesity and metabolic syndrome.

There are certain limitations and flaws to this study: this retrospective research was conducted in a single medical centre, which introduced unavoidable time bias and selection bias. The gender composition and age composition of the study subjects were more focused, which resulted in a certain bias for consequences. The study can be improved by using a more accurate sampling method, which would further reduce the impact of sampling

error on results and improve the accuracy of the model. At the same time, by enlarging the sample size to conduct the propensity score matching design, more accurate research results can be obtained.

Funding

This work was supported by grants from the Guangxi Natural Science Foundation (Grant No. 2022JJA140014).

Disclosure statement

The authors declare no conflict of interest.

References

[1] Abbasi A, Gupta SS, Sabharwal N, et al., 2021, A Comprehensive Review of Obstructive Sleep Apnea. Sleep Science, 2021(14): 142–154.

[2] Benjafield AV, Ayas NT, Eastwood PR, et al., 2019, Estimation of the Global Prevalence and Burden of Obstructive Sleep Apnoea: A Literature-Based Analysis. Lancet Respiratory Medicine, 2019(7): 687–698.

[3] Zhang J, Zhao D, Zhou Z X et al., 2021, Value of Night Pulse Oximetry Monitoring in Obstructive Sleep Apnea Hypopnea Syndrome Prediction and Classification. Zhonghua Jie He He Hu Xi Za Zhi, 2021(44): 101–107.

[4] Cheng P, Yu W, Chu CL, 2018, Decision Tree Algorithm-Based Medical Big Data. Information technology and informatization, 2018(9): 70–74.

[5] Li YX, Chai Y, Hu YQ, et al., 2019, Review of Imbalanced Data Classification Methods. Control and Decision, 34(4): 673–688.

[6] Liu SK, He XQ, Xia LY, 2022, Discussion of the Validity of Model Evaluation Indicators under Unbalanced Data. Control and Decision, 38(19): 5–9.

[7] Kapur VK, Auckley DH, Chowdhuri S, et al., 2017, Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline. Journal of Clinical Sleep Medicine, 13(3): 479–504.

[8] Bonsignore MR, 2022, Obesity and Obstructive Sleep Apnea. Handbook of Experimental Pharmacology, 2022(274): 181–201.

[9] Locke BW, Lee JJ, Sundar KM, 2022, OSA and Chronic Respiratory Disease: Mechanisms and Epidemiology. International Journal of Environmental Research and Public Health, 2022(19): 5473.

[10] Hou HF, Zhao YG, Yu WQ et al., 2018, Association of Obstructive Sleep Apnea with Hypertension: A Systematic Review and Meta-analysis. Journal of Global Health, 2018(8): 010405.

[11] Payab M, Tayanloo-Beik A, Falahzadeh K, et al., 2022, Metabolomics Prospect of Obesity and Metabolic Syndrome; A Systematic Review. Journal of Diabetes and Metabolic Disorders, 2022(21): 889–917.

[12] Rana S, Ali S, Wani HA, et al., 2022, Metabolic Syndrome and Underlying Genetic Determinants: A Systematic Review. Journal of Diabetes and Metabolic Disorders, 2022(21): 1095–1104.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.