



Developing and validating a chronic obstructive pulmonary disease quick screening questionnaire using statistical learning models

Chronic Respiratory Disease
Volume 19: 1–9
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14799731221116585
journals.sagepub.com/home/crd


Xiaoyue Wang^{1,2,†}, Hong He^{3,4,†}, Liang Xu^{5,†}, Cuicui Chen^{1,2,†}, Jieqing Zhang^{2,6}, Na Li⁵, Xianxian Chen⁵, Weipeng Jiang^{1,2}, Li Li^{1,2}, Linlin Wang^{1,2}, Yuanlin Song^{1,2}, Jing Xiao⁵, Jun Zhang^{3,4} and Dongni Hou^{1,2} 

Abstract

Background: Active targeted case-finding is a cost-effective way to identify individuals with high-risk for early diagnosis and interventions of chronic obstructive pulmonary disease (COPD). A precise and practical COPD screening instrument is needed in health care settings.

Methods: We created four statistical learning models to predict the risk of COPD using a multi-center randomized cross-sectional survey database ($n = 5281$). The minimal set of predictors and the best statistical learning model in identifying individuals with airway obstruction were selected to construct a new case-finding questionnaire. We validated its performance in a prospective cohort ($n = 958$) and compared it with three previously reported case-finding instruments.

Results: A set of seven predictors was selected from 643 variables, including age, morning productive cough, wheeze, years of smoking cessation, gender, job, and pack-year of smoking. In four statistical learning models, generalized additive model had the highest area under curve (AUC) value both on the developing cross-sectional data set (AUC = 0.813) and the prospective validation data set (AUC = 0.880). Our questionnaire outperforms the other three tools on the cross-sectional validation data set.

Conclusions: We developed a COPD case-finding questionnaire, which is an efficient and cost-effective tool for identifying high-risk population of COPD.

¹Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai, China

²Shanghai Key Laboratory of Lung Inflammation and Injury, Shanghai, China

³Department of Anesthesiology, Fudan University Shanghai Cancer Center, Shanghai, China

⁴Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China

⁵AI Center, Ping An Technology (Shenzhen) Co. Ltd, Shenzhen, China

⁶Department of Pharmacy, Zhongshan Hospital, Fudan University, Shanghai, China

[†]Those authors contributed equally to this work.

Corresponding authors:

Jing Xiao, Ping An Insurance(Group) Company of China, Ltd, 23/F, Ping An Financial Center, 5033 Yitian Road, Futian District, Shenzhen, China.
Email: XIAOJING661@pingan.com.cn

Jun Zhang, Department of Anesthesiology, Fudan University Shanghai Cancer Center; Department of Oncology, Shanghai Medical College, Fudan University, 270 Dong an Road, Shanghai 200032, China.
Email: snazhang@aliyun.com&emsp

Dongni Hou, Department of Pulmonary and Critical Care Medicine, Zhongshan Hospital, Fudan University, 180 Fenglin Road, Xuhui District, Shanghai 200032, China.
Email: houdn2014@126.com



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Keywords

chronic obstructive pulmonary disease, machine learning, generalized additive model, screening, smoking

Date received: 21 March 2022; accepted: 1 July 2022

Introduction

Underdiagnosis is a major challenge worldwide. A recent national cross sectional study in China reported less than 3% of COPD patients were aware of their condition and few of them received a previous pulmonary function test.¹ Other community-based population studies from North and South America, Europe, and Australia have revealed that about 70–80% of these subjects have not been diagnosed with COPD.^{2–4}

Inadequate-utilization of spirometry contributes most to the high-rate of underdiagnosis of COPD.⁵ Patients' poor access to spirometers and lack of expertise in performing and interpreting spirometry limit spirometry-based diagnosis especially in primary care. In addition, a considerable part of the early stage COPD patients with only mild airflow limitation have few or nonspecific symptoms or poor perception of their symptoms.⁶ A precise, practical and cost-effective screening strategy is urgently needed for promoting early diagnosis and interventions, especially in high risk population area.

Active targeted case finding in health care setting with screening questionnaires prior to spirometry test has been demonstrated a cost-effective way to identify undiagnosed patients and was recommended in GOLD 2020.⁷ Different questionnaires have been reported, e.g. the COPD Population Screener Questionnaire (COPD-PS),⁸ based only on a few priori selected items and does not account for risk factors such as occupation, education level, living conditions, and prenatal maternal smoking. These risk factors vary in different countries and regions, and their effects have not been evaluated in COPD screening. Moreover, the previous tools only provided a risk score rather than a predicted probability of having COPD.

In this study, we used statistical learning algorithms on a database of national cross-sectional survey study from China to identify novel predictive patterns of significant clinical characters. We then compared the performance of four statistical learning models to find the best predictive model. Finally, we developed a COPD screening questionnaire and validated its efficacy in a prospective cohort of specialist population.

Methods

Study subjects and study design

The developing data set consists of 5281 participants in the China Pulmonary Health (CPH) cohorts from Shanghai.

The design of this national randomized cross-sectional study has been previously described.¹ All participants were >20 years old and received standardized spirometry measurement between June 2012 and February 2014. To further validate our case-finding instrument, we established another prospective observational cohort. Participants who underwent spirometry tests in a tertiary teaching hospital were enrolled between April 2020 and September 2020. Participants were excluded if they were <40 years old; had history of asthma or lung cancer; had history of thoracic or abdominal surgery; admitted to hospital for any cardiac condition in the past month; had heart rate greater than 120 beats per min; were under antibacterial chemotherapy for tuberculosis; were pregnant or breastfeeding; or they did not receive bronchial dilation test. Information for the selected predictors was collected in a face-to-face interview using a questionnaire before or after they underwent spirometry tests. The diagnosis of COPD was based on the objective measure post-bronchodilator FEV1/FVC with value <70%.⁹ All participants were provided written informed consent, and the ethics review committees of Beijing Capital Medical University (No. 11-ke-42) and Zhongshan Hospital Fudan University (No. B2019-248 (2)) approved this work.

Data processing

The cross-sectional database consists a total of 643 variables except for spirometry data. We eliminated the variables and samples of high missing rate (Supporting Methods and Supplemental Figure S1), remaining 159 candidate predictors of interest and 4736 participants for modeling. The predictors cover patient demographics (e.g. age, gender, body-mass index (BMI)), respiratory symptoms, activity limitation, depression and anxiety symptoms, medical history and medication, cigarette smoke exposure, occupation and living environment, quality of life, results of physical examination and laboratory tests. We used different impute methods according to the intrinsic logic of original questions, see details in Supplementary Materials - Methods - Data processing and Supplemental Table S1.

Statistical learning models

Firstly, the cross-sectional database were used to establish models for predicting risk of COPD. We evaluated the performance of four statistical learning models on

predicting the risk of COPD using the cross-sectional database, including logistic regression (LR), generalized additive model (GAM), extreme gradient boosting (XGBoost), and random forest (RF). Before using the training data to train the model, we tested and determined the value of the hyperparameters for GAM, XGBoost and RF models. After selecting the feature set, the 10-fold cross-validation method was used to determine the values of hyperparameters according to the highest AUC score on the cross-sectional training set using the grid search. Then we retrained the model on the whole cross-sectional training data. Interaction was added accounting for the relationship between age and years of smoking cessation in LR and GAM.

Selection of predictors

To screen for the most effective predictors and optimal prediction model, we randomly selected one-tenth of the cross-sectional survey data as the internal validation data, and the remaining nine-tenths as training data (Supplemental Figure S2). The minimal set of predictors providing the highest average AUC was selected as final predictors based on their integrated importance in the four statistical models (Supporting Methods). The optimal prediction model was selected from the four prediction models for best AUC on cross-sectional internal validation data. Finally, we combined the selected predictors and statistical learning model into a COPD case-finding instrument which directly predicts the probability of having irreversible airway obstruction in spirometry tests for an individual.

Validation

To evaluate the reliability of our COPD case-finding model, we compared performance of our model and other three previously reported approaches by Zarowitz et al. (2011),¹⁰ Kotz et al. (2008)¹¹ and Price et al. (2006)¹² in COPD case-finding. For each participant in our cross-sectional cohort, we used our model and three previously reported models to predict if he/she had COPD. Then we compared the receiver operating characteristic (ROC) curve and AUC of the four models in identifying spirometry-confirmed COPD cases. The features included in the three previous approaches were presented in Supplemental Table S2.

Determination of cut-off values

We calculated the cut-off values of predicted probability for high-risk population to make positive predict value (PPV) higher than 0.5 and for low-risk population to make negative predictive value (NPV) < 0.02.^{1,13}

Data were expressed as frequencies (percentages) for categorical variables and as means \pm SD or median (IQR).

Student's t-test or Mann-Whitney *U* test were used for comparison of continuous variables as appropriate. Chi squared test was used to compare parametric and categorical variables, respectively. LR and GAM models were implemented in R (stats, mgcv 1.8.33) and the other analysis is carried out by Python (Pandas 1.0.1, Scikit-Learn 0.23.2, pygam 0.8.0, XGBoost 1.0.2). *p*-values < 0.05 were interpreted as statistically significant.

Results

Participant characteristics

Data from 4736 individuals were used for modeling. Demographics and clinical characteristics were summarized in Table 1. Approximately 44% of the sample was male. A percentage of 11.6 of the participants had spirometry-defined COPD. The proportions of GOLD stage I, II, III, and IV were 53.7%, 38.0%, 7.4%, and 0.9%, respectively. Current and former smokers made up 25.1% of non-COPD and 43.1% of the COPD subgroup. Only 59 (10.7%) participants had previous diagnosis of COPD in COPD subgroup. The prospective validation data set included 958 patients undergoing spirometry test. Compared with patients in the cross-sectional data set, those in the prospective data set were more likely to have smoke exposure history (45.7%) and suffer more respiratory symptoms (43.9%) and more severe airway obstruction.

Selection of predictors

A total of 157 enrolled candidate predictors were used for final analysis (see details in online Appendix Table S3). In stepwise logistic regression, 48 predictors were selected as for the smallest AIC. Eight most important predictors were selected from the pool of 48 predictors based on the summed ranking of four predictive models (Supplemental Table S4): age, saturation of peripheral Oxygen (SpO₂), morning productive cough, wheeze, years of smoking cessation, gender, job, and pack-years of smoking, which provided the highest average AUC and smallest set of predictors (Supplemental Figure S3). The descriptive statistics of the eight selected predictors were shown in Table 2. All of the predictors had significant difference between COPD and non-COPD groups in cross-sectional data. Considering SpO₂ is not widely available in primary care settings, we compared the final AUC of the four models with the eight predictors and seven predictors without SpO₂, the difference in average AUC was 0.008 (0.784 vs 0.792). Thus, SpO₂ was excluded, remaining seven predictors in the final predictors set. The odds ratio (OR) of each predictors in LR and GAM model were shown in Supplemental Table S6.

Table 1. Clinical characteristics of participants in cross-sectional data set and prospective validation data set.^a

Characteristics	Cross-sectional data set			Prospective data set		
	Non-COPD (N = 4185)	COPD (N = 551)	p-value	Non-COPD (N = 766)	COPD (N = 192)	p-value
Age (year)	53.2 ± 12.3	63.9 ± 10.1	<.001	61.1 ± 9.7	69.1 ± 9.3	<.001
Male (%)	1719 (41.1)	344 (62.4)	<.001	403 (52.6)	163 (84.9)	<.001
Height (cm)	161.4 ± 8.0	162.2 ± 8.2	.009	163.6 ± 8.0	165.9 ± 7.0	<.001
Weight (kg)	63.3 ± 10.8	64.1 ± 11.1	.11	63.4 ± 10.4	65.4 ± 10.5	0.04
BMI (kg/m ²)	24.2 ± 3.3	24.3 ± 3.5	.58	23.7 ± 3.2	23.7 ± 3.2	.88
Cigarette smoke exposure						
Current smoker	869 (20.8)	171 (31.0)	<.001	153 (20.0)	41 (21.4)	.74
Former smoker with passive smoking	178 (4.2)	65 (11.7)	<.001	85 (11.1)	60 (31.2)	<.001
Former smoker without passive smoking	6 (0.1)	2 (0.4)	.53	59 (7.7)	40 (20.8)	<.001
Never-smoker with passive smoking	2817 (67.3)	287 (52.0)	<.001	264 (34.4)	17 (8.9)	<.001
Never-smoker without passive smoking	315 (7.6)	26 (4.9)	.02	205 (26.8)	34 (17.7)	.01
Pack-year in current smoker (pack-years)	6.7 ± 15.1	15.0 ± 21.7	<.001	12.1 ± 21.8	29.7 ± 28.4	<.001
Respiratory symptoms						
Dyspnea	199 (4.8)	64 (11.6)	<.001	31 (4.0)	74 (38.5)	<.001
Wheeze	145 (3.5)	97 (17.6)	<.001	81 (10.6)	123 (64.1)	<.001
mMRC grade ≥3	353 (8.4)	179 (32.4)	<.001	—	—	—
Chronic cough	286 (6.8)	84 (15.2)	<.001	122 (15.9)	68 (35.4)	<.001
Chronic phlegm	269 (6.4)	94 (17.0)	<.001	130 (17.0)	104 (54.2)	<.001
Any of the above respiratory symptoms	894 (21.4)	269 (48.7)	<.001	262 (34.2)	159 (82.8)	<.001
COPD grade						
I	—	334 (60.6)	—	—	32 (16.7)	—
II	—	173 (31.4)	—	—	79 (41.1)	—
III	—	38 (6.9)	—	—	65 (33.9)	—
IV	—	6 (1.1)	—	—	16 (8.3)	—
FEV1 (mL)	2693.3 ± 658.3	2085.0 ± 669.6	<.001	2771.1 ± 711.0	1530.3 ± 625.1	<.001
FVC (mL)	3276.5 ± 798.4	3382.3 ± 1057.3	.03	3333.4 ± 920.9	2685.1 ± 759.4	<.001
FEV1/FVC (%)	82.4 ± 6.3	61.5 ± 9.2	<.001	83.7 ± 5.2	55.8 ± 11.4	<.001
Previous diagnosis of respiratory conditions						
Asthma	34 (0.8)	34 (6.2)	<.001	12 (1.6)	22 (11.5)	<.001
COPD	11 (0.3)	17 (3.1)	<.001	0 (0)	136 (70.8)	<.001
Tuberculosis	13 (0.3)	5 (0.9)	.08	21 (2.7)	16 (8.3)	.007
Chronic bronchitis	144 (3.4)	66 (12.0)	<.001	42 (5.5)	58 (30.2)	<.001

^a Data are presented as % or mean ± SD, p-value are calculated based on Chi-square or Mann–Whitney U test.

Modelling

The results of four models with seven selected predictors were shown in Table 3. On the cross-sectional data set, GAM model had the highest AUC value (AUC 0.813), followed by LR (AUC 0.811) and XGBoost (AUC 0.810). On the prospective validation data set, the GAM model also achieved the best performance on clinical test data with the AUC value of 0.880 (Table 3). In addition, the width of confidence band of the GAM model was significantly smaller than the other three models, which indicated that the GAM model was more robust. Therefore, we used GAM model to construct a new COPD case-finding instrument called COPD Quick Screening Questionnaire (COPD-QSQ). The questions

included in COPD-QSQ were presented in Table 4 and details for calculating the risks using final GAM model were listed in Supplemental Table S6 and S7 and Supplemental Figure S4. Details of the estimated effects of predictors in GAM and LR model were presented in Supplemental Table S5 and S6 and Supplemental Figure S4. The feature importance and SHAP values of XGBoost and RF model were shown in Supplemental Figure S5.

Comparison with previous questionnaires

Compared with other three instruments previously reported by Ref. Zarowitz et al. (2011), Kotz et al. (2014)

Table 2. Statistics of selected predictors of COPD in cross-sectional and prospective data sets.^a

Predictors	Cross-sectional data set			Prospective data set		
	Non-COPD (N = 4185)	COPD (N = 551)	p-value ^b	Non-COPD (N = 766)	COPD (N = 192)	p-value ^b
Age	53.2 ± 12.3	63.9 ± 10.1	<.001	61.1 ± 9.7	69.1 ± 9.3	<.001
Gender						
Female	2466 (58.9)	207 (37.6)	<.001	363 (47.4)	29 (15.1)	<.001
Male	1719 (41.1)	344 (62.4)		406 (52.6)	163 (84.9)	
Pack-years	0.0 (0.0 - 0.0)	0.0 (0.0 - 30.25)	<.001	0.0 (0.0 - 20.75)	25.5 (0.0 - 45.0)	<.001
Job						
Unemployed	565 (13.5)	118 (21.4)	<.001	435 (56.8)	144 (75.0)	<.001
Worker	1086 (25.9)	73 (13.2)		38 (5.0)	9 (4.69)	
Farmer	842 (20.1)	146 (26.5)		62 (8.1)	21 (10.9)	
Technical stuffs	197 (4.7)	16 (2.9)		41 (5.4)	4 (2.1)	
Housekeeper	89 (2.1)	9 (1.6)		9 (1.2)	2 (1.0)	
Official	76 (1.8)	4 (0.7)		37 (4.8)	2 (1.04)	
Driver	69 (1.6)	7 (1.3)		5 (0.7)	1 (0.5)	
Cook	38 (0.9)	1 (0.2)		1 (0.1)	0 (0)	
Student	19 (0.5)	1 (0.2)		0 (0)	0 (0)	
Others	1190 (28.4)	174 (31.6)		138 (18.0)	9 (4.69)	
Years of smoking cessation	0.0 (0.0 - 0.0)	0.0 (0.0 - 0.0)	<.001	0.0 (0.0 - 0.0)	1.00 (0.00 - 5.25)	.34
Morning productive cough	571 (13.6)	102 (18.5)	<.001	80 (10.4)	68 (35.4)	<.001
Wheeze	127 (3.0)	94 (17.1)	<.001	81 (10.6)	123 (64.1)	<.001

^aData are presented as n (%), median (IQR1~ IQR3), or mean ± SD.

^bp-value are calculated based on Chi-square test or Mann-Whitney U test.

Table 3. Predictive ability of different models on cross-sectional validation dataset and prospective cohort dataset.

Models ^a	AUC (95% CI)	Sensitivity	Specificity	PPV	NPV	Accuracy	p-value
Cross-sectional validation dataset							
GAM	0.813 (0.753–0.867)	0.91	0.55	0.21	0.98	0.59	<.001
LR	0.811 (0.747–0.867)	0.89	0.51	0.21	0.98	0.61	<.001
RF	0.702 (0.628–0.777)	0.4	0.88	0.39	0.92	0.86	N/A
XGBoost	0.810 (0.750–0.864)	0.98	0.19	0.14	0.99	0.3	N/A
Prospective cohort dataset							
GAM	0.880 (0.848–0.910)	0.98	0.23	0.24	0.98	0.38	<.001
LR	0.869 (0.836–0.901)	0.97	0.26	0.24	0.97	0.4	<.001
RF	0.875 (0.844–0.906)	0.84	0.71	0.41	0.95	0.73	
XGBoost	0.869 (0.832–0.901)	1	0.06	0.21	1	0.25	

AUC: area under the receiver operating characteristic curve; PPV: positive predictive value; NPV: negative predictive value; NA: not available for the model; LR: logistic regression; GAM: generalized additive model; RF: random forest; XGBoost: extreme gradient boosting; COPD: chronic obstructive pulmonary disease.

^aFor each model, we defined probability higher than 0.075 as with risk for COPD, and the others without for COPD to calculate the metrics. We used spirometry-defined COPD as gold standard.

and Price et al. (2006), our COPD-QSQ outperformed the other instruments with an AUC of 0.813 on the cross-sectional data set. (Figure 1) Tools by Kotz et al. (2014) and Price et al. (2006) used LR model and similar

predictors and showed similar results on both data sets (AUC 0.770 and 0.774, respectively). Zarowitz et al. (2011) contained the smallest number of questions, while the AUC was lower than other instruments.

Table 4. COPD quick screening questionnaire (COPD-QSQ).

Characteristics	Items in questionnaire	Answers
1. Sex	What is your sex?	Male/Female
2. Wheeze	Have you ever wheezed?	Yes/No
3. Morning productive cough	Do you often cough up sputum when you wake up in the morning?	Yes/No
4. Job	What is your current job?	Official/Unemployed/Farmer/Technical stuffs/Worker, cooker, or housekeeper/Others
5. Years of smoking cessation	How many years have you quit smoking?	Number
6. Age	How old are you (years)?	Number
7. Pack-years	How many packs of cigarettes do you smoke a year?	Number

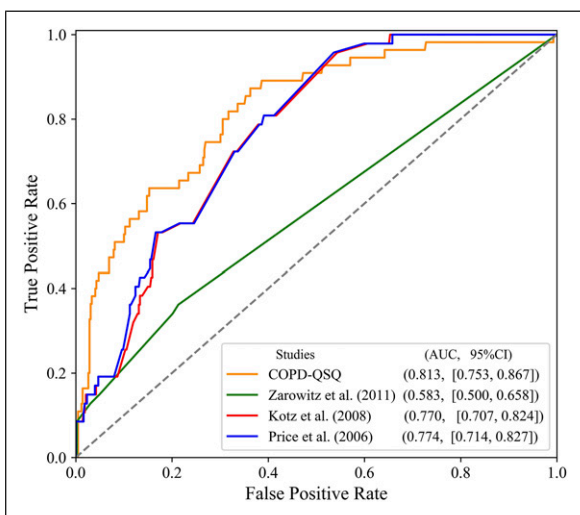


Figure 1. Comparisons of the area under curve (AUC) between generalized additive model and three previous approaches on the cross-sectional validation data. The models included seven predictors: age, morning productive cough, wheeze, years of smoking cessation, gender, job, and pack-years of smoking.

Optimal cut off value for spirometry test

Our questionnaire is aimed at identifying the individuals with high risk of COPD who need further validation by spirometry test, and those with low risk of COPD who should not receive spirometry test. To determine the cut-off value of defining high-risk population using our model, balancing between cost and effectiveness, we adopted an optimal PPV of 0.5 to define “high-risk” group, where the corresponding cut-off value in GAM model was 0.265. (Figure 2) It means that individuals who had value of 0.265 need further confirmative spirometry test, and to identify a case with airflow obstruction, two high-risk individuals were required to receive spirometry test. In addition, we also defined a low-risk population who should not receive

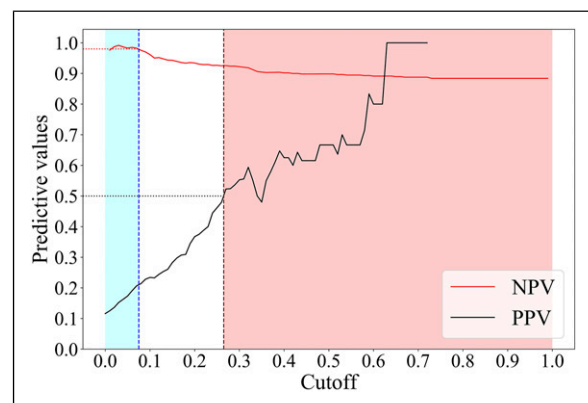


Figure 2. Positive predictive value (PPV) and negative predictive value (NPV) of generalized additive model prediction results.

The red vertical line presents the high-risk cut-off value of 0.265 using generalized additive model (GAM) with an optimal PPV of 0.5. The blue vertical line presents the low-risk cut-off value of 0.075 with an optimal NPV of 0.98 using the same model. The model included seven predictors: age, morning productive cough, wheeze, years of smoking cessation, gender, job, and pack-years of smoking.

spirometry test for COPD screening. We used the NPV of 0.98 and the corresponding cut-off value was 0.075, that is, one case with airflow obstruction would be missed among 50 low-risk individuals without spirometry test. (Figure 2) The predicted risk between 0.075 and 0.265 was defined as moderate risk.

Discussion

To our knowledge, this is the first study using advanced statistical learning models to predict the risk of COPD in general population and develop a case-finding instrument for COPD. With the abundant variables in dataset collected

from cross-sectional study, we identified seven important predictors showing high predicting power with an average AUC of 0.784 for detecting spirometry-defined COPD patients. The highest AUC was reached by the GAM model as 0.813. The case-finding instrument derived from the selected predictors and GAM model had higher AUC of 0.880 for risk prediction of COPD in our prospective validation cohort.

Our developing data set covered a wide range of potential predictors of COPD in general population, including age, living conditions, income, job, biomass usage, childhood lung infections, and maternal smoking during pregnancy. The participants were not restricted to smokers, which permits the applicability of our model in screening for non-smokers with high risk of COPD. We found predictors (such as age, gender, symptoms and smoking) which have been commonly used in previous case-finding tools,^{7,8,12,14} and also occupational exposure which has not been included in other case-finding tools. In line with the result, a statement from American Thoracic Society and European Respiratory Society reported that occupational exposure contributed 14% to the burden of COPD, which was twice higher in non-smokers.¹⁵ To keep the simplicity of our case-finding instruments, we classified the occupations into nine classes according to their estimated OR for COPD. In addition, we found years of smoking cessation was of high importance which was not included in other instruments.

Difference in detailed definition of predictors should be accounted for in a screening questionnaire. For each item, the most efficient question were selected from several related candidates using AIC value of stepwise regression and the sum importance ranking from four statistical models. For example, the candidate question related to job included: “your current job?”, “how many years did you had this job?”, “the job you ever had for longest time?”, and “did you exposed to dust, allergens, or noxious gas in your working place?”. Current job performed best than other questions. In spite of cautious selection of predictors, there are special conditions to be considered in clinical application, such as for retired individuals, the last job before retirement is a reasonable substitute to account for occupation factor. In addition, the question “do you ever wheeze” is not limited to a recent period, which would not miss individuals who have chronic and recently onset wheeze, but may also screen out those who ever had acute wheeze and have recovered.

Previously published COPD case-finding models were mostly based on logistic or multivariate regression techniques.^{16–18} Statistical learning models have been used to predict the risk of mortality after acute stroke and acute myocardial infarction,^{19,20} the risk of drug toxicity,²¹ and the deterioration of patients in critical care units.²² The large sample size of our study permits usage of advanced statistical learning strategies in developing a case-finding model. The final AUCs of our prediction model were

between 0.80 and 0.90, which were higher than those of other instruments. In comparison, our final models outperformed other previously reported models both on the cross-sectional survey data.

Different from ensemble models, GAM adds a non-parametric part to characterize the sub-linearity of factors and has a strict penalty for non-parametric smoothness, which permits better fitting and prediction competence. In addition, GAM provides a calculated probability for each individual, which allows physicians to assess the risk of having COPD and to make clinical decisions accordingly. Also, GAM permits analyzing the pathogenic factors of the examinee. For example, the 2D regression curve in our GAM model ([Supplemental Figure S4](#)) provided a moderate high-risk region of age and smoking cessation.

In addition to the advantages of GAM model, our study has some inherent data strengths in both cardinality and degree of database. The training database was derived from a cross-sectional study with strict multi-stage randomized sampling, the sample size of which, to the best of our knowledge, is larger than that of any previous studies.^{23–25} Also, our predictors were selected from a large candidate pool with four different statistical learning models, underlying the importance of their role in screening. The COPD-QSQ questionnaire developed in this study classifies high-and low-risk population with the probability of having COPD as 50% and 2%, respectively. It provides a feasible, cost-effective, and precise case-finding tool for clinical use in health care settings.

There are also limitations in our study. Firstly, our study population was from a highly industrialized city of China, while the importance of risk factors for COPD differs in different economic and cultural context. There may not be universal equation/questionnaire that fit all countries and regions. Large datasets from different countries and regions are required to further test the generalization ability of our proposed method. Secondly, the cross-sectional database had variables and samples with missing values. Albeit efforts in balancing data saturation and sample size and cautiously imputation, the missing values may still introduce confounding in the final model. However, we noted that the selected features in the final model were biologically plausible and mostly reported as common risk factors of COPD, suggesting the elimination of variables did not miss important information for prediction. Thirdly, our validation cohort included individuals underwent spirometry tests in a generalized tertiary hospital for diverse reasons (eg. mild or severe respiratory symptoms, risk of respiratory disease, pre-operation assessment, and routine respiratory health examination, ect.) Despite of their variety in health background, the population may still different from that of primary or secondary care settings. Forth, individuals with previously diagnosed COPD patients were included in our cohorts, where the prevalence of COPD may higher than its aimed population of COPD

screening. The performance of our questionnaire needs further evaluation in multi-center prospective COPD screening studies at different health care settings.

Conclusions

In this study, we proposed a set of predictors to accurately predict risk of COPD and designed a new case-finding questionnaire for COPD called COPD-QSQ. This questionnaire has potential applications in different health care settings to assist physicians in identifying individuals of high risk for COPD.

Author contribution

Xiaoyue Wang and Liang Xu contributed substantially to data collection, patient management, statistical analysis and interpretation, and the writing of the manuscript. Hong He contributed to statistical analysis and writing of the manuscript. Dr Na Li and Ms. Xianxian Chen is the artificial intelligence expert who was responsible for data processing, statistical analysis, methodology design and implementation of the algorithm, and participated in manuscript writing and diagram plotting. Cuicui Chen, Weipeng Jiang, Li Li, Linlin Wang, Jian Wang, Mengzhen Cheng, and Jieqing Zhang contributed to the acquisition, analysis, or interpretation of data. Yuanlin Song contributed to statistical analysis and writing of the manuscript. Zhang Jun and Jing Xiao contributed to administrative, technical, and material support and to critical revision of the manuscript for important intellectual content. Dongni Hou contributed substantially to the study design, data collection, statistical analysis and interpretation, and the writing of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by The National Natural Science Foundation of China (81800008, 81770075, 82041003), National key R&D plan (2020YFC2003700), Science and Technology Commission of Shanghai Municipality (20DZ2261200, 20Z11901000, 20XD1401200), and Shanghai Municipal Key Clinical Specialty (shslczdk02201).

Ethics statement

All participants were provided written informed consent, and the ethics review committees of Beijing Capital Medical University (No. 11-ke-42) and Zhongshan Hospital Fudan University (No. B2019-248(2)) approved this work.

Availability of data and materials

The corresponding authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. The data are available from the corresponding authors upon reasonable request.

ORCID iD

Dongni Hou  <https://orcid.org/0000-0001-8332-7321>

Supplemental Material

Supplemental material for this article is available online.

References

1. Wang C, Xu J, Yang L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. *Lancet* 2018; 391: 1706–1717. DOI: [10.1016/S0140-6736\(18\)30841-9](https://doi.org/10.1016/S0140-6736(18)30841-9).
2. Diab N, Gershon AS, Sin DD, et al. Underdiagnosis and overdiagnosis of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2018; 198: 1130–1139. DOI: [10.1164/rccm.201804-0621CI](https://doi.org/10.1164/rccm.201804-0621CI).
3. Kaplan A and Thomas M. Screening for COPD: the gap between logic and evidence. *Eur Respir Rev* 2017; 26. DOI: [10.1183/16000617.0113-2016](https://doi.org/10.1183/16000617.0113-2016).
4. Mannino DM, Gagnon RC, Petty TL, et al. Obstructive lung disease and low lung function in adults in the United States: data from the National Health and Nutrition Examination Survey, 1988–1994. *Arch Intern Med* 2000; 160: 1683–1689. DOI: [10.1001/archinte.160.11.1683](https://doi.org/10.1001/archinte.160.11.1683).
5. Lamprecht B, Soriano JB, Studnicka M, et al. Determinants of underdiagnosis of COPD in national and international surveys. *Chest* 2015; 148: 971–985. DOI: [10.1378/chest.14-2535](https://doi.org/10.1378/chest.14-2535).
6. Labaki WW and Han MK. Improving detection of early chronic obstructive pulmonary disease. *Ann Am Thorac Soc* 2018; 15: S243–S248. DOI: [10.1513/AnnalsATS.201808-529MG](https://doi.org/10.1513/AnnalsATS.201808-529MG).
7. Jordan RE, Adab P, Sitch A, et al. Targeted case finding for chronic obstructive pulmonary disease versus routine practice in primary care (TargetCOPD): a cluster-randomised controlled trial. *Lancet Respir Med* 2016; 4: 720–730. DOI: [10.1016/s2213-2600\(16\)30149-7](https://doi.org/10.1016/s2213-2600(16)30149-7).
8. Martinez FJ, Raczek AE, Seifer FD, et al. Development and initial validation of a self-scored COPD population screener questionnaire (COPD-PS). *COPD* 2008; 5: 85–95. DOI: [10.1080/15412550801940721](https://doi.org/10.1080/15412550801940721).
9. Nishino M. Perinodular radiomic features to assess nodule microenvironment: does it help to distinguish malignant versus benign lung nodules? *Radiology* 2019; 290: 793–795. DOI: [10.1148/radiol.2018182619](https://doi.org/10.1148/radiol.2018182619).

10. Zarowitz BJ, O'Shea T, Lefkowitz A, et al. Development and validation of a screening tool for chronic obstructive pulmonary disease in nursing home residents. *J Am Med Directors Assoc* 2011; 12: 668–674. DOI: [10.1016/j.jamda.2010.11.007](https://doi.org/10.1016/j.jamda.2010.11.007).
11. Kotz D, Simpson CR, Viechtbauer W, et al. Development and validation of a model to predict the 10-year risk of general practitioner-recorded COPD. *NPJ Prim Care Respir Med* 2014; 24: 14011. DOI: [10.1038/npjpcrm.2014.11](https://doi.org/10.1038/npjpcrm.2014.11).
12. Price DB, Tinkelman DG, Nordyke RJ, et al. Scoring system and clinical application of COPD diagnostic questionnaires. *Chest* 2006; 129: 1531–1539. DOI: [10.1378/chest.129.6.1531](https://doi.org/10.1378/chest.129.6.1531).
13. Llordés M, Zurdo E, Jaén Á, et al. Which is the best screening strategy for COPD among smokers in primary care? COPD. *J Chronic Obstructive Pulm Dis* 2017; 14: 43–51. DOI: [10.1080/15412555.2016.1239703](https://doi.org/10.1080/15412555.2016.1239703).
14. Yawn BP, Mapel DW, Mannino DM, et al. Development of the lung function questionnaire (LFQ) to identify airflow obstruction. *Int J Chron Obstruct Pulmon Dis* 2010; 5: 1–10.
15. Blanc PD, Annesi-Maesano I, Balmes JR, et al. The occupational burden of nonmalignant respiratory diseases. An official American thoracic society and European respiratory society statement. *Am J Respir Crit Care Med* 2019; 199: 1312–1334. DOI: [10.1164/rccm.201904-0717ST](https://doi.org/10.1164/rccm.201904-0717ST).
16. Spyrtos D, Haidich AB, Chloros D, et al. Comparison of three screening questionnaires for chronic obstructive pulmonary disease in the primary care. *Respiration* 2017; 93: 83–89. DOI: [10.1159/000453586](https://doi.org/10.1159/000453586).
17. Ma X, Wu Y, Zhang L, et al. Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population. *J Transl Med* 2020; 18: 146. DOI: [10.1186/s12967-020-02312-0](https://doi.org/10.1186/s12967-020-02312-0).
18. Inoue H, Tsukuya G, Samukawa T, et al. Comparison of the COPD population screener and international primary care airway group questionnaires in a general Japanese population: the hisayama study. *Int J Chronic Obstructive Pulm Dis* 2016; 11: 1903–1909. DOI: [10.2147/copd.S110429](https://doi.org/10.2147/copd.S110429).
19. Heo J, Yoon JG, Park H, et al. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke* 2019; 50: 1263–1265. DOI: [10.1161/strokeaha.118.024293](https://doi.org/10.1161/strokeaha.118.024293).
20. Goldstein BA, Navar AM and Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges, 2017, 1805–1814.
21. Zhang L, Zhang H, Ai H, et al. *Applications of machine learning methods in drug toxicity prediction*, 2018, 987–997.
22. Shillan D, Sterne JAC, Champneys A, et al. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review, 2019, 284.
23. Yan L, Zhang H-T, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Machine Intelligence* 2020; 2: 283–288. DOI: [10.1038/s42256-020-0180-7](https://doi.org/10.1038/s42256-020-0180-7).
24. Hu C, Liu Z, Jiang Y, et al. Early prediction of mortality risk among patients with severe COVID-19, using machine learning. *Int J Epidemiol* 2020; 49: 1918–1929. DOI: [10.1093/ije/dyaa171](https://doi.org/10.1093/ije/dyaa171).
25. Peng J, Chen C, Zhou M, et al. A Machine-learning Approach to Forecast Aggravation Risk in Patients with Acute Exacerbation of Chronic Obstructive Pulmonary Disease with Clinical Indicators. *Sci Rep* 2020; 10: 3118. DOI: [10.1038/s41598-020-60042-1](https://doi.org/10.1038/s41598-020-60042-1).