



Research Paper

A new classifier constructed with platelet features for malignant and benign pulmonary nodules based on prospective real-world data

Ruiling Zu^{1*}, Lin Wu^{2*}, Rong Zhou^{3*}, Xiaoxia wen^{4*}, Bangrong Cao^{5*}, Shan Liu⁴, Guishu Yang¹, Ping Leng⁴, Yan Li⁴, Li Zhang¹, Xiaoyu Song¹, Yao Deng¹, Kaijiong Zhang¹, Chang Liu¹, Yuping Li¹, Jian Huang⁷, Dongsheng Wang¹, Guiquan zhu⁶, Huaichao Luo^{1,7}

1. Department of clinical laboratory, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
2. Chiang Mai University College of International College of Digital Innovation, Chiang Mai, Thailand
3. Department of Clinical Laboratory, Sichuan Academy of Medical Sciences & Sichuan provincial People's Hospital, Chengdu, China
4. Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, China
5. Radiation Oncology Key Laboratory of Sichuan Province, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China
6. State Key Laboratory of Oral Diseases, National Clinical Research Centre for Oral Diseases, Department of Head and Neck Oncology, West China Hospital of Stomatology, Sichuan University, Chengdu, China
7. Center for Informational Biology, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

*Contributed equally

 Corresponding authors: Dongsheng Wang, Department of Clinical Laboratory, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, No 55, Section 4, South People's Road, Chengdu 610041, Sichuan, People's Republic of China; Tel: + 862885420344; Email: wangdongshengnc@163.com; Guiquan zhu, Department of Head and Neck Oncology, West China Hospital of Stomatology, Sichuan University, No. 14, Section 3, Renmin South Road, 610041 Chengdu, Sichuan, PRChina. Email: zhugq@scu.edu.cn; ORCID ID: <https://orcid.org/0000-0002-8316-5809>; Huaichao Luo, Department of Clinical Laboratory, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, No 55, Section 4, South People's Road, Chengdu 610041, Sichuan, People's Republic of China; Tel: + 862885420344; Email: luo1987cc@163.com.

© The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>). See <http://ivyspring.com/terms> for full terms and conditions.

Received: 2021.09.24; Accepted: 2022.04.20; Published: 2022.05.09

Abstract

Objectives: As the pulmonary nodules were hard to be discriminated as benignancy or malignancy only based on imageology, a prospective and observational real-world research was devoted to develop and validate a predictive model for managing the diagnostic challenge.

Methods: This study started in 2018, and a predictive model was constructed using eXtreme Gradient Boosting (XGBoost) based on computed tomographic, clinical, and platelet data of all the eligible patients. And the model was evaluated and compared with other common models using ROC curves, continuous net reclassification improvement (NRI), integrated discrimination improvement (IDI), and net benefit (NB). Subsequently, the model was validated in an external cohort.

Results: The development group included 419 participants, while there were 62 participants in the external validation cohort. The most accurate XGBoost model called SCHC model including age, platelet counts in platelet rich plasma samples (pPLT), plateletcrit in platelet rich plasma samples (pPCT), nodule size, and plateletcrit in whole blood samples (bPCT). In the development group, the SCHC model performed well in whole group and subgroups. Compared with VA, MC, BU model, the SCHC model had a significant improvement in reclassification as assessed by the NRI and IDI, and could bring the patients more benefits. For the external validation, the model performed not as well. The algorithm of SCHC, VA, MC, and BU model were first integrated using a web tool (<http://i.uestc.edu.cn/SCHC>).

Conclusions: In this study, a platelet feature-based model could facilitate the discrimination of early-stage malignancy from benignancy patients, to ensure accurate diagnosis and optimal management. This research also indicated that common laboratory results also had the potential in diagnosing cancers.

Key words: lung cancer; pulmonary nodules; platelets; diagnosis; XGBoost; clinical laboratory

Introduction

All over the world, lung cancer has become the most frequent cancer and cause of cancer death. According to the report of Chinese cancer registration in China, there were 733.3 thousands of new lung cancer cases in 2015, which exceeded the number in 2007. The reports also revealed that the 5-year survival of lung cancer in 2012-2015 was 23.8% and in 2009-2011 was 20.3% [1, 2]. Low-dose computed tomography (LDCT) has been accepted as an effective screening method in high-risk individuals to reduce lung cancer mortality [3, 4]. The limits of LDCT screening included the inappropriate detection of lung nodules, for their high possibilities of being early lung cancers. Pathologically, those equivocal nodules could be only confirmed by invasive biopsy or postoperative biopsy, while a part of nodules were determined to be benign [3, 4]. Therefore, some researchers were devoted to construct clinical models or liquid biopsies increasing the appropriate detection of LDCT.

The clinical models usually combine LDCT-detected nodule characteristics with clinical data of patients to estimate the probabilities of malignancy in pulmonary nodules, such as VA, MC and BU model [5-8]. Those diagnostic models partly improved the diagnostic value of LDCT, but were still hard to distinguish benign from malignant nodules [9, 10]. An accurate and practical model could indicate the probability that the equivocal nodule was malignant or benign and would reduce the misdiagnosis in screening programs.

Routine blood test (RBT) is a low-cost test including more than dozens of features and revealing the complications of individuals, which can be characterized as an easy and basic liquid biopsy and provides a large data for researches [11]. In our previous research, we identified that platelet features based on RBT had potential to diagnose lung cancer, which meant the platelet features could complement LDCT to distinguish benign nodules from malignant nodules [12]. In this research, we collected real-world data to construct a diagnostic model combining platelet features with clinical and LDCT information, which could be used to discriminate malignancy from benign nodules, and to expand the treatment options. The model was also presented firstly as a web tool.

2. Material and Methods

This prospective, observational and multi-center real-world study was started from 2018 until now in Sichuan Cancer Hospital and Sichuan Provincial People's Hospital. The data of development group were collected consecutively between January 2018

and December 2020 in Sichuan Cancer Hospital. The new external validation data were collected between March 2021 and April 2021 in Sichuan Provincial People's Hospital. The study was approved by the medical ethical committee of Sichuan Cancer Hospital (SCCHEC-02-2018-031, SCCHEC-02-2020-043), and Sichuan Provincial People's Hospital (2021-NO.404).

2.1 Study Design

The work-flow chart of this study was shown in Figure 1. The individuals consecutively diagnosed with a first occurrence of pulmonary nodules were enrolled in this study. Patients should be observed for the duration of the diagnosis and therapy line (up to the day the patient discharged from hospital). The blood samples were collected and determined at the same time participants admitted hospital [12]. The pathological examination of resected tissue was finished after surgery or biopsy. The clinical information (age, gender, smoking history, anamnesis, family history, personal history), LDCT information (nodule size, number, characteristic, position), pathology information (pathological type, pathological stage), and laboratory information (platelet features in whole blood samples and platelet rich plasma samples) were captured and summarized.

In the development group, seventy percent of all participants were enrolled in the training cohort, while thirty percent of participants were enrolled in the testing cohort randomly. The diagnostic model achieved in training cohort, and validated in the testing and external cohort. Then the model was compared with other common models in the whole development group.

2.2 Study Participants

The development and external validation of the model based on the data of patients as follows. All the patients whose pulmonary nodules were first found in LDCT, and accepted the surgery treatment or histologic examination in the two hospitals were enrolled. According to the pathological results, the patients were divided into two groups: the benign group (patients with benign nodules) and the malignant group (patients with malignant nodules). The histological type and TNM staging were based on the guidelines of Chinese society of clinical oncology (2019 version) [34]. All the LDCT information, pathology information, and laboratory information were also recorded.

And the exclusion criteria were as followed: 1) patients who were not confirmed by the surgery or biopsy; 2) patients with other cancers and acute inflammation; 3) patients who had received the medicine influencing the platelets such as aspirin; 4)

patients who received platelet or blood transfusion in half of a month; 5) patients without complete clinical and laboratory data.

2.3 Laboratory Measures

In Sichuan Cancer Hospital, the pulmonary nodules were detected by SIEMENS Definition Flash CT (SIEMENS Healthineers Co., Ltd, Erlangen, GER) and PHILIPS Brilliance iCT (Koninklijke Philips N.V., Amsterdam, NED). As previously, the fresh whole blood samples were collected to detect platelet features and separate the platelet rich plasma (PRP) from nucleated blood cells by a 20-minute 120×g centrifugation step (Co., Ltd. Shuke, Chengdu, Sichuan, CHN). Meanwhile, the platelet features in the PRP samples were detected. The platelet features

in whole blood samples and PRP samples were detected with a Mindray cell counter BC-6600 (Mindray Bio-Medical Electronics Co., Ltd, Shenzhen, CHN) [12]. The pathological examination of resected tissue was completed in pathology department of Sichuan Cancer Hospital.

In Sichuan Provincial People’s Hospital, the SOMATOM Force CT (SIEMENS Healthineers Co., Ltd, Erlangen, GER) was used to detect pulmonary nodules. The platelet features were detected with automatic hemacytology analyzer (Sysmex, Shanghai, Ltd, Shanghai, CHN). The pathological examination was completed in pathology department of Sichuan Provincial People’s Hospital.

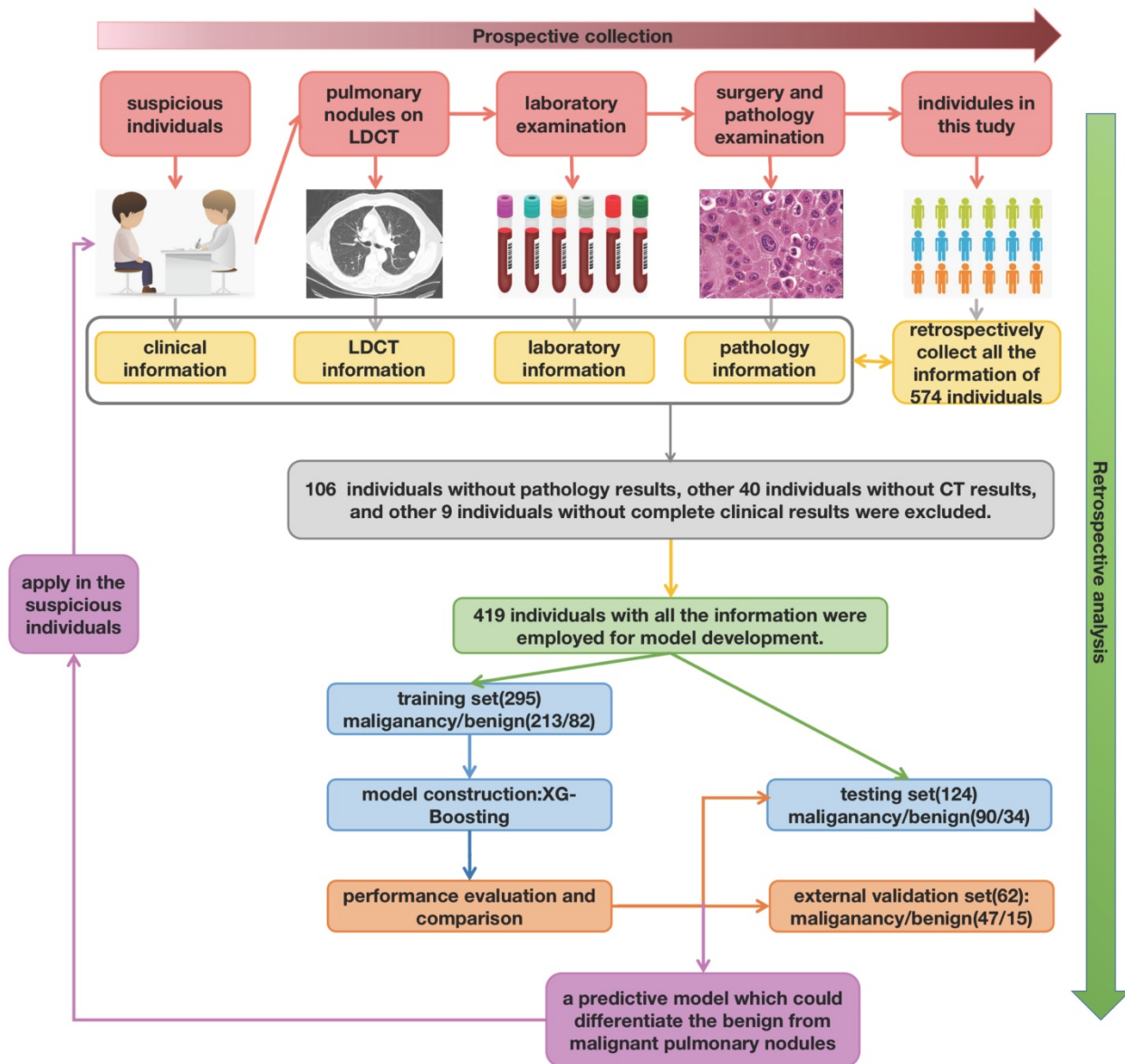


Figure 1. The work-flow chart of this study. Primarily, there were total of 574 individuals enrolled in this research. After retrospectively analyzed, 106 individuals without pathology results, other 40 individuals without CT results, and other 9 without complete clinical information were excluded.

To make sure the reliability of all the measures, all the measurements were finished according to the standard operation procedure (SOP). The internal quality control (IQC) for LDCT, automatic hemacytology analyzer, and pathology examination needed to be acceptable. And the results of personnel parallel experiments for all the departments also needed to be acceptable. Moreover, the external quality assessment (EQA) results of clinical laboratory and pathology department needed to be 91 to 100 points from 2018 to 2021.

2.4 Model Construction

A prediction model was constructed using eXtreme Gradient Boosting (XGBoost). XGBoost is a machine learning technique, which generates a series of decision trees in a gradient boosting manner. In this research, a total of 24 features were entered into multi-tree XGBoost, including gender, smoking status, nodule characteristics on LDCT scan, anamnesis, personal history, family history, and platelet features. And all features were ranked by this model based on the values of their importance. Subsequently, the top 5 features were added to the XGBoost model. After adjusting the features, the final model was created with the importance scores re-calculating.

2.5 Model Performance and Validation

The receiver operating characteristic (ROC) curve was used to estimate the performance of the model, which reflected the sensitivity (Sens), specificity (Spec), accuracy (ACC), and the area under the curve (AUC) in the plot. Prediction model performance was compared with other clinical models by using measures of reclassification (NRI and IDI) [13]. NRI was referred to the movement of patients from benign nodules to malignant nodules based on different models. IDI was calculated to measure the discrimination, which referred to the ability of a model correctly distinguishing between benign and malignant group. In this research, net benefit (NB), assessed the increase in the proportion of patients with pulmonary nodules receiving the appropriate surgery, after estimation by the predictive models [14]. The net benefit of each model over a whole range of threshold probabilities of outcome was graphically displayed as a decision curve.

2.6 Statistical Analysis

All analyses were performed by using the SPSS 22.0 software (SPSS Inc., Chicago, IL) and R 4.0.4 (Version 1.74) (R Foundation for Statistical Computing, <http://www.R-project.org>). The data from malignant patients were compared to the benign

patients using the Wilcoxon test. Thus, $p < 0.05$ were considered significant with two sided, and values were indicated in medians. The construction, performance and validation of the model were implemented under R either.

3. Results

3.1 The demographics and characteristics of participants

After excluding the patients who were lack of pathology results or leaving hospital without surgery, a total of 419 patients with pulmonary nodules were enrolled in the development group finally, which included 295 in training cohort and 124 in testing cohort respectively. The basic clinical information and CT information were summarized in Table 1. The age, gender and pathological stage were adjusted between two cohorts.

Table 1. The demographics and characteristics of all participants

	Traning			Testing		
	benign	malignant	p.overall	benign	malignant	p.overall
gender:	82 (27.8%)	213 (72.2%)		34 (27.4%)	90 (72.6%)	
Female	32 (39.0%)	102 (47.9%)	0.215	11 (32.4%)	46 (51.1%)	0.095
Male	50 (61.0%)	111 (52.1%)		23 (67.6%)	44 (48.9%)	
up:			0.096			0.746
down	41 (50.0%)	82 (38.5%)		15 (44.1%)	35 (38.9%)	
up	41 (50.0%)	131 (61.5%)		19 (55.9%)	55 (61.1%)	
GGN:			0.276			0.287
GGN	10 (12.2%)	39 (18.3%)		1 (2.9%)	10 (11.1%)	
non-GGN	72 (87.8%)	174 (81.7%)		33 (97.1%)	80 (88.9%)	
glitch:			0.024			0.109
glitch	14 (17.1%)	66 (31.0%)		6 (17.6%)	31 (34.4%)	
non-glitch	68 (82.9%)	147 (69.0%)		28 (82.4%)	59 (65.6%)	
age:			0.091			0.014
<=60	52 (63.4%)	110 (51.6%)		26 (76.5%)	45 (50.0%)	
>60	30 (36.6%)	103 (48.4%)		8 (23.5%)	45 (50.0%)	
stage:			<0.001			<0.001
benign	82 (100%)	0 (0.00%)		34 (100%)	0 (0.00%)	
I-II	0 (0.00%)	135 (63.4%)		0 (0.00%)	57 (63.3%)	
III-IV	0 (0.00%)	59 (27.7%)		0 (0.00%)	25 (27.8%)	
NS	0 (0.00%)	19 (8.92%)		0 (0.00%)	8 (8.89%)	
smoking:			0.888			1.000
Ever/current	29 (35.4%)	79 (37.1%)		13 (38.2%)	35 (38.9%)	
Never	53 (64.6%)	134 (62.9%)		21 (61.8%)	55 (61.1%)	
size:			1.000			0.620
<=3cm	46 (56.1%)	118 (55.4%)		22 (64.7%)	52 (57.8%)	
>3cm	36 (43.9%)	95 (44.6%)		12 (35.3%)	38 (42.2%)	

3.2 Construction of the SCHC model

The importance of all 24 features was ranked by XGBoosting model (Figure 2A). Subsequently, the adjusted top 5 features were added to the final model called SCHC model, including age, platelet counts in PRP sample (pPLT), plateletcrit in PRP sample (pPCT), nodule size, and plateletcrit in whole blood sample (bPCT) (Figure 2B).

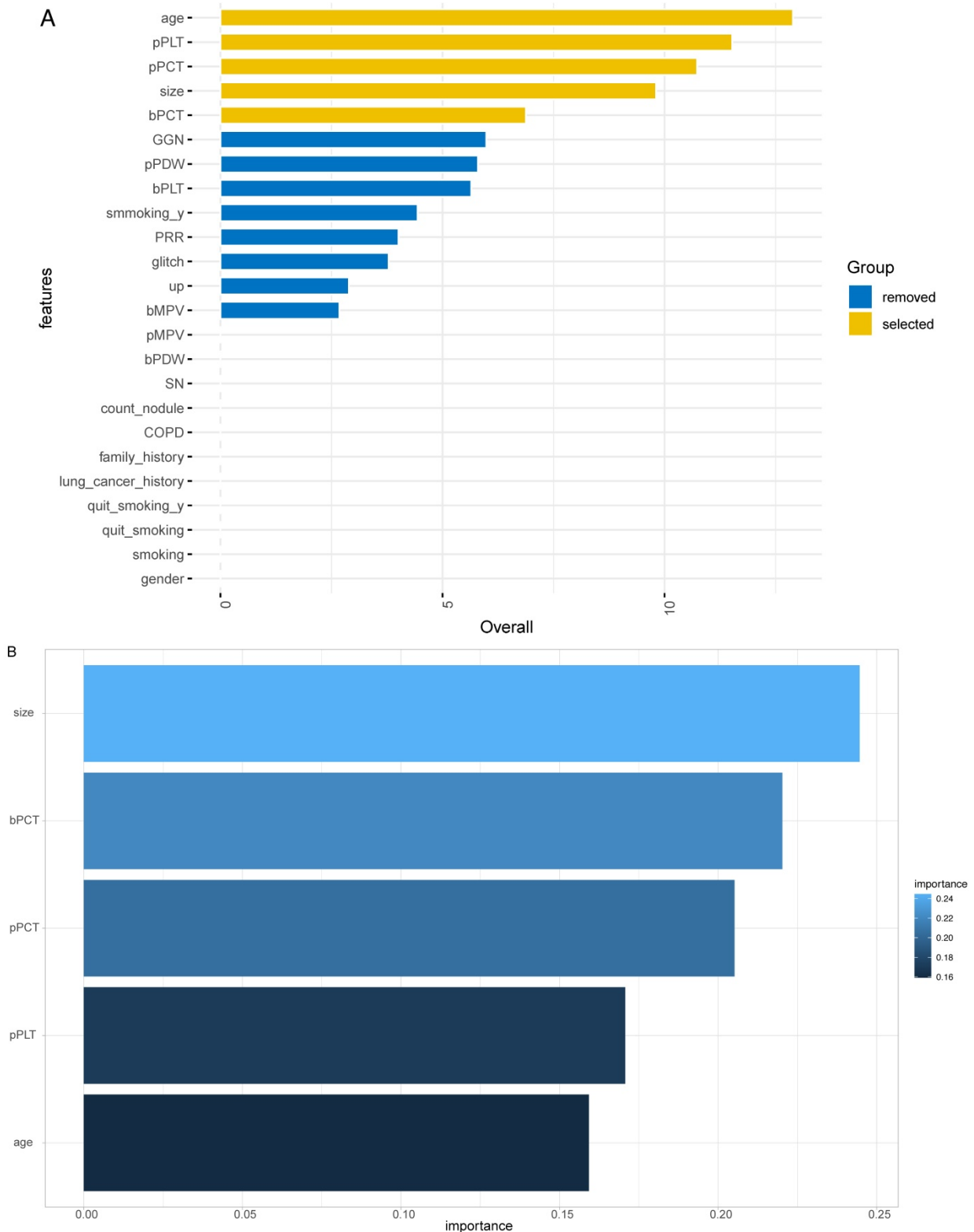


Figure 2. ALL the feature importance scores for the XGBoost model (A). Feature importance scores for the final model (B). Size: the largest diameter of the nodules in millimeter detected by LDCT; bPLT: platelet counts in whole blood sample; bMPV: mean platelet volume in whole blood sample; bPDW: platelet distribution width in whole blood sample; bPCT: plateletcrit in whole blood sample; pPLT: platelet counts in PRP samples; pMPV: mean platelet volume in PRP samples; pPDW: platelet distribution width in PRP sample; pPCT: plateletcrit in PRP sample; GGN: the nodule is ground glass/nonsolid; smoking_y: the years of smoking; glitch: the edge of nodule has spicules; up: the nodule is located in an upper lobe; SN: the nodule is solid; COPD: the patient has a history of COPD; family_history: the patient has a family history of lung cancer; lung_cancer_history: the patient has a history of extrathoracic cancer that was diagnosed 5 years ago; quit_smoking_y: years since the patient has been quitting smoking; quit_smoking: whether the patient has quit smoking; smoking: whether the patient has smoked.

3.3 Performance and internal validation of the SCHC model

The performance of SCHC model in two cohorts was presented in Figure 3. In two cohorts, SCHC model exhibited a good performance (AUC: 0.76 and 0.72, respectively). The predicted probabilities of malignant set were higher than the benign set in the two cohorts ($p < 0.001$), which indicated the model could differentiate the malignant nodules from the benign nodules.

ROC-AUCs, specificities and sensitivities were calculated for different-size nodules, when the AUC was maximized, or the sensitivity was held at a performance of 90 % (Figure 4). SCHC model

performed well on the individuals who were older than 60 with nodules in the 20-30 mm range, achieving 87.5% specificity and 81.8% sensitivity with highest AUC. The SCHC model also performed well on male with same size nodules. And the performance on the LSCC patients with nodules in the 10-20 mm range was particularly well (AUC=0.856, with the 81.8% specificity and 85.4% sensitivity). As the small size nodules were hard to distinguish, when the sensitivity was held at a performance of 90 %, the specificity of SCCH model were at the range from 9.8% to 51.2%. In the clinical applications, the thresholds could be chosen from the different goals in different individuals.

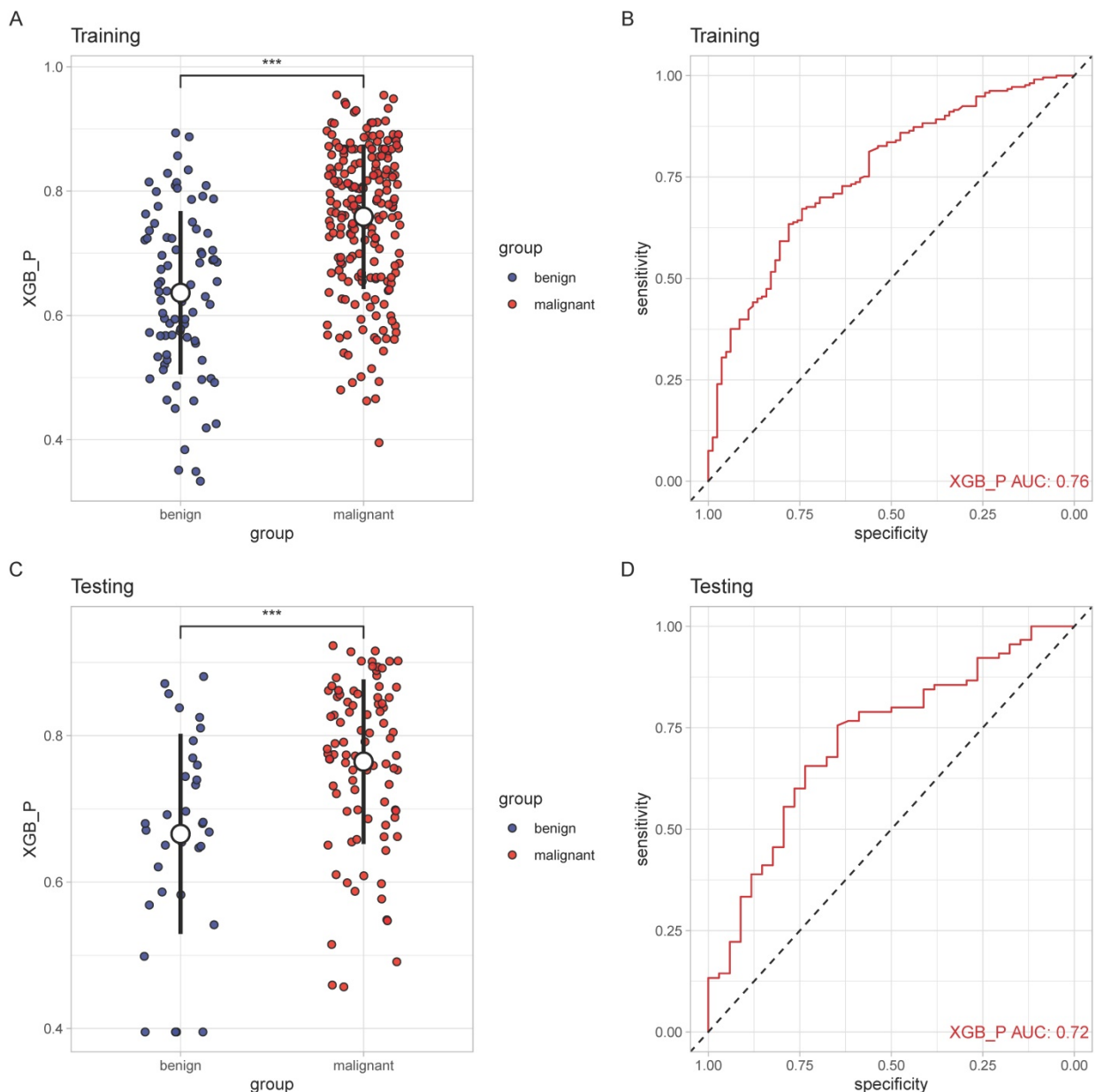


Figure 3. The performance of SCHC model in training and testing cohort. Scatter plots indicating the predictive probabilities calculated by SCHC from the benign and malignant groups of training cohort(A) and testing cohort(C). Receiver operating characteristic curves for performance of SCHC model in training cohort (B) and testing cohort(D). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

	cut-off max		specificity at 90% sensitivity			
	age		age			
	<=60	>60	<=60	>60		
1-10mm	AUC=0.718 (CI:0.592-0.844) SE=0.667,SP=0.677	AUC=0.721 (CI:0.496-0.947) SE=0.643,SP=0.800	AUC=0.648 (CI:0.548-0.748) SE=0.909,SP=0.387	AUC=0.664 (CI:0.490-0.839) SE=0.929,SP=0.400		
10-20mm	AUC=0.709 (CI:0.596-0.821) SE=0.596,SP=0.778	AUC=0.662 (CI:0.502-0.822) SE=0.678,SP=0.714	AUC=0.586 (CI:0.494-0.678) SE=0.912,SP=0.259	AUC=0.529 (CI:0.427-0.631) SE=0.915,SP=0.143		
20-30mm	AUC=0.724 (CI:0.515-0.933) SE=0.909,SP=0.556	AUC=0.813 (CI:0.623-1.000) SE=0.818,SP=0.875	AUC=0.732 (CI:0.553-0.912) SE=0.909,SP=0.556	AUC=0.580 (CI:0.408-0.751) SE=0.909,SP=0.250		
30-60mm	AUC=0.696 (CI:0.516-0.876) SE=0.469,SP=0.909	AUC=0.657 (CI:0.367-0.948) SE=0.811,SP=0.667	AUC=0.635 (CI:0.477-0.793) SE=0.906,SP=0.364	AUC=0.619 (CI:0.409-0.830) SE=0.906,SP=0.333		
	gender		gender			
	female	male	female	male		
	1-10mm	AUC=0.692 (CI:0.541-0.843) SE=0.800,SP=0.524	AUC=0.753 (CI:0.593-0.913) SE=0.765,SP=0.700	AUC=0.633 (CI:0.521-0.746) SE=0.933,SP=0.333	AUC=0.696 (CI:0.570-0.821) SE=0.941,SP=0.450	
10-20mm	AUC=0.758 (CI:0.619-0.896) SE=0.852,SP=0.571	AUC=0.674 (CI:0.553-0.795) SE=0.673,SP=0.741	AUC=0.629 (CI:0.494-0.765) SE=0.902,SP=0.357	AUC=0.529 (CI:0.450-0.607) SE=0.909,SP=0.148		
20-30mm	AUC=0.600 (CI:0.227-0.973) SE=0.650,SP=0.750	AUC=0.800 (CI:0.659-0.941) SE=0.714,SP=0.846	AUC=0.600 (CI:0.350-0.850) SE=0.950,SP=0.250	AUC=0.649 (CI:0.504-0.795) SE=0.914,SP=0.385		
30-60mm	AUC=0.892 (CI:0.781-1.000) SE=0.784,SP=1.000	AUC=0.654 (CI:0.470-0.837) SE=0.604,SP=0.769	AUC=0.709 (CI:0.423-0.996) SE=0.919,SP=0.500	AUC=0.612 (CI:0.476-0.749) SE=0.917,SP=0.308		
	stage		stage			
	I-II	III-IV	I-II	III-IV		
	1-10mm	AUC=0.739 (CI:0.629-0.849) SE=0.889,SP=0.512	AUC=0.739 (CI:0.513-0.964) SE=0.857,SP=0.683	AUC=0.690 (CI:0.600-0.780) SE=0.917,SP=0.463	AUC=0.549 (CI:0.503-0.595) SE=1.000,SP=0.098	
10-20mm	AUC=0.664 (CI:0.563-0.765) SE=0.608,SP=0.756	AUC=0.780 (CI:0.660-0.899) SE=0.731,SP=0.756	AUC=0.541 (CI:0.475-0.607) SE=0.911,SP=0.171	AUC=0.608 (CI:0.520-0.696) SE=0.923,SP=0.293		
20-30mm	AUC=0.741 (CI:0.598-0.884) SE=0.703,SP=0.824	AUC=0.761 (CI:0.588-0.933) SE=0.714,SP=0.765	AUC=0.607 (CI:0.486-0.727) SE=0.919,SP=0.294	AUC=0.641 (CI:0.504-0.777) SE=0.929,SP=0.353		
30-60mm	AUC=0.706 (CI:0.559-0.853) SE=0.575,SP=0.824	AUC=0.723 (CI:0.569-0.723) SE=0.676,SP=0.824	AUC=0.563 (CI:0.464-0.663) SE=0.950,SP=0.176	AUC=0.636 (CI:0.511-0.761) SE=0.919,SP=0.353		
	pathology			pathology		
	LACC	LSCC	other	LACC	LSCC	other
	1-10mm	AUC=0.711 (CI:0.598-0.824) SE=0.821,SP=0.537	AUC=0.543 (CI:0.000-1.000) SE=0.500,SP=0.927	AUC=0.829 (CI:0.665-0.993) SE=0.667,SP=0.902	AUC=0.657 (CI:0.570-0.743) SE=0.923,SP=0.390	AUC=0.549 (CI:0.503-0.595) SE=1.000,SP=0.098
10-20mm	AUC=0.691 (CI:0.597-0.785) SE=0.629,SP=0.756	AUC=0.856 (CI:0.728-0.983) SE=0.818,SP=0.854	AUC=0.476 (CI:0.211-0.740) SE=0.500,SP=0.707	AUC=0.563 (CI:0.493-0.634) SE=0.907,SP=0.220	AUC=0.833 (CI:0.721-0.944) SE=0.909,SP=0.756	AUC=0.500 (CI:0.500-0.500) SE=1.000,SP=0.000
20-30mm	AUC=0.724 (CI:0.573-0.874) SE=0.675,SP=0.824	AUC=0.752 (CI:0.543-0.961) SE=0.667,SP=0.824	AUC=0.892 (CI:0.752-1.000) SE=0.833,SP=0.824	AUC=0.639 (CI:0.515-0.763) SE=0.925,SP=0.353	AUC=0.647 (CI:0.534-0.759) SE=1.000,SP=0.294	AUC=0.824 (CI:0.706-0.941) SE=1.000,SP=0.647
30-60mm	AUC=0.730 (CI:0.592-0.868) SE=0.685,SP=0.824	AUC=0.748 (CI:0.590-0.906) SE=0.571,SP=0.824	AUC=0.709 (CI:0.489-0.929) SE=0.700,SP=0.824	AUC=0.601 (CI:0.483-0.719) SE=0.907,SP=0.294	AUC=0.658 (CI:0.522-0.795) SE=0.905,SP=0.412	AUC=0.559 (CI:0.480-0.638) SE=1.000,SP=0.118

Figure 4. Classification performance in different subgroup patients across different nodule size ranges. The ROC-AUC, specificity and sensitivity (AUC at max, yellow, left; at 90% sensitivity, green, right) were given for the different age, gender, stage and pathology. The vertical axis was different range of nodule size in millimeter. The color intensity was related to the AUC values (AUC at max, yellow) and specificity (at 90% sensitivity, green). SE: sensitivity; SP: specificity.

3.4 Comparison among SCHC model with other models

SCHC model was performed and compared with other clinical models on all participants in development group (Figure 5). Alluvial diagrams indicating the misclassification for all the models were plotted in Figure 5B and Figure 5D. SCHC model

misclassified approximately 100 of malignant individuals as benign individuals, which was same as MC and BU models. But the VA model misclassified almost 200 of malignant individuals as benign individuals, whose misdiagnosis rate was higher than the other three models. Instead, the miss rate of VA model was lower than the other three models. SCHC model misclassified approximately 20 of malignant

individuals as benign individuals, while the MC and BU model misclassified about 30 of malignant individuals as benign individuals. SCHC model could be seen as the most extreme model in misclassifying the malignant and benign nodules. Compared with the MC model, SCHC had significant improvement in reclassification as assessed by the IDI (0.1274, 95%CI: 0.0909 - 0.1638), and NRI (0.2053, 95%CI: 0.1177 - 0.2928). SCHC also showed a better reclassification ability than VA and BU model with significant NRI (0.1906, 95%CI: 0.0955 - 0.2857, 0.2053; 95%CI: 0.1177 - 0.2928, respectively) and IDI (0.1371, 95%CI: 0.0971 - 0.1772; 0.1235, 95%CI: 0.0883 - 0.1586, respectively) (Figure 5E). The decision curves plotted the net benefits achieved by making decisions based on the four models (Figure 5F). The decision curve displayed that SCHC model with the highest net benefit at all threshold probabilities, which meant SCHC model had the highest clinical value. Thus, no matter what probability threshold was chosen by clinicians, the SCHC model could bring the patients more benefits than the other three clinical models.

3.5 External Validation

Based on the results above, SCHC model was further assessed in external validation cohort. 15 benign and 47 malignant participants were included in the external validation cohort. The basic information was presented in Table S1. The accuracy of SCHC model presented limited in the external cohort with an AUC of 0.52 (95% CI, 0.354 to 0.684), which was similar with VA model (AUC=0.47, 95% CI: 0.326 to 0.651), while the MC and BU model outperformed with the AUC of 0.68 (95% CI, 0.520 to 0.833) and 0.65 (95% CI, 0.479 to 0.813) respectively

(Figure 5G).

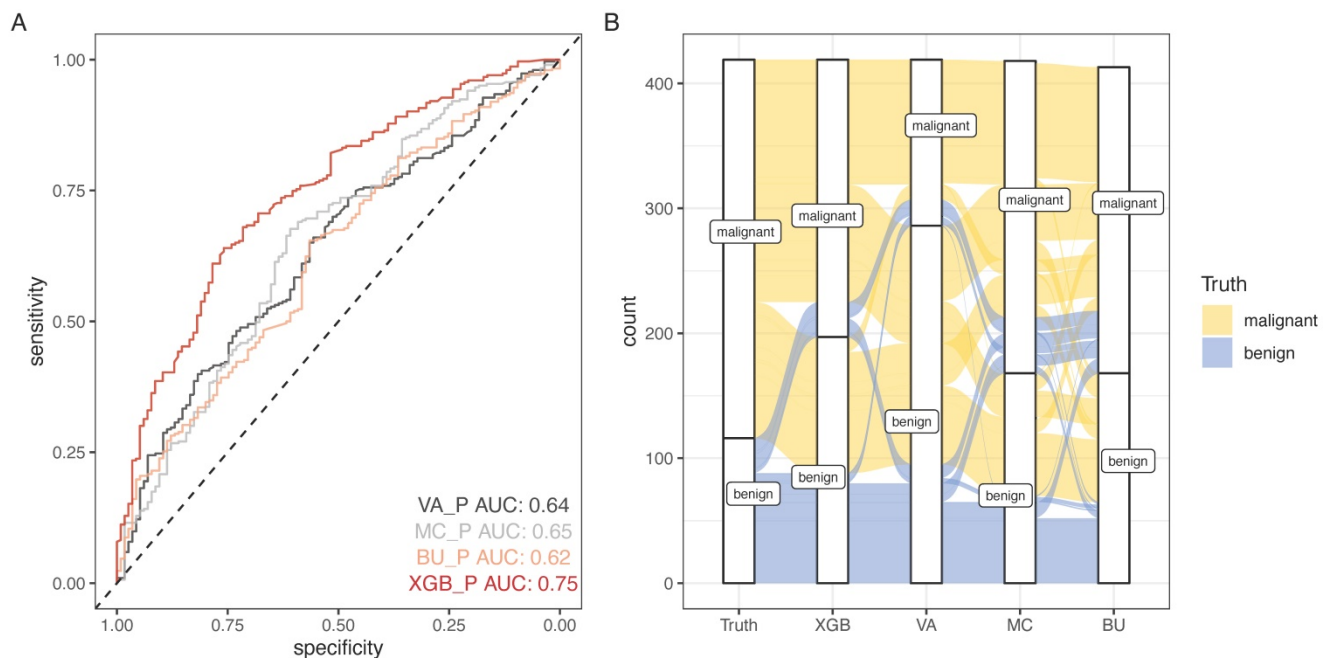
3.6 Web tool based on SCHC model

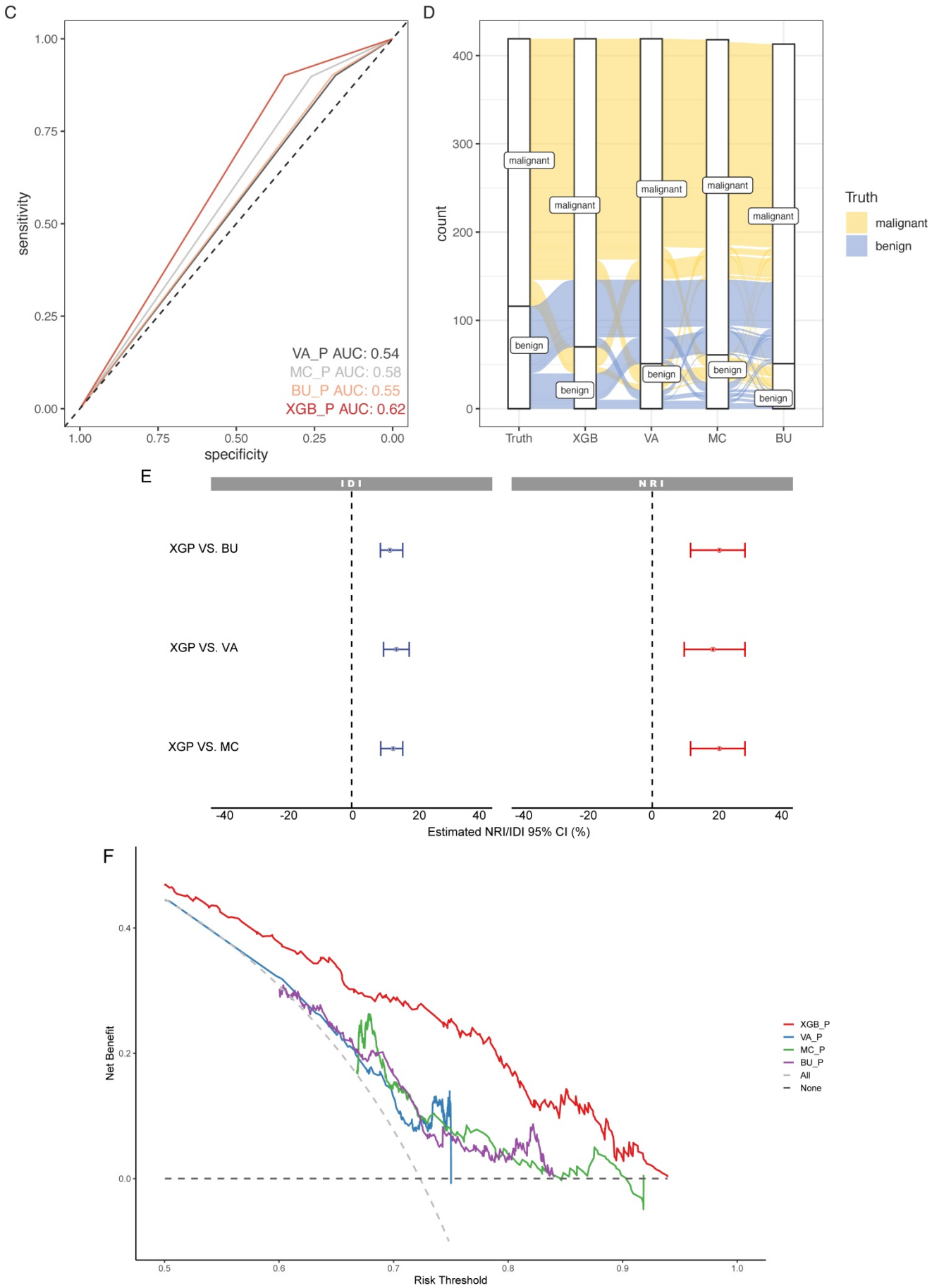
Accessing <http://i.uestc.edu.cn/SCHC> in a web browser displayed directly the main page of SCHC model, where Age, pPLT, pPCT, size, bPCT could be input. By clicking the "Get-Result" button, the probability of malignancy would display with the blue bar, and "malignant" would output if the probability exceeded the threshold. As shown in Fig. 6, users would be allowed to choose SCCH, BU, MC and VA model at the top of the Web page. The web tool would also be updated in the future work along with our follow-up work.

4. Discussion

This prospective and observational real-world study provided a reliable model, which could distinguish the benign nodules from malignant nodules based on platelet features, age and nodules size. Compared with other common models, the incremental improvements in AUC, NRI, IDI and NB indicated an accurate diagnostic value of our SCHC model.

As the lung cancer has become an important public health problem, the LDCT was a widely used screening test for the suspicious pulmonary nodules. A large part of the suspicious pulmonary nodules was proved to be false positive, which might associate with an unnecessary surgery. In order to help clinicians improve the predictive accuracy for the pulmonary nodules, a number of clinical models were published.





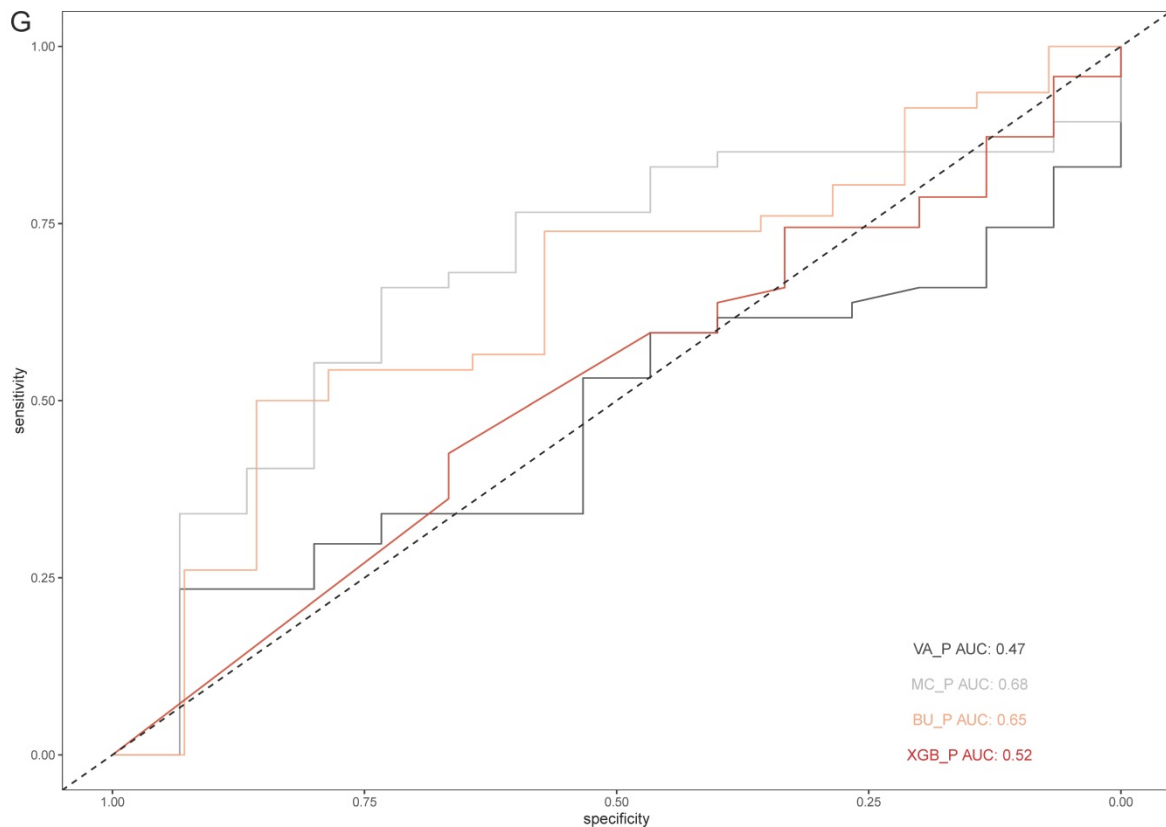


Figure 5. The performance of SCHC model in development cohort compared with other three models. Receiver operating characteristic curves for performance of the model, when the AUC was held at a performance of the max (A), and the sensitivity was held at a performance of 90 % (C). Alluvial diagrams indicating the misclassification for all the models, when the AUC was held at a performance of the max (B), and the sensitivity was held at a performance of 90 % (D). VA_P: the probabilities calculated using VA model; MC_P: the probabilities calculated using MC model; BU_P: the probabilities calculated using BU model; XGB_P: the probabilities calculated using SCHC model. Integrated discrimination improvement for discrimination (IDI), net reclassification improvement (NRI) in the development cohort for SCHC model and other three models (E). Decision curve shown for all models across all threshold probabilities (F). Receiver operating characteristic curves for performance of the models in validation cohort (G).

The clinical models, like VA model, MC model and BU model, were widely cited in previous researches. The VA model was developed with the data from Department of Veterans Affairs (VA), which calculated the probabilities of malignancy in patients with solid pulmonary nodules (SPNs) using the patients' smoking history, age and nodule size. The accuracy of VA model presented well in the internal cohort with an AUC of the ROC curve of 0.79 (95% CI, 0.74 to 0.84), but not well in the external researches [15, 16], which was same in this research. The MC model was proposed by Swensen [7] at the Mayo Clinic, which expressed the probabilities of malignant nodules as a function of the variables including age, smoking history, cancer history and characteristics of the nodules. But the MC model performed not much stable in other validation researches, with the AUC range from 0.60 to 0.89 [10, 17-19]. The BU model also named Brock model was developed from the Pan-Canadian Early Detection of Lung Cancer Study, which included two sets of models: a parsimonious model and a full model. The BU model calculated the probabilities using multivariable logistic regression based on the

variables including age, gender, family history of lung cancer, presence of COPD, nodule size, nodule location, nodule counts, and nodule characteristics. In the other external validation researches, the BU model performed excellent discrimination, with an AUC greater than 0.80 [18-21]. Most of the clinical models were retrospective researches and set up with malignant and healthy individuals, so the three models showed limited values in this research. And SCHC model was set up with malignant and benign nodule patients, so it performed better than the other models in our research.

Except for the clinical models, the liquid biopsy has also been hot in the researches of diagnosing malignancy. Especially in lung cancer, multiple biomarkers like DNA methylation and serum microRNAs have been proved great diagnostic values with high sensitivities and specificities [15, 22]. But the detection of DNA methylation or serum microRNAs is complicated and costly, which is not suitable for the widely screening. Moreover, the variables of our model including age, platelet features and nodule size are all easy to estimate. The age and nodule size have been reported in other models,

which reveals the two factors taking roles in the prediction of malignant nodules. And the platelet features are new factors applied in predicting the malignancy. Platelets are the second most components of blood, which not only play the key role in hemostasis, but also in cancer progression. As the concepts of tumor educated platelets (TEPs) were presented, it have been proven that the tumor cells could activate the platelets, and then the activated platelets release molecules to impact the tumor cells in turn [23-25]. And platelet features are accessible by hematology analyzer, so the detection could be automatic, standard, recurring and low-cost. In our previous research, we set up a multivariate logistic regression model to diagnose lung cancer based on

the platelet features, with the AUC of 0.92 for the training cohort and 0.79 for the testing cohort, which means platelets could be noninvasive biomarkers for a new liquid biopsy of classifying the pulmonary nodules. This research also indicated that common laboratory results also have the potential in diagnosing cancers. The clinical laboratory results include clinical biochemistry, hematology and microbiology results, which offer complexity and a variety of information of patients. So, combination the common laboratory results could be used in diagnosing different diseases, not only cancers, which also provide a supererogation of the clinical laboratory science.

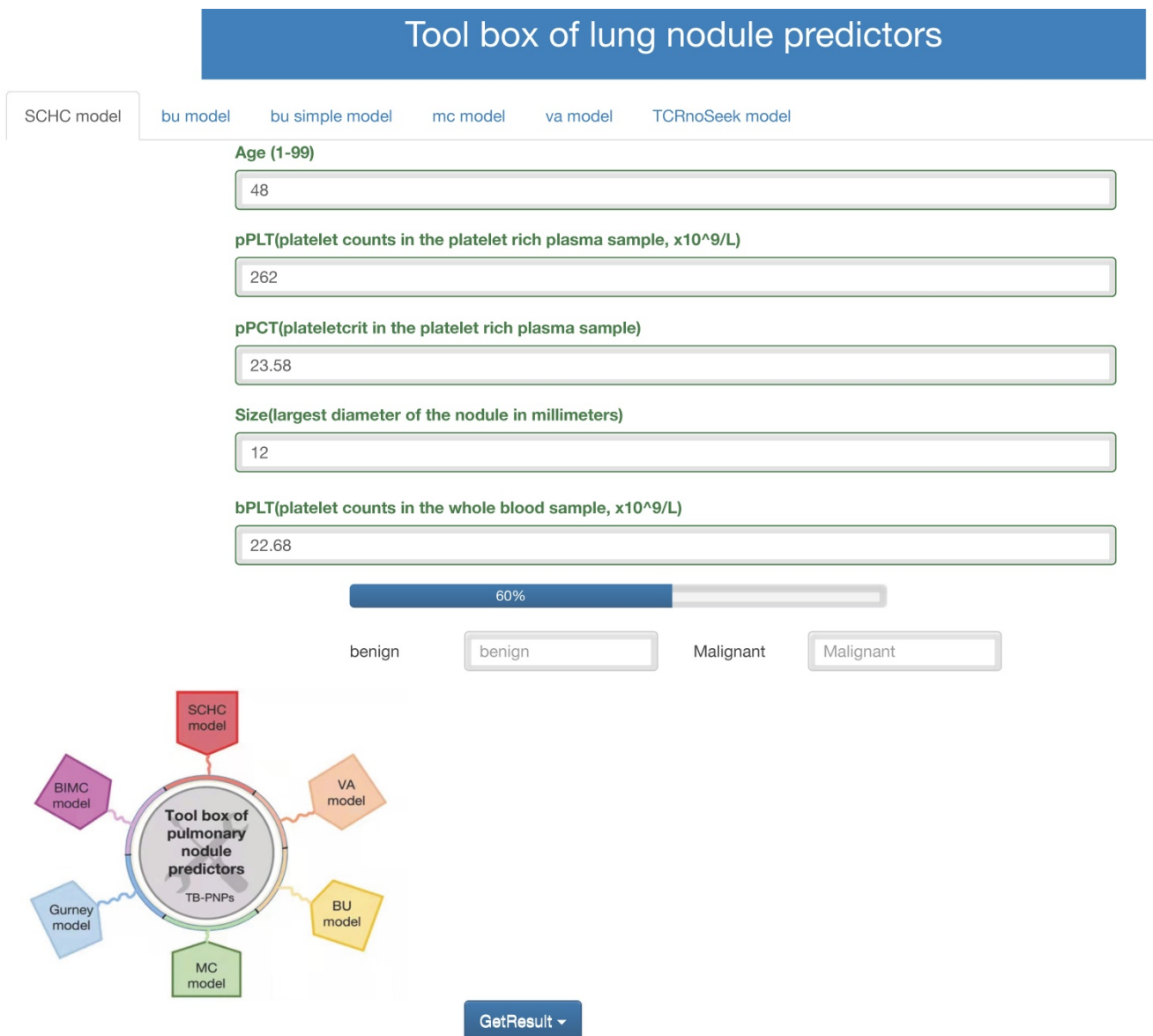


Figure 6. Web tool of the SCCH model for calculating the probability of malignant pulmonary nodules.

This research was a prospective and observational study, which could provide more reliable data to influence the predicted probabilities. Recently, real-world studies (RWS) are gaining increasing attention since the United States Food and Drug Administration (FDA) and the Chinese Food and Drug Administration (CFDA) presented that RWS could be used in etiology, diagnosis, therapy and prognosis [33]. Machine learning has been used in real-world data for numerous disease diagnoses. The XGBoost model is a machine learning method used in many medical prediction models for the high precision prediction [26-29]. As reported in the previous researches, the XGBoost algorithm always achieved more excellent performance compared with other algorithm such as Logistic regression, Bayesian analysis and so on [30-32]. XGBoost could provide the importance score of each variable, and then the top 5 of the variables were selected into the final model. And compared with VA, MC, and BU model, the XGBoost model had best prediction effect, which also could bring more benefits to the patients with pulmonary nodules.

Furthermore, a web page was used to display the SCHC model, and totally free to support both patients and clinicians. The input included the platelet features which were accessible and inexpensive, while the output contained the probability values. This web tool also first integrated the VA, MC, and BU model together, which was easy to calculate the probability of malignancy requiring no programming knowledge. While choosing a strategy for evaluating patients with lung nodules, the clinicians could consider the probability calculated with the web tool. Those nodules with low-risk probabilities, the “benign” outputs could allow the nodules to be surveyed by serial imaging.

There were still some limitations for the SCHC model. In order to validate our model, we further collected an external cohort from Sichuan Provincial People’s Hospital. But the performance of the SCHC model was not good. The SCHC model was constructed with platelet features in PRP samples, which might be influenced by the different centrifuges, laboratory environment, and manipulators. The platelets were easily activated by external force, and the activated platelets were more likely to aggregate, relating to the incorrect results of plateletcrit. This might be the reason for the poor performance in the external validation. The small size of validation data might also contribute to the limitation of the SCHC model. In the future work, we would increase the validation sample size, and construct a simpler and more constant operating manual to minimize the influence in the different

system.

5. Conclusion

We have developed and validated a predictive model with good performance characteristics based on the real-world data for identifying malignant nodules from benign nodules more accurately. And SCHC model used the age, platelet features, and nodule size, which could be easily implemented and would not take additional cost of patients. Thus, SCHC model could be widely applied in the screening of patients with pulmonary nodules, and might reduce the invasive procedures by recognizing benign nodules to surveillance. We developed a user-friendly web tool that integrated SCHC model and common clinical models together making a broadly availability (<http://i.uestc.edu.cn/SCHC>). And this research also provided evidence for the supererogatory uses of common laboratory results.

Supplementary Material

Supplementary methods, figures and tables.
<https://www.jcancer.org/v13p2515s1.pdf>

Acknowledgments

This work was supported by the Health and Family Planning Commission of Sichuan Province universal application project (NO. 19PJ275), and the Science and Technology Bureau of Chengdu technological innovation project (NO. 2019-YF05-01279-SN).

Competing Interests

The authors have declared that no competing interest exists.

References

1. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. *CA Cancer J Clin.* 2016; 66: 115-32.
2. Zeng H, Chen W, Zheng R, Zhang S, Ji JS, Zou X, et al. Changing cancer survival in China during 2003-15: a pooled analysis of 17 population-based cancer registries. *Lancet Glob Health.* 2018; 6: e555-e67.
3. Sawada S, Yamashita N, Sugimoto R, Ueno T, Yamashita M. Long-term Outcomes of Patients With Ground-Glass Opacities Detected Using CT Scanning. *Chest.* 2017; 151: 308-15.
4. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med.* 2020; 382: 503-13.
5. Gould MK, Ananth L, Barnett PG. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest.* 2007; 131: 383-8.
6. McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med.* 2013; 369: 910-9.
7. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Arch Intern Med.* 1997; 157: 849-55.
8. Chung K, Mets OM, Gerke PK, Jacobs C, den Harder AM, Scholten ET, et al. Brock malignancy risk calculator for pulmonary nodules: validation outside a lung cancer screening population. *Thorax.* 2018; 73: 857-63.
9. Soardi GA, Perandini S, Motton M, Montemezzi S. Assessing probability of malignancy in solid solitary pulmonary nodules with a new Bayesian calculator: improving diagnostic accuracy by means of expanded and updated features. *Eur Radiol.* 2015; 25: 155-62.

10. Nair VS, Sundaram V, Desai M, Gould MK. Accuracy of Models to Identify Lung Nodule Cancer Risk in the National Lung Screening Trial. *Am J Respir Crit Care Med*. 2018; 197: 1220-3.
11. Wu J, Zan X, Gao L, Zhao J, Fan J, Shi H, et al. A Machine Learning Method for Identifying Lung Cancer Based on Routine Blood Indices: Qualitative Feasibility Study. *JMIR Med Inform*. 2019; 7: e13476.
12. Zu R, Yu S, Yang G, Ge Y, Wang D, Zhang L, et al. Integration of platelet features in blood and platelet rich plasma for detection of lung cancer. *Clin Chim Acta*. 2020; 509: 43-51.
13. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010; 21: 128-38.
14. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *Bmj*. 2016; 352: i6.
15. Liang W, Chen Z, Li C, Liu J, Tao J, Liu X, et al. Accurate diagnosis of pulmonary nodules using a noninvasive DNA methylation test. *J Clin Invest*. 2021; 131.
16. Li Y, Wang J. A mathematical model for predicting malignancy of solitary pulmonary nodules. *World J Surg*. 2012; 36: 830-5.
17. Soardi GA, Perandini S, Larici AR, Del Ciello A, Rizzardi G, Solazzo A, et al. Multicentre external validation of the BIMC model for solid solitary pulmonary nodule malignancy prediction. *Eur Radiol*. 2017; 27: 1929-33.
18. She Y, Zhao L, Dai C, Ren Y, Jiang G, Xie H, et al. Development and validation of a nomogram to estimate the pretest probability of cancer in Chinese patients with solid solitary pulmonary nodules: A multi-institutional study. *J Surg Oncol*. 2017; 116: 756-62.
19. Al-Ameri A, Malhotra P, Thygesen H, Plant PK, Vaidyanathan S, Karthik S, et al. Risk of malignancy in pulmonary nodules: A validation study of four prediction models. *Lung Cancer*. 2015; 89: 27-30.
20. Winkler Wille MM, van Riel SJ, Saghir Z, Dirksen A, Pedersen JH, Jacobs C, et al. Predictive Accuracy of the PanCan Lung Cancer Risk Prediction Model -External Validation based on CT from the Danish Lung Cancer Screening Trial. *Eur Radiol*. 2015; 25: 3093-9.
21. White CS, Dharaiya E, Campbell E, Boroczky L. The Vancouver Lung Cancer Risk Prediction Model: Assessment by Using a Subset of the National Lung Screening Trial Cohort. *Radiology*. 2017; 283: 264-72.
22. Chen X, Hu Z, Wang W, Ba Y, Ma L, Zhang C, et al. Identification of ten serum microRNAs from a genome-wide serum microRNA expression profile as novel noninvasive biomarkers for nonsmall cell lung cancer diagnosis. *Int J Cancer*. 2012; 130: 1620-8.
23. McAllister SS, Weinberg RA. The tumour-induced systemic environment as a critical regulator of cancer progression and metastasis. *Nat Cell Biol*. 2014; 16: 717-27.
24. Calverley DC, Phang TL, Choudhury QG, Gao B, Oton AB, Weyant MJ, et al. Significant downregulation of platelet gene expression in metastatic lung cancer. *Clin Transl Sci*. 2010; 3: 227-32.
25. Nilsson RJ, Balaj L, Hulleman E, van Rijn S, Pegtel DM, Walraven M, et al. Blood platelets contain tumor-derived RNA biomarkers. *Blood*. 2011; 118: 3680-3.
26. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*. 2019; 19: 211.
27. Liu L, Yu Y, Fei Z, Li M, Wu FX, Li HD, et al. An interpretable boosting model to predict side effects of analgesics for osteoarthritis. *BMC Syst Biol*. 2018; 12: 105.
28. Liu Z, Zhou T, Han X, Lang T, Liu S, Zhang P, et al. Mathematical models of amino acid panel for assisting diagnosis of children acute leukemia. *J Transl Med*. 2019; 17: 38.
29. Zou LS, Erdos MR, Taylor DL, Chines PS, Varshney A, Parker SCJ, et al. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genomics*. 2018; 19: 390.
30. Wang L, Wang X, Chen A, Jin X, Che H. Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. *Healthcare (Basel)*. 2020; 8.
31. Chen T, Li X, Li Y, Xia E, Qin Y, Liang S, et al. Prediction and Risk Stratification of Kidney Outcomes in IgA Nephropathy. *Am J Kidney Dis*. 2019; 74: 300-9.
32. Li W, Yin Y, Quan X, Zhang H. Gene Expression Value Prediction Based on XGBoost Algorithm. *Front Genet*. 2019; 10: 1077.
33. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Center for Biologics Evaluation and Research. Use of Real World Evidence to Support Regulatory Decision Making for Medical Devices. Guidance for industry and FDA staff. Document issued on August 31, 2017.
34. CSCO.Guidelines of Chinese society of clinical oncology-Primary lung cancer(vision 1,2019).Beijing: PMPH, 2019.