PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Moderate-severe OSA screening based on support vector machine of the Chinese population facio-cervical measurements dataset: a cross-sectional study
AUTHORS	zhang, liu; Yan, Ya Ru; Li, Shi Qi; Li, Hong Peng; Lin, Ying Ni; Li, Ning; Sun, Xian Wen; Ding, Yong Jie; Li, Chuan Xiang; Li, Qing Yun

VERSION 1 – REVIEW

REVIEWER	Palm, Andreas Uppsala Universitet, Department of Medical Sciences, Respiratory, Allergy and Sleep Research
REVIEW RETURNED	29-Jan-2021

GENERAL COMMENTS	This was a single center cross-sectional observational study of 481 Chinese patients with suspected OSA. The study aimed to analyze how accurate a machine-learning model could predict the presence of moderate to severe obstructive sleep apnea. The authors conclude that the machine-learning model is accurate in predicting moderate to severe OSA, especially in patients without significant daytime symptoms. I have seven issues that require clarification from the authors, along with some minor suggestions for modifications. Major issues: 1. In Methods section page 6 line 3 it is mentioned that certain patients are excluded from the study if they had severe cognitive impairment, severe heart failure, severe respiratory failure or other serious acute or chronic diseases and neuromuscular diseases. How did you define the severity of these conditions? Did the patients have any comorbid conditions? 2. Description of statistical methods should be concentrated to a “Statistics paragraph”. 3. In Table 1, variables with P<0.05 found in chi-2 and t-tests are then used in the machine-learning model. p-values crude OR and OR after adjustment for BMI are presented. Why did you adjust for BMI? What does the crude logistic regression add? In page 9, line 29 it is stated that MID and H/TSD are independent risk-factors for OSA. Results found in cross-sectional studies says us nothing about causality and therefore it is more accurate to write about associations instead of risk-factors. In addition, there is an association to” moderate/severe OSA”, not “OSA”. 4. Why did the authors use the training set as reference group and not the AHI from the PSGs? With this design, there would be very strange if the results from the STOP-BANG

	questionnaires would be closer to the training set than the validation set. In the methods section the terms “training set” and “validation set” are used. In the results section, the “validation set” in suddenly called “testing set”. Very confusing. 5. The description of the model construction is confusing. Is the output of the model binary? What are the results from the machine-learning model? I miss a step between table 1 and table 2. Cross tables with outcome of PSG, the machine learning model, and the STOP-BANG results would be clarifying. 6. Flow chart, Figure 2. According to the methods section, p-values from t-tests and chi-2 tests were used to chose variables to include into the machine-learning model, not logistic regression. Only 95% Cis, not p-values are presented for the logistic regression in table 1 (good so). 7. I wish that the authors suggest how this machine-learning model can be used in a clinical context. How can a clinician interpret the results? Minor issues: 1. This manuscript should improve with an English language review. 2. In the abstract, the result section is a bit confusing. Sensitivity and specificity are graded first with percentage and then with decimal numeral. There is no consistency in the way the different modalities are compared. The same problem exists in the results section. 3. Page 6, line 55. Should say “other caffeine-containing products, instead of “Cola-containing products” . 4. Page 6, line 59. Should say “thoracal”, instead of ribcage. 5. Page 7, line 5. The sentence “Sleep data were scored manually by registered polysomnographer of the United States according to American Academy of Sleep Medicine Recommendations (AASM)” is confusing. 6. The STOP-BANG questionnaire is a central part of the narrative of this study. It would be polite to the reader with a brief overview of this scale in the methods-section. 7. Table 1 should contain a row with characteristics of all patients. In row 4, it should say “p-value”, instead of “P”. Some values in the Mallampti test-rows are missing. Should say “N/A: not applicable”, instead of “NA: not available” 8. AHI-groups are categorized into “Deal” and “No-deal”. It would be more informative to just name the categories “Mild AHI” and “Moderate-severe OSA”. 9. In the methods section the terms “training set” and “validation set” are used. In the results section, the “validation set” in suddenly called “testing set”. Very confusing. Please also describe the rationale for dividing the patients in these groups.

REVIEWER	Vena, Daniel Brigham and Women's Hospital, Sleep and Circadian Disorders
REVIEW RETURNED	01-Mar-2021

GENERAL COMMENTS	The current paper analyzes the relationship between facio-cervical measurements and the presence of OSA (AHI > 15 /hr). They then used the significant predictors and used them to develop a model to predict the presence of OSA. Significant predictors were: height to thyro-sternum distance, maximum interincisal distance, as well as male sex, age, neck circumference, waist circumference, and BMI. The SVM model which included these variables predicted

	OSA with 88% accuracy (sens = 87% and spec = 70%) and performed substantially better than the STOP-BANG, especially in asymptomatic patients. The methodology and the results are very good. Missing from the paper are good physiological explanations for the variables selected and how they related to OSA. Limiting clinical utility of the results is the lack of availability of the model to clinicians, unlike the STOP-BANG which is a simple questionnaire. Additional comments are made in the appropriate sections below. Introduction: Can you provide detail on the scale of the problem of OSA patients in the context of surgery? What is the role of OSA severity in perioperative outcomes? For example, do moderate OSA patients have similar outcomes to severe? Can you clarify the significance of Asians having a small upper-airway compared to Caucasian and its link with facio-cervical characteristics? Why is this not applicable to other races/airway sizes? Methods: Why did you choose AHI > 15 /hr as the cutoff. Is this the optimal outcome for perioperative outcomes? Why do you call apnea and non-apnea patients deal or no-deal? Not sure this is common nomenclature. Maybe OSA+ versus OSA-? Please specify how many principal components you selected and the criteria for selecting them. Overall, I'm not clear on the role of principal components in the analysis. Why normalize TMD and TSD to height? I understand that TMD and TSD will be greater due to height, but does this matter? If these measures relate to OSA severity, does it matter that these measures are naturally elevated in taller patients? I do not follow the role of the logistic regression in the flow chart in Figure 2. I understand you used it to understand the bivariate relationship between predictor variables and presence of OSA, but how did it feed into the SVM model, as illustrated in Figure 2? Results: Regarding Figure 3, what is the significance of the small "island" of non-OSA patients. Is this the data from all patients? Testing? Training? What are the axes? Please report the results of the principal component analysis How much better is the model compared to if you just used Age, Sex, NC, WC, BMI? In table 1, what are the odds rations relative to? 2SD change in predictor variable? Please clarify in the notes below the table. Somewhere in the results, clarify the direction of the odd ratios for the significant variables. I.e. MID odds ratio of 1.3 means 1.3x greater odds of having OSA per X increase in MID (is that right?). While odds ratio of 0.3 for H/TSD means 3 times lower odds of having OSA per X increase in TSD for a given height. Discussion Can you provide a physiological explanation for why MID, TMD and TSD are associated with OSA and difficult intubation? Another limitation of the study is that it is more challenging to use your model in the clinical environment compared to the STOP-BANG. In this way, the method has limited clinical utility in its current state. Is it possible to develop such a questionnaire that would assess OSA risk? This would provide excellent clinical applicability of your work. Further, it would allow other groups to test the questionnaire on their populations.

VERSION 1 – AUTHOR RESPONSE

Dear Reviewer 1,

Thank you very much for your time involved in reviewing the manuscript. Your comments were highly insightful and enabled us to greatly improve the quality of our manuscript.

Comments:

This was a single center cross-sectional observational study of 481 Chinese patients with suspected OSA. The study aimed to analyze how accurate a machine-learning model could predict the presence of moderate to severe obstructive sleep apnea. The authors conclude that the machine-learning model is accurate in predicting moderate to severe OSA, especially in patients without significant daytime symptoms.

I have seven issues that require clarification from the authors, along with some minor suggestions for modifications.

Major issues:

1. In Methods section page 6 line 3 it is mentioned that certain patients are excluded from the study if they had severe cognitive impairment, severe heart failure, severe respiratory failure or other serious acute or chronic diseases and neuromuscular diseases. How did you define the severity of these conditions? Did the patients have any comorbid conditions?

Response: Thank you for your reminding.

(1) The detailed corrections are listed below, and the corresponding revision is on Page 5, Line 9. Exclusion criteria: (1) Patients showing complications with severe respiratory diseases, such as severe COPD, interstitial lung disease or acute asthma; (2) Patients showing complications with serious cardiovascular diseases such as acute myocardial infarction, acute heart failure or chronic congestive heart failure (Grade III and IV); (3) Patients with mental illnesses who could not cooperate with the examination; (4) Patients who receiving non-invasive positive pressure ventilation therapy; (5) Patients who might had other sleep disorders under clinical evaluation.

(2) Some patients have comorbid conditions, such as hypertension.

2. Description of statistical methods should be concentrated to a “Statistics paragraph”.

Response: Thank you for this advice. The description of statistical methods has been concentrated to the “Statistics paragraph” (Page 6, Line 17).

3. In Table 1, variables with $P<0.05$ found in chi-2 and t-tests are then used in the machine-learning model. p-values crude OR and OR after adjustment for BMI are presented. Why did you adjust for BMI? What does the crude logistic regression add? In page 9, line 29 it is stated that MID and H/TSD

are independent risk-factors for OSA. Results found in cross-sectional studies says us nothing about causality and therefore it is more accurate to write about associations instead of risk-factors. In addition, there is an association to" moderate/severe OSA", not "OSA".

Response: We feel great thanks for your advice.

(1) The crude OR and OR adjusted for BMI were used to understand the bivariate relationship between variables and presence of OSA. The OR adjusted for BMI aims to avoid the confounding effect of BMI.
(2) We have corrected "risk factors" to "associations", and "OSA" to "moderate or severe OSA".

4. Why did the authors use the training set as reference group and not the AHI from the PSGs? With this design, there would be very strange if the results from the STOP- BANG questionnaires would be closer to the training set than the validation set. In the methods section the terms "training set" and "validation set" are used. In the results section, the "validation set" in suddenly called "testing set". Very confusing.

Response: Thank you for your nice comments.

(1) The SABIHC2 model was set up on the training dataset and validated on the testing dataset. In both datasets, the moderate or severe OSA were reference group.
(2) We conclude that the model based on training dataset has best predictive ability, closely followed by the model based on testing dataset, while STOP- BANG questionnaire comes next. So, the results from the STOP- BANG questionnaires would be closer to testing dataset than training dataset.
(3) All "validation dataset" were corrected to "testing dataset" to ensure uniformity.

5. The description of the model construction is confusing. Is the output of the model binary? What are the results from the machine-learning model? I miss a step between table 1 and table 2. Cross tables with outcome of PSG, the machine learning model, and the STOP-BANG results would be clarifying.

Response: The model produces a binary output. The cross tables were showed as follow.

		PSG		
		AHI<15	AHI≥15	Total
SABIHC2	0	123	23	146
	1	48	287	335
Total		171	310	481

	PSG		
	AHI<15	AHI≥15	Total
STOP-Bang≤4	132	159	291
STOP-Bang>4	39	151	190

Total	171	310	481

6. Flow chart, Figure 2. According to the methods section, p-values from t-tests and chi- 2 tests were used to chose variables to include into the machine-learning model, not logistic regression. Only 95% Cis, not p-values are presented for the logistic regression in table 1 (good so).

Response: Thank you for this advice.

(1) We had corrected “logistic regression” to “t-test or chi-square test” in Figure 2 and updated the figure.
(2) We add the p-values of logistic regression in Table 1.

7. I wish that the authors suggest how this machine-learning model can be used in a clinical context. How can a clinician interpret the results?

Response: Many thanks for this comment.

(1) As most models work, we are planning to develop a software or application in the future, to allow health care worker friendly installation and application. And we have added the sentence in the main text (page 12, line 3).
(2) The model produces a binary output with a sensitivity of 0.874 and specificity of 0.700. If the output is 1, the patient is more likely to be a moderate or severe OSA with an accuracy rate of 0.928, which may require further examination, such as PSG.

Minor issues:

1. This manuscript should improve with an English language review.

Response: Thanks for your suggestions. We have carefully edited the entire manuscript and the manuscript has been polished by a professional editor before resubmission. We hope the revised manuscript would satisfy you.

2. In the abstract, the result section is a bit confusing. Sensitivity and specificity are graded first with percentage and then with decimal numeral. There is no consistency in the way the different modalities are compared. The same problem exists in the results section.

Response: Thanks for your correction.

All sensitivity and specificity were change to decimal numeral to ensure uniformity in the full text.

3. Page 6, line 55. Should say “other caffeine-containing products, instead of “Cola- containing products”.

Response: We have corrected the “Cola- containing products” into “other caffeine-containing products”.

4. Page 6, line 59. Should say “thoracal”, instead of ribcage.

Response: We have corrected the “ribcage” into “thoracal”.

5. Page 7, line 5. The sentence “Sleep data were scored manually by registered polysomnographer of the United States according to American Academy of Sleep Medicine Recommendations (AASM)” is confusing.

Response: The sentence was corrected to “the PSG data were scored according to the American Academy of Sleep Medicine (AASM) criteria”.

6. The STOP-BANG questionnaire is a central part of the narrative of this study. It would be polite to the reader with a brief overview of this scale in the methods-section.

Response: Thanks for your suggestions. We have added the sentences to the methods-section (page 5, line 23).

The STOP-BANG questionnaire is a scoring model consisting of eight questions and its scores are based on Yes/No answers (score: 1/0). The eight questions included snoring, tiredness, observed apnea and high blood pressure, BMI, age, neck circumference and gender.

7. Table 1 should contain a row with characteristics of all patients. In row 4, it should say “p-value”, instead of “P”. Some values in the Mallampti test-rows are missing. Should say “N/A: not applicable”, instead of “NA: not available”

Response: According to your suggestion, we have added a row of all patients.

(1) The “P” and “NA” were corrected to “p-value” and “N/A”.
(2) We added “N/A” to the Mallampti test-rows.

8. AHI-groups are categorized into “Deal” and “No-deal”. It would be more informative to just name the categories “Mild AHI” and “Moderate-severe OSA”.

Response: Your suggestion really means a lot to us. We changed the deal group and no-deal group to moderate or severe OSA and non or mild OSA in the full text, respectively.

9. In the methods section the terms “training set” and “validation set” are used. In the results section, the “validation set” in suddenly called “testing set”. Very confusing. Please also describe the rationale for dividing the patients in these groups.

Response: According to your suggestion, we have corrected the “validation dataset” into “testing dataset” to ensure uniformity.

Dear Reviewer2,

Thank you very much for your time involved in reviewing the manuscript and your very encouraging comments on the merits. Your comments were highly insightful and enabled us to greatly improve the quality of our manuscript.

Comments:

The current paper analyzes the relationship between facio-cervical measurements and the presence of OSA (AHI > 15 /hr). They then used the significant predictors and used them to develop a model to predict the presence of OSA. Significant predictors were: height to thyro-sternum distance, maximum interincisal distance, as well as male sex, age, neck circumference, waist circumference, and BMI. The SVM model which included these variables predicted OSA with 88% accuracy (sens = 87% and spec = 70%) and performed substantially better than the STOP-BANG, especially in asymptomatic patients. The methodology and the results are very good. Missing from the paper are good physiological explanations for the variables selected and how they related to OSA. Limiting clinical utility of the results is the lack of availability of the model to clinicians, unlike the STOP-BANG which is a simple questionnaire. Additional comments are made in the appropriate sections below.

Introduction:
Can you provide detail on the scale of the problem of OSA patients in the context of surgery?
What is the role of OSA severity in perioperative outcomes? For example, do moderate OSA patients have similar outcomes to severe?

Response: We added the sentence to the manuscript in Introduction section (page 4, line 8).

The rates of postoperative cardiovascular events show a rise in moderate or severe OSA (25.1%) compared to no or mild OSA (16.8%).[1]

Reference:

1. Chan MTV, Wang CY, Seet E, et al. Postoperative Vascular Complications in Unrecognized Obstructive Sleep Apnea (POSA) Study Investigators. Association of Unrecognized Obstructive Sleep Apnea With Postoperative Cardiovascular Events in Patients Undergoing Major Noncardiac Surgery. JAMA. 2019 May 14;321(18):1788-1798.

Can you clarify the significance of Asians having a small upper-airway compared to Caucasian and its link with facio-cervical characteristics? Why is this not applicable to other races/airway sizes?

Response: Many thanks for this comment. We added the sentence to the manuscript in Introduction section.

Only in Asians, smaller upper airways are predictors of upper airway collapsibility, and an anatomic imbalance between tongue and mandible volume influenced upper airway collapsibility among Caucasians. [1]

The above evidence prompts facio-cervical characteristics may predict OSA for Asians. Whether it is applicable to other races needs further studies.

Reference:

1. Schorr F, Kayamori F, Hirata RP, et al. Different Craniofacial Characteristics Predict Upper Airway Collapsibility in Japanese-Brazilian and White Men. Chest. 2016; 149: 737-46.

Methods:
Why did you choose AHI > 15 /hr as the cutoff. Is this the optimal outcome for perioperative outcomes?

Response: Thanks for your reminding.

An AHI≥15 /hr is associated with complications,[1] which is always used as inclusion criteria in most studies on OSA.[2,3] Further, moderate-severe is considered as the index of perioperative safety, [4,5] and AHI≥15 means moderate-severe ones.

Reference:

1. Veasey SC, Rosen IM. Obstructive Sleep Apnea in Adults. N Engl J Med. 2019 Apr 11;380(15):1442-1449.

2. Lee CH, Sethi R, Li R, et al. Obstructive Sleep Apnea and Cardiovascular Events After Percutaneous Coronary Intervention. Circulation. 2016 May 24;133(21):2008-17.

3. Pépin JL, Letesson C, Le-Dong NN, et al. Assessment of Mandibular Movement Monitoring With Machine Learning Analysis for the Diagnosis of Obstructive Sleep Apnea. JAMA Netw Open. 2020 Jan 3;3(1):e1919657.

4. American Society of Anesthesiologists Task Force on Perioperative Management of patients with obstructive sleep apnea. Practice guidelines for the perioperative management of patients with obstructive sleep apnea: an updated report by the American Society of Anesthesiologists Task Force on Perioperative Management of patients with obstructive sleep apnea. Anesthesiology. 2014 Feb;120(2):268-86.

5. Chan MTV, Wang CY, Seet E, et al. Postoperative Vascular Complications in Unrecognized Obstructive Sleep Apnea (POSA) Study Investigators. Association of Unrecognized Obstructive Sleep Apnea With Postoperative Cardiovascular Events in Patients Undergoing Major Noncardiac Surgery. JAMA. 2019 May 14;321(18):1788-1798.

Why do you call apnea and non-apnea patients deal or no-deal? Not sure this is common nomenclature. Maybe OSA+ versus OSA-?

Response: Your suggestion really means a lot to us. We changed the deal group and no-deal group to non or mild OSA and moderate or severe OSA in the full text, respectively.

Please specify how many principal components you selected and the criteria for selecting them. Overall, I'm not clear on the role of principal components in the analysis.

Response: We feel great thanks for your professional review work on our article. We added the sentences to the manuscript.

We performed a significant principal component taking in to account the strong collinearity among parameters, such as BMI and WC. Five principal components were selected according to the accumulative variance contribution more than 90% and scree plot.[1] The results of the principal component analysis were showed in the second question of Results of yours.

Reference:

1.Björklund M. Be careful with your principal components. Evolution. 2019 Oct;73(10):2151-2158.

Why normalize TMD and TSD to height? I understand that TMD and TSD will be greater due to height, but does this matter? If these measures relate to OSA severity, does it matter that these measures are naturally elevated in taller patients?

Response: Thank you for your reminding.

As you stated, TMD and TSD will be greater due to height. Schmitt et al. reported that the ratio of height to thyromental distance was a more sensitive indictor of difficult intubation than the thyromental distance alone.[1] Results obtained based on our dataset were consistent, as showed in the table below. So, we choose the ratio to analysis.

	All patients (n=481)	AHI<15 (n=171)	AHI≥15 (n=310)	p-value*	crude OR (95% CI) #
TMD	9.13±1.27	9.11±1.11	9.13±1.36	0.878	1.011(0.877-1.164)
H/TMD	18.98±2.75	18.69±2.30	19.14±2.95	0.085	1.064(0.991-1.142)
TSD	9.16±1.42	9.74±1.38	8.84±1.34	<0.001	0.610(0.525-0.709)
H/TSD	18.92±2.97	17.49±2.06	19.70±3.10	<0.001	1.448(1.311-1.599)

*: t test or chi-square test as appropriate. #: Odds ratios are depicted for moderate or severe OSA relative to no or mild OSA. N/A: not applicable.

Reference:

1. Schmitt HJ, Kirmse M, Radespiel-Troger M. Ratio of patient's height to thyromental distance improves prediction of difficult laryngoscopy. Anaesth Intensive Care 2002;30:763–5.

I do not follow the role of the logistic regression in the flow chart in Figure 2. I understand you used it to understand the bivariate relationship between predictor variables and presence of OSA, but how did it feed into the SVM model, as illustrated in Figure 2?

Response: Thanks for your careful checks.

About the logistic regression in figure2, it is a mistake in statistical method. We had corrected logistic regression to t-test or chi-square test in Figure 2 and uploaded the figure. We thank the reviewer for changing this error.

Results:
Regarding Figure 3, what is the significance of the small “island” of non-OSA patients. Is this the data from all patients? Testing? Training? What are the axes?

Response: Thanks for you reminding.

(1) Figure3 showed classification results of all patients based on SVM. The optimal hyperplane was obtained from the SVM classifier using Python. The small “island” may due to that the algorithm regard this part as green area, because of three green dots around and without blue dot. The small island is not meaningful, due to the limited dataset.

(2) The abscissa and the ordinate are the first principal component and the second principal component respectively. We have added the titles of axes in Figure3.

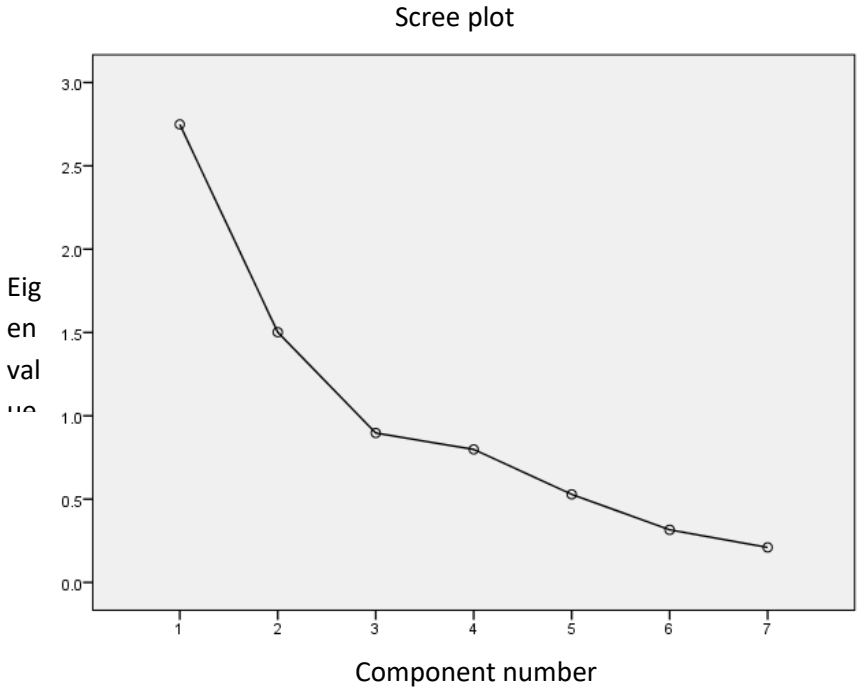
Please report the results of the principal component analysis

Response: The results of the principal component analysis were showed as follows. Five principal components were selected according to the accumulative variance contribution more than 90% and scree plot.[1]

Reference:

1.Björklund M. Be careful with your principal components. Evolution. 2019 Oct;73(10):2151-2158.

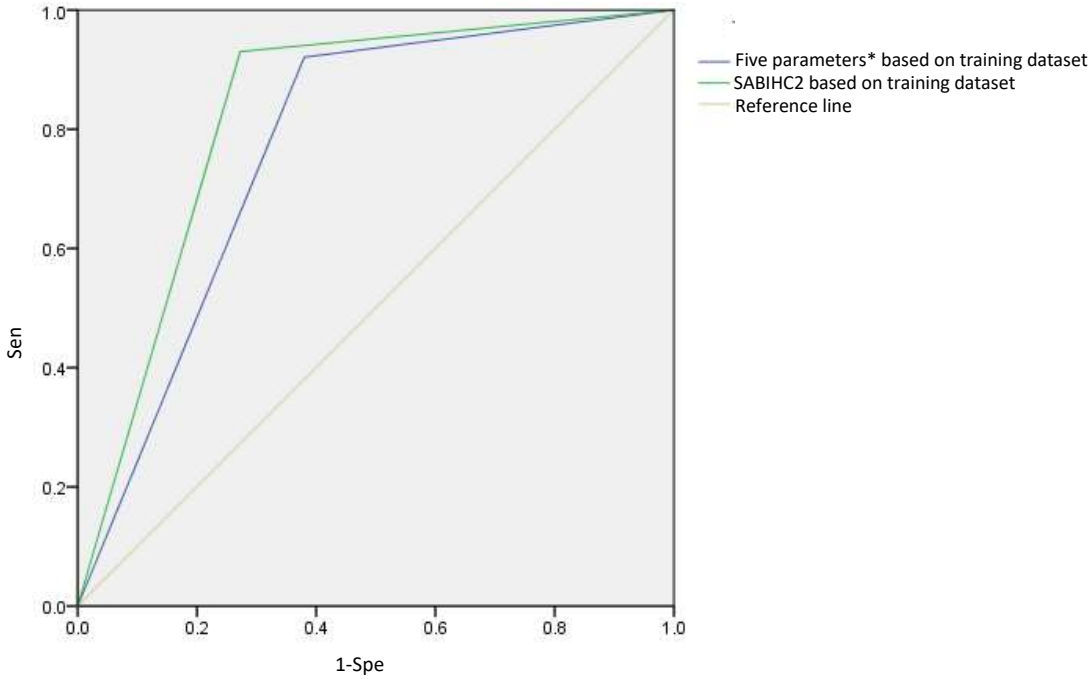
Component	Initial eigenvalues		
	Total	% of variance	Cumulative %
1	2.749	39.267	39.267
2	1.502	21.45	60.718
3	0.897	12.813	73.53
4	0.798	11.405	84.935
5	0.529	7.551	92.486
6	0.316	4.512	96.997
7	0.21	3.003	100

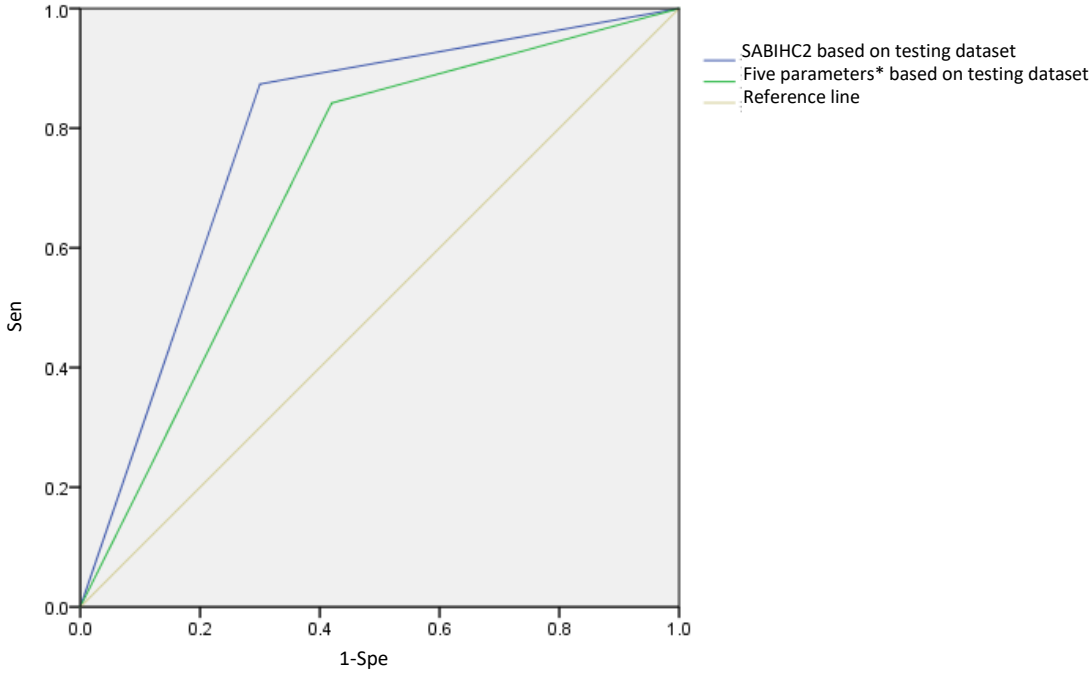


	Components				
	1	2	3	4	5
age	-0.509	0.469	0.519	0.175	0.469
NC	0.846	0.149	0.126	0.194	0.017
WC	0.849	0.304	0.135	-0.124	0.054
BMI	0.787	0.185	0.035	-0.483	0.145
MID	0.139	-0.594	0.749	-0.026	-0.251
H/TSD	-0.062	0.849	0.122	0.192	-0.457
Sex	0.64	-0.243	-0.128	0.666	0.109

How much better is the model compared to if you just used Age, Sex, NC, WC, BMI?

Response: We appreciate the kind advice from the reviewer. We supplemented the resulet of the model based on five parameters (Age, Sex, NC, WC, BMI). The SABIHC2 model performed better than the model based on five parameters, see the figures and table below.


Five parameters* based on training dataset
SABIHC2 based on training dataset
Reference line
Sen
1-Spe
0.0
0.2
0.4
0.6
0.8
1.0


SABIHC2 based on testing dataset
Five parameters* based on testing dataset
Reference line
Sen
1-Spe
0.0
0.2
0.4
0.6
0.8
1.0

		Sen	Spe	AUC	95% CI	+LR	-LR
SABIHC2	training dataset	0.930	0.727	0.829	0.777-0.880	3.407	0.096
	testing dataset	0.874	0.700	0.787	0.702-0.872	2.913	0.180
Five parameters*	training dataset	0.921	0.620	0.770	0.713-0.828	2.424	0.127
	testing dataset	0.842	0.580	0.711	0.617-0.805	2.005	0.272

*Five parameters mean the model based on Age, Sex, NC, WC and BMI.

In table 1, what are the odds rations relative to? 2SD change in predictor variable? Please clarify in the notes below the table.

Response: We appreciate the kind advice from the reviewer. We have added the sentence to the notes below the table in manuscript as follow. Odds ratios are depicted for moderate or severe OSA relative to no or mild OSA. And the data were presented as mean± SD.

Somewhere in the results, clarify the direction of the odd ratios for the significant variables. I.e. MID odds ratio of 1.3 means 1.3x greater odds of having OSA per X increase in MID (is that right?). While odds ratio of 0.3 for H/TSD means 3 times lower odds of having OSA per X increase in TSD for a given height.

Response: We appreciate the kind advice from the reviewer.

A mistake about the calculate method of H/TMD and H/TSD was found, which incorrectly calculate H/TMD as TMD to height. We corrected the mistake, and recalculated all results including t-test, logistic regression and machine learning model. Meanwhile, the two tables and three figures were updated. Thanks to the reviewer for pointing this mistake to our attention.

Discussion
Can you provide a physiological explanation for why MID, TMD and TSD are associated with OSA and difficult intubation?

Response:

Difficult intubation is related to limited mouth opening and head up, which can represent by MID and H/TMD, respectivity.[1] Some reports reported H/TSD to reflect the length of neck, which also influence intubation. [1,2] OSA is characterized by repeated occurrences of upper airway collapse and obstruction during sleep. Those parameters may partly represent upper airway structure, which could be used to screen OSA.

Reference:

1. Khan ZH, Mohammadi M, Rasouli MR, Farrokhnia F, Khan RH. The diagnostic value of the upper lip bite test combined with sternomental distance, thyromental distance, and interincisor distance for prediction of easy laryngoscopy and intubation: a prospective study. Anesth Analg. 2009 Sep;109(3):822-4.

2. Naguib M, Scamman FL, O'Sullivan C, Aker J, Ross AF, Kosmach S, Ensor JE. Predictive performance of three multivariate difficult tracheal intubation models: a double-blind, case-controlled study. Anesth Analg. 2006 Mar;102(3):818-24.

Another limitation of the study is that it is more challenging to use your model in the clinical environment compared to the STOP-BANG. In this way, the method has limited clinical utility in its current state. Is it possible to develop such a questionnaire that would assess OSA risk? This would provide excellent clinical applicability of your work. Further, it would allow other groups to test the questionnaire on their populations.

Response: Your suggestion really means a lot to us.

Compared to STOP-Bang questionnaire, the machine learning model is able to select and integrate numerous variables to optimize prediction tasks. As most models work, we are planning to develop a software or application in the future, to allow health care worker friendly installation and application.

VERSION 2 – REVIEW

REVIEWER	Palm, Andreas Uppsala Universitet, Department of Medical Sciences, Respiratory, Allergy and Sleep Research
REVIEW RETURNED	12-May-2021

GENERAL COMMENTS	Comments to author 12 May 2021 The revised manuscript has improved since the first submission but I miss comments from the author with clarifications of my previous issues. Some of my issues have been corrected in the manuscript, some have been red-marked and some have not. This aggravates the re-review process. The description of the statistical methods is still very confusing and I recommend an additional statistical review. In addition to my former issues I have some other minor issues Minor issues 1) Abstract, Results paragraph. Please be consequent with order, first AUC, then sensitivity followed by specificity throughout the manuscript. 2) Abstract, Conclusion section. Only a Chinese population is studied. The external validity is limited to this group. Please clarify. 3) Introduction, line 2. I would prefer the phrase “are estimated to be undiagnosed” instead of “are underdiagnosed patients”.

	4) Introduction line 3. These figures are also estimations, please clarify that. 5) Introduction, line 8. Which are the cardiovascular events? 6) Statistics. Please specify which regression model was used. Logistic regession? Why did you only adjust for BMI in the regression model and not also with gender and age, two important risk factors for OSA. 7) Results section. Please define MAI and LSpO2 and specify which units MAI and LSpO2 use. This was a single center cross-sectional observational study of 481 Chinese patients with suspected OSA. The study aimed to analyze how accurate a machine-learning model could predict the presence of moderate to severe obstructive sleep apnea. The authors conclude that the machine-learning model is accurate in predicting moderate to severe OSA, especially in patients without significant daytime symptoms. I have seven issues that require clarification from the authors, along with some minor suggestions for modifications. Major issues: 1. In Methods section page 6 line 3 it is mentioned that certain patients are excluded from the study if they had severe cognitive impairment, severe heart failure, severe respiratory failure or other serious acute or chronic diseases and neuromuscular diseases. How did you define the severity of these conditions? Did the patients have any comorbid conditions? 2. Description of statistical methods should be concentrated to a “Statistics paragraph”. 3. In Table 1, variables with P<0.05 found in chi-2 and t-tests are then used in the machine-learning model. p-values crude OR and OR after adjustment for BMI are presented. Why did you adjust for BMI? What does the crude logistic regression add? In page 9, line 29 it is stated that MID and H/TSD are independent risk-factors for OSA. Results found in cross-sectional studies says us nothing about causality and therefore it is more accurate to write about associations instead of risk-factors. In addition, there is an association to” moderate/severe OSA”, not “OSA”. 4. Why did the authors use the training set as reference group and not the AHI from the PSGs? With this design, there would be very strange if the results from the STOP-BANG questionnaires would be closer to the training set than the validation set. In the methods section the terms “training set” and “validation set” are used. In the results section, the “validation set” in suddenly called “testing set”. Very confusing. 5. The description of the model construction is confusing. Is the output of the model binary? What are the results from the machine-learning model? I miss a step between table 1 and table 2. Cross tables with outcome of PSG, the machine learning model, and the STOP-BANG results would be clarifying. 6. Flow chart, Figure 2. According to the methods section, p-values from t-tests and chi-2 tests were used to chose variables to include into the machine-learning model, not logistic regression. Only 95% Cis, not p-values are presented for the logistic regression in table 1 (good so).

	7. I wish that the authors suggest how this machine-learning model can be used in a clinical context. How can a clinician interpret the results? Minor issues: 1. This manuscript should improve with an English language review. 2. In the abstract, the result section is a bit confusing. Sensitivity and specificity are graded first with percentage and then with decimal numeral. There is no consistency in the way the different modalities are compared. The same problem exists in the results section. 3. Page 6, line 55. Should say “other caffeine-containing products, instead of “Cola-containing products” . 4. Page 6, line 59. Should say “thoracal”, instead of ribcage. 5. Page 7, line 5. The sentence “Sleep data were scored manually by registered polysomnographer of the United States according to American Academy of Sleep Medicine Recommendations (AASM)” is confusing. 6. The STOP-BANG questionnaire is a central part of the narrative of this study. It would be polite to the reader with a brief overview of this scale in the methods-section. 7. Table 1 should contain a row with characteristics of all patients. In row 4, it should say “p-value”, instead of “P”. Some values in the Mallampti test-rows are missing. Should say “N/A: not applicable”, instead of “NA: not available” 8. AHI-groups are categorized into “Deal” and “No-deal”. It would be more informative to just name the categories “Mild AHI” and “Moderate-severe OSA”. 9. In the methods section the terms “training set” and “validation set” are used. In the results section, the “validation set” in suddenly called “testing set”. Very confusing. Please also describe the rationale for dividing the patients in these groups.

REVIEWER	Gallo, Crescenzio University of Foggia, Clinical and Experimental Medicine
REVIEW RETURNED	22-Jun-2021

GENERAL COMMENTS	1) The English of the article should be revised and improved. 2) Comparison tests were used (eg. t-test) assuming that the distributions of the variables were normal; in fact, no normality test was performed on distributions. 3) Only the SVM method was used in the study: the authors could have considered other possible Machine Learning methods (such as Random Forest, CN2, kNN, Naive Bayes, AdaBoost and Artificial Neural Networks) that could have offered better performance in AUC, sensitivity and specificity. 4) Subdivision of patients into the two training and testing datasets was performed randomly, but it is not specified whether stratified subdivision was also performed. 5) Results were validated through the hold out method (70-30). It would have been better to use a more sophisticated approach, such as ten-fold cross-validation or even leave-one-out.

VERSION 2 – AUTHOR RESPONSE

Dear Reviewer 1, Dr. Andreas Palm, Uppsala Universitet

Comments to author 12 May 2021

The revised manuscript has improved since the first submission but I miss comments from the author with clarifications of my previous issues.

Some of my issues have been corrected in the manuscript, some have been red-marked and some have not. This aggravates the re-review process.

The description of the statistical methods is still very confusing and I recommend an additional statistical review.

Response: Thank you very much for your time involved in reviewing the manuscript. Your comments were highly insightful and enabled us to greatly improve the quality of our manuscript. We are very sorry for not accurately marking the modified part. And we apologize for the lack of clarity in the statistical methods and we have elaborated on this in the text (page 6, line 19).

In addition to my former issues I have some other minor issues

Minor issues

1) Abstract, Results paragraph. Please be consequent with order, first AUC, then sensitivity followed by specificity throughout the manuscript.

Response: Thank you for this advice. We have corrected the order throughout the manuscript.

2) Abstract, Conclusion section. Only a Chinese population is studied. The external validity is limited to this group. Please clarify.

Response: Thanks for your suggestions. The limitation was added in the Abstract and Conclusion section.

3) Introduction, line 2. I would prefer the phrase “are estimated to be undiagnosed” instead of “are underdiagnosed patients”.

Response: Thank you for this advice. We have corrected the “are underdiagnosed patients” into “are estimated to be undiagnosed”

4) Introduction line 3. These figures are also estimations, please clarify that.

Response: Thank you for your reminder. We have added “approximately” and “about” to clarify the estimated figures.

5) Introduction, line 8. Which are the cardiovascular events?

Response: Thank you for your reminder. According to the reference, cardiovascular events include myocardial injury, cardiac death, congestive heart failure, thromboembolism, atrial fibrillation, and stroke.

Reference:

1. Chan MTV, Wang CY, Seet E, et al. Postoperative Vascular Complications in Unrecognized Obstructive Sleep Apnea (POSA) Study Investigators. Association of Unrecognized Obstructive Sleep Apnea With Postoperative Cardiovascular Events in Patients Undergoing Major Noncardiac Surgery. JAMA. 2019 May 14;321(18):1788-1798.

6) Statistics. Please specify which regression model was used. Logistic regression? Why did you only adjust for BMI in the regression model and not also with gender and age, two important risk factors for OSA.

Response: Many thanks for this comment. The logistic regression was used. And we apologize for the lack of clarity in the statistical methods and we have elaborated on this in the text (page 6, line 23).

We agree with this comment that gender and age are two important risk factors for OSA. Consideration of the potential correlation between BMI and facio-cervical measurements, we adjusted for BMI in the regression model. We can add a Table including the adjusted OR for gender and age if deemed necessary.

7) Results section. Please define MAI and LSpO2 and specify which units MAI and LSpO2 use.

Response: Thank you for your reminder. We added the full names and units in the main text (page 9, line 8)

Dear Reviewer2, Dr. Crescenzio Gallo, University of Foggia

Thank you very much for your time involved in reviewing the manuscript and your very encouraging comments on the merits. Your comments were highly insightful and enabled us to greatly improve the quality of our manuscript. In the remainder of this letter, we discuss each of your comments individually along with our corresponding responses. We hope that the explanation has fully addressed all of your concerns.

Comments:

1) The English of the article should be revised and improved.

Response: Thanks for your suggestions. We have carefully edited the entire manuscript and the manuscript has been polished by a professional editor before resubmission. We hope the revised manuscript would satisfy you.

2) Comparison tests were used (eg. t-test) assuming that the distributions of the variables were normal; in fact, no normality test was performed on distributions.

Response: Many thanks for this comment. We apologize for the lack of clarity in the statistical methods and we have elaborated on this in the text (page 6, line 19).

3) Only the SVM method was used in the study: the authors could have considered other possible Machine Learning methods (such as Random Forest, CN2, kNN, Naive Bayes, AdaBoost and Artificial Neural Networks) that could have offered better performance in AUC, sensitivity and specificity.

Response: We thank the reviewer for this excellent suggestion and we agree it makes total sense.

Other studies have shown that compared to the other machine learning methods SVM is very powerful at recognizing subtle patterns in complex datasets.[1] so we choose the SVM as the method to predict moderate-severe OSA.
Reference:

1. Aruna S and Rajagopalan SP: A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. Int J Comput Appl 31(8): 14-20, 2011.

4) Subdivision of patients into the two training and testing datasets was performed randomly, but it is not specified whether stratified subdivision was also performed.

Response: Thanks for your reminder. We used stratified random sampling to divide patients and added the sentences in the Method part (page 7, line 15).

5) Results were validated through the hold out method (70-30). It would have been better to use a more sophisticated approach, such as ten-fold cross-validation or even leave-one-out.

Response: Thanks for your suggestions. We used the ten-fold cross-validation as the validated method (page 6, line 14) and changed the Results, Table 2, and Figures.

Special thanks to you for your good comments.

VERSION 3 – REVIEW

REVIEWER	Gallo, Crescenzio University of Foggia, Clinical and Experimental Medicine
REVIEW RETURNED	21-Jul-2021

GENERAL COMMENTS	The paper is now clearly stated and fully described.