·健康与社会行为·

基于三种机器学习方法的慢性阻塞性肺疾病人群早筛模型的建立与验证

母应姣[1]，王子云[1]，苏旭[2]，李凌[2]，周婕[2]，王艺颖[2]，刘涛[1,2]
1. 贵州医科大学公共卫生与健康学院，环境污染与疾病监控教育部重点实验室，贵州 贵阳 561113
2. 贵州省疾病预防控制中心，贵州 贵阳 550004

摘要：目的　构建慢性阻塞性肺疾病（chronic obstructive pulmonary disease，COPD）患者筛检模型。**方法**　采用多阶段分层随机抽样的方法，抽取贵州省≥40 岁的常住居民 4 587 名，对其进行问卷调查、体格检查及肺功能检查。经过单因素分析初步筛选模型纳入变量，经多因素 logistic 回归确定最终纳入变量。分别应用 logistic 回归（logistic regression，LR）、随机森林（random forest，RF）、支持向量机（support vector machine，SVM）构建 COPD 患者筛检模型，使用受试者工作曲线下面积（area under the curve，AUC）评价模型效果。使用 delong 法检验模型之间 AUC 的差异。**结果**　根据多因素 logistic 回归分析结果，本研究将年龄、14 岁前经常咳嗽、哮喘、每日吸烟量（支）、烹饪燃料与排风、有害气体暴露 6 种因素纳入 LR、RF、SVM 模型。三种模型训练集 AUC 分别为 73.64%、87.14%、73.30%，测试集 AUC 分别为 76.10%、70.96%、76.08%，均具有较好的筛检效果。Delong 法结果显示，三种模型的筛检效果在训练集与测试集均存在一定差异。**结论**　本研究通过年龄、哮喘等 6 个简单变量建立经济、快捷且有效的 COPD 患者筛检模型。

关键词：Logistic 回归；随机森林；支持向量机；慢性阻塞性肺疾病；筛检
中图分类号：R563.9；R18　　**文献标志码：**A　　**文章编号：**1003-8507（2024）09-0677-07
DOI：10.20043/j.cnki.MPM.202312018

Establishment and verification of early screening model of chronic obstructive pulmonary disease based on three machine learning methods

MU Ying-jiao*, WANG Zi-yun, SU Xu, LI Ling, ZHOU Jie, WANG Yi-ying, LIU Tao
Key Laboratory of Environmental Pollution and Disease Monitoring, Ministry of Education, School of Public Health and Health, Guizhou Medical University Guiyang, Guiyang 561113, China

Abstract: Objective To establish a screening model for patients with chronic obstructive pulmonary disease (COPD). **Methods** By using the method of multi-stage stratified random sampling, 4 587 permanent residents ≥ 40 years old in Guizhou Province were investigated by questionnaire, physical examination, and pulmonary function examination. Variables to be included into the model were screened by univariate analysis and then further screened by multivariate Logistic regression. Logistic regression (LR), random forest (RF) and support vector machine (SVM) were used to construct the screening model of COPD patients, and the area under the curve (AUC) was used to evaluate the effect of the model. Delong method was used to test the difference of AUC between models. **Results** According to the results of multivariate Logistic regression analysis, age, frequent cough before 14 years old, asthma, daily smoking, cooking fuel and exhaust, and harmful gas exposure were included in LR, RF and SVM models. The AUC of the three model training sets were 73.64%, 87.14%, and 73.30%, respectively, and the AUC of the test set were 76.10%, 70.96%, and 76.08%, respectively, all of which had good screening results. The results of Delong method showed that the screening effects of the three models were different between the training set and the test set. **Conclusion** This study established an economical, rapid, and effective screening model for COPD patients through six simple variables such as age and asthma.

Keywords: Logistic regression; Random forest; Support vector machine; Chronic obstructive pulmonary disease; Screening

慢性阻塞性肺疾病（chronic obstructive pulmonary disease，COPD）是严重危害我国人民健康的呼吸系统疾病之一，《中国死因监测数据集（2021）》表明因 COPD 死亡人数占主要呼吸系统疾病首位[1]。由于 COPD 在早期无明显症状，多数 COPD 患者并未及时得到发现，从而错过最佳治疗时间。目前 COPD 筛查

基金项目：基于医防融合的健康管理中心大数据平台研究与示范（黔科合支撑［2021］一般 447）；贵州省 2019 年中央补助地方重大疾病防治项目；贵州省卫生健康委省级重点建设学科项目
作者简介：母应姣（1999—），女，硕士在读，研究方向：慢性病预防与控制
通信作者：刘涛，E-mail: liutao9099@163.com

常使用肺功能检测，但肺功能检测成本高、操作难，在基层医疗机构开展较为困难。吕学莉等[2]研究显示我国≥40 岁人群中肺功能检测率仅为 5.9%。因此学者开始建立模型[3-4]以筛检 COPD 患者，随后建议筛查 COPD 患者到上级医院确诊，以做好 COPD 患者的“早发现、早诊断、早治疗”。近年来，早筛模型得到良好发展，一些模型[5-6]纳入年龄、海拔、结核病史、吸烟、COPD 家族史等多种变量，AUC 为 67.8% ~ 96.70%。但应注意到，这些模型的变量较为复杂，且同时不同地区存在环境与气候、生活习惯、经济水平等方面的差异，故应探索适宜各地防制工作的筛检模型。因此，本研究基于 2019—2020 年贵州省 COPD 调查数据，尝试通过 logistic 回归（logistic regression，LR）、随机森林（random forest，RF）、支持向量机（support vector machine，SVM）三种模型建立变量少、操作简单且性能较好的 COPD 患者筛检工具，为 COPD 患者的“三早预防”提供指导。

1 对象与方法

1.1 研究对象 于 2019 年 10 月 14 日—2020 年 5 月 9 日，将贵州省的县区按照性别、城镇化水平进行分层，随机抽取 9 个县区，抽中的县区按照与人口规模成比例（probability proportionate to size sampling，PPS）的方法抽取 3 个街道 / 乡镇，使用 PPS 抽样在抽中的街道 / 乡镇中随机抽取 2 个村，每个村随机抽取 1 个组（>150 户），每个组随机抽取 100 户家庭（含≥40 岁居民），采用 KISH 表法在每个家庭抽取 1 名居民进行调查。本研究开始前经中国疾控中心慢病中心伦理审查委员会审查通过（编号：201901），所有研究对象均签署知情同意书。

纳入标准：调查前 12 个月在调查地区居住 6 个月以上且年龄≥40 岁的常住居民。

排除标准：（1）居住在功能区中的居民，如工棚、军队、学生宿舍、养老院等；（2）精神疾患或认知障碍，如痴呆、理解能力障碍、聋哑等；（3）新近发现和正在治疗的肿瘤；（4）高位截瘫；（5）妊娠期或哺乳期女性人群。

1.2 调查内容 调查内容由问卷调查、体格检查及肺功能检查组成。问卷调查使用《中国居民慢性阻塞性肺疾病监测》问卷[7]，由经过贵州省疾病预防控制中心及区（县）疾控中心培训合格的调查员使用电子平板面对面询问调查对象基本人口学信息、个人及家族史、呼吸道症状、吸烟情况、烹饪燃料等内容。在安静、平整的房间测量身高、体重等体格检查。使用德国耶格公司的便携式肺功能仪进行肺功能检测，内容包含第一秒用力呼气量（FEV1）、用力肺活量（FVC）、六秒用力呼气容积（FEV6）等。

1.3 相关定义 （1）COPD 诊断：在支气管舒张试验后肺功能测试中，调查对象肺功能检测合格且 FEV1/FVC＜70%，即诊断为 COPD[5]；（2）烹饪污染燃料：使用无烟煤等煤燃料、木头或动物粪便等生物燃料进行烹饪；（3）烹饪清洁燃料：指烹饪时使用液化气、燃气或电等清洁燃料；（4）烹饪排风：烹饪使用抽油烟机、排风扇或烟囱等排风装置；（5）有害气体：对身体有害的气体和蒸汽，如汽油、农药、油烟及二氧化硫等；（6）粉尘：工作环境中的灰尘、烟尘、烟雾、粉末、金属及化合物粉尘等；（7）体质指数（body mass index，BMI）：体重（kg）/ 身高的二次方（m^2）；（8）肺功能测试不合格：肺功能检测时可接受操作≤1 次，其中可接受操作是指呼吸迅速，起始无犹豫或有效的 FEV6（用力时间＞6 s，如呼气时间在＜6 s，则要求其时间容量曲线须显示呼气相平台出现且超过 2 s）。

1.4 统计学方法 本研究连续型资料的描述与组间比较分别采用（$\bar{x} \pm s$）和 t 检验，分类资料则采用频数（构成比）和 χ^2 检验。将 COPD 可能相关因素进行 t 检验和 χ^2 检验分析，具有统计学差异的变量进行多因素 logistic 回归分析筛选模型构建变量。为使结果对贵州省 40 岁及以上人群有代表性，统计学指标的计算均经复杂加权调整。

研究数据按照 8:2 分为训练集与测试集，使用 ROSE 包对训练集数据进行平衡。以是否患 COPD（0= 否，1= 是）为结局变量，基于 LR、RF、SVM 建立筛检模型。LR 由 caret 包的 glm 函数构建；RF 由 randomForest 包的 randomForest 函数构建，使用 bootstrap 对样本进行重采样训练；SVM 由 e1071 包的线性核函数构建。使用 AUC、灵敏度、特异度等评价模型性能。使用 delong 法比较不同模型 AUC 差异。采用 R(4.2.3)统计软件对所有资料进行统计学分析。采用双侧检验，检验水准 α=0.05。

2 结 果

2.1 基本信息 本次完成所有调查人群共 5 092 人，最终纳入 4 587 人进行分析，见图 1。本次分析人群年龄（56.3 ± 9.60）岁，COPD 患者年龄（62.32 ± 9.36）岁，非 COPD 患者（55.76 ± 9.42）岁。男性 2 479 人（54.04%）、女性 2 108 人（45.96%），男性患者占比为 72.32%，多于女性患者的 27.68%。

不同年龄、不同性别、城乡、BMI 等人口学特征，14 岁前是否经常咳嗽、15 ~ 17 岁因肺炎或支气管炎住院、患哮喘、患支气管扩张症、患高血压等个人疾病史，父母患哮喘、父母患支气管扩张症等家族疾病史，是否经常咳嗽、是否起床咳嗽、是否晚上咳嗽、是否经

常咳痰、是否起床咳痰、是否反复发作的喘息、是否气短或呼吸困难等个人相关症状，每日吸烟量增加、烹饪燃料与排风、有害气体暴露等个人相关危险因素暴露在 COPD 患者与非 COPD 患者中，差异具有统计学意义（P 均＜0.05），见表 1。

2.2　COPD 多因素 logistic 回归分析　将年龄、性别、城乡、BMI 等 21 个变量纳入非条件多因素 logistic 回归分析。结果显示，年龄每增加 1 岁，COPD 的患病风险增加 5.3%；14 岁前经常咳嗽的人群患 COPD 的风险是未咳嗽人群的 2.41 倍；哮喘人群患 COPD 的风险是未患哮喘人群的 2.65 倍；每日吸烟量处于 0.1 ~ 19.9 支的人群，其患 COPD 的风险是不吸烟的人群的 1.96 倍；相较于烹饪使用清洁燃料且排风的人群，烹饪使用清洁燃料但不排风人群患 COPD 的风险增加 60.9%，烹饪使用污染燃料且不排风人群患 COPD 风险增加 36.3%；相较于没有有害气体暴露史的人群，具有有害气体暴露史人群患 COPD 的风险增加 35.8%。见表 2。

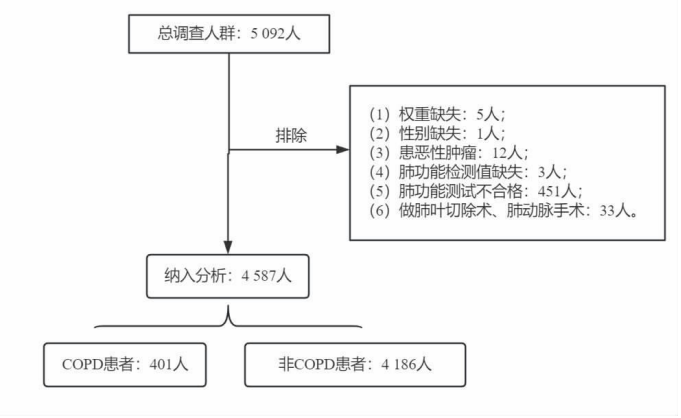

图 1　研究对象的纳入及排除流程图

Figure 1　Inclusion and exclusion of research subjects

表 1　贵州省人群 COPD 相关因素比较[$(\bar{x}\pm s)$, n(%)]

Table 1　Comparison of COPD related factors among the population in Guizhou Province [$(\bar{x}\pm s)$, n(%)]

变量	非 COPD 患者(n=4 186)	COPD 患者(n=401)	总人群(n=4 587)	χ^2/t 值	P 值
年龄（岁）	55.76 ± 9.42	62.32 ± 9.36	56.33 ± 9.60	9.523	＜0.001
性别				40.813	＜0.001
女	1 997(94.73)	111(5.27)	2 108		
男	2 189(88.30)	290(11.70)	2 479		
民族				1.593	0.207
汉族	3 131(91.52)	290(8.48)	3 421		
少数民族	1 055(90.48)	111(9.52)	1 166		
教育程度				0.145	0.862
文盲 / 半文盲	2 028(91.11)	198(8.89)	2 226		
小学 / 初中	1 797(90.99)	178(9.01)	1 975		
高中及以上	361(93.52)	25(6.48)	386		
城乡				11.032	0.001
农村	3 632(90.73)	371(9.27)	4 003		
城市	554(94.86)	30(5.14)	584		
婚姻				0.806	0.446
单身	56(90.32)	6(9.68)	62		
已婚 / 同居	3 719(91.47)	347(8.53)	4 066		
离异 / 丧偶 / 分居	411(89.54)	48(10.46)	459		
职业				2.826	0.093
务农	2 028(90.50)	213(9.50)	2 241		
非务农	2 158(91.99)	188(8.01)	2 346		
BMI（kg/m^2）	24.41 ± 3.45	23.39 ± 3.24	24.32 ± 3.45	4.029	＜0.001
相关疾病史					
14 岁前经常咳嗽				22.482	＜0.001
否	4 103(91.56)	378(8.44)	4 481		
是	83(78.30)	23(21.70)	106		
14 岁前因肺炎或支气管炎住院				3.419	0.065
否	4 137(91.34)	392(8.66)	4 529		
是	49(84.48)	9(15.52)	58		
15 ~ 17 岁因肺炎或支气管炎住院				11.932	0.001
否	4 164(91.36)	394(8.64)	4 558		
是	22(75.86)	7(24.14)	29		
患哮喘				39.383	＜0.001
否	4 104(91.75)	369(8.25)	4 473		
是	82(71.93)	32(28.07)	114		
患支气管扩张症				11.379	＜0.001
否	4 166(91.36)	394(8.64)	4 560		
是	20(74.07)	7(25.93)	27		

（续表）

变量	非 COPD 患者（n=4 186）	COPD 患者（n=401）	总人群（n=4 587）	χ^2/t 值	P 值
患高血压				11.781	0.001
否	3 368(91.90)	297(8.10)	3 665		
是	818(88.72)	104(11.28)	922		
相关家族史					
父母患哮喘				3.987	0.046
否	3 886(91.33)	369(8.67)	4 255		
是	300(90.36)	32(9.64)	332		
父母患 COPD				1.987	0.159
否	3 664(91.42)	344(8.58)	4 008		
是	522(90.16)	57(9.84)	579		
父母患慢性肺源性心脏病				0.002	0.961
否	4 108(91.33)	390(8.67)	4 498		
是	78(87.64)	11(12.36)	89		
父母患支气管扩张症				4.983	0.026
否	4 145(91.32)	394(8.68)	4 539		
是	41(85.42)	7(14.58)	48		
相关症状					
经常咳嗽				12.431	<0.001
否	3 968(91.75)	357(8.25)	4 325		
是	218(83.21)	44(16.79)	262		
起床咳嗽				11.642	0.001
否	3 988(91.76)	358(8.24)	4 346		
是	198(82.16)	43(17.84)	241		
晚上咳嗽				5.887	0.015
否	4 008(91.59)	368(8.41)	4 376		
是	178(84.36)	33(15.64)	211		
经常咳痰				17.787	<0.001
否	3 731(92.08)	321(7.92)	4 052		
是	455(85.05)	80(14.95)	535		
起床咳痰				25.803	<0.001
否	3 847(91.95)	337(8.05)	4 184		
是	339(84.12)	64(15.88)	403		
反复发作的喘息				33.286	<0.001
否	3 977(92.00)	346(8.00)	4 323		
是	209(79.17)	55(20.83)	264		
气短或呼吸困难				32.29	<0.001
否	3 838(92.19)	325(7.81)	4 163		
是	348(82.08)	76(17.92)	424		
相关危险因素					
每日吸烟量（支）				18.942	<0.001
0	2 466(94.19)	152(5.81)	2 618		
0.1 ~ 19.9	835(87.16)	123(12.84)	958		
20.0 ~ 39.9	760(87.56)	108(12.44)	868		
≥40	125(87.41)	18(12.59)	143		
14 岁前每天和吸烟者生活				1.351	0.245
否	1 533(91.74)	138(8.26)	1 671		
是	2 653(90.98)	263(9.02)	2 916		
14 岁后每天和吸烟者生活				0.887	0.346
否	1 430(90.91)	143(9.09)	1 573		
是	2 756(91.44)	258(8.56)	3 014		
接触烹饪油烟				0.268	0.605
否	2 719(91.00)	269(9.00)	2 988		
是	1 467(91.74)	132(8.26)	1 599		
烹饪燃料与排风				8.883	<0.001
清洁燃料且排风	2 006(93.43)	141(6.57)	2 147		
清洁燃料不排风	1 420(90.56)	148(9.44)	1 568		
污染燃料且排风	381(86.59)	59(13.41)	440		
污染燃料不排风	379(87.73)	53(12.27)	432		
粉尘暴露				2.252	0.134
否	696(23.63)	249(76.37)	2 945		
是	1 490(90.74)	152(9.26)	1 642		
有害气体暴露				20.916	<0.001
否	2 989(92.31)	249(7.69)	3 238		
是	1 197(88.73)	152(11.27)	1 349		

表 2　COPD 多因素 logistic 回归分析结果

Table 2　Results of multivariate logistic regression analysis for COPD

变量	比较组	参照组	β	$s_{\bar{x}}$	t/Waldχ^2 值	P 值	OR 值(95%CI)
年龄			0.052	0.008	6.881	<0.001	1.053(1.038 ~ 1.069)
性别	男	女	0.412	0.308	1.337	0.181	1.510(0.825 ~ 2.764)
城乡	城市	农村	−0.261	0.212	−1.227	0.220	0.771(0.508 ~ 1.169)
BMI			−0.029	0.022	−1.304	0.192	0.972(0.930 ~ 1.015)
14 岁前经常咳嗽	是	否	0.881	0.321	2.745	0.006	2.413(1.286 ~ 4.526)
15 ~ 17 岁因肺炎或支气管炎住院	是	否	0.757	0.521	1.453	0.146	2.132(0.768 ~ 5.922)
患哮喘	是	否	0.975	0.384	2.542	0.011	2.651(1.250 ~ 5.624)
患支气管扩张症	是	否	0.513	0.674	0.760	0.447	1.670(0.445 ~ 6.264)
患高血压	是	否	0.120	0.168	0.718	0.473	1.128(0.812 ~ 1.567)
父母患哮喘	是	否	0.138	0.270	0.512	0.609	1.148(0.677 ~ 1.948)
父母患 COPD	是	否	0.342	0.478	0.715	0.475	1.407(0.551 ~ 3.593)
父母患支气管扩张症	是	否	0.433	0.390	1.110	0.267	1.543(0.717 ~ 3.317)
经常咳嗽	是	否	−0.373	0.395	−0.945	0.344	0.689(0.318 ~ 1.493)
起床咳嗽	是	否	−0.441	0.366	−1.204	0.229	0.644(0.314 ~ 1.319)
晚上咳嗽	是	否	−0.200	0.240	−0.833	0.405	0.819(0.511 ~ 1.311)
经常咳痰	是	否	0.485	0.305	1.589	0.112	1.625(0.893 ~ 2.958)
起床咳痰	是	否	0.361	0.249	1.452	0.146	1.435(0.881 ~ 2.336)
反复发作的喘息	是	否	0.328	0.222	1.477	0.140	1.389(0.898 ~ 2.147)
气短或呼吸困难	是	否	0.441	0.316	1.395	0.163	1.554(0.837 ~ 2.885)
每日吸烟量(支)	0.1 ~ 19.9	0	0.672	0.301	2.235	0.025	1.957(1.086 ~ 3.528)
	20.0 ~ 39.9		0.441	0.413	1.069	0.285	1.554(0.692 ~ 3.491)
	≥40		0.242	0.160	1.519	0.129	1.274(0.932 ~ 1.742)
烹饪燃料与排风	清洁燃料不排风	清洁燃料且排风	0.476	0.224	2.127	0.033	1.609(1.038 ~ 2.495)
	污染燃料且排风		0.106	0.221	0.482	0.630	1.112(0.722 ~ 1.715)
	污染燃料不排风		0.310	0.147	2.115	0.034	1.363(1.023 ~ 1.817)
有害气体暴露	是	否	0.306	0.147	2.079	0.038	1.358(1.018 ~ 1.812)

2.3　COPD 患病筛检模型效果与比较　本研究训练集 COPD 患者和非 COPD 患者人数分别为 323 和 3 346 例，测试集人数分别为 78 和 840 例。通过ROSE 包 ovun.sample 函数对训练集进行平衡处理，将非 COPD 患者数量减少，患者数量增加，平衡后训练集的 COPD 患者和非 COPD 患者分别为 1 782 和 1 887 例。以是否患 COPD 作为因变量，基于多因素 logistic 回归结果，年龄、14 岁前经常咳嗽、哮喘、每日吸烟量(支)、有害气体暴露、烹饪燃料与排风 6 个因素作为筛检变量纳入模型。

训练集结果显示：RF 模型的 AUC 最大，为 87.14%，其次是 LR 模型，AUC 为 73.64%，经 delong 法检验，两者差异有统计学意义（Z=26.954，P<0.001）；最后为 SVM 模型，AUC 为 73.30%，与 RF 模型相比，差异具有统计学意义（Z=28.091，P<0.001），与 LR 模型相比，差异无统计学意义（Z=1.814，P=0.070）。表明 LR 模型和 SVM 模型在 AUC 性能方面表现相当，RF 模型性能最好。见表 3、图 2。

测试集结果显示：LR 模型的 AUC 最大，为 76.10%，其次是 SVM 模型，AUC 为 76.08%，经 delong 法检验，两者差异无统计学意义（Z=0.026，P=0.980）；最后为 RF 模型，AUC 为 70.96%，与 LR 模型相比，差异有统计学意义（Z=−2.925，P=0.003），与 SVM 模型相比，差异有统计学意义（Z=−3.078，P=0.002）。表明 LR 模型和 SVM 模型在 AUC 性能方面表现相当，RF 模型性能最差。见表 3、图 3。

表 3　LR、RF、SVM 模型筛检效果(%)

Table 3　The screening effectiveness of LR, RF, and SVM models (%)

数据集	参数	LR	RF	SVM
训练集	AUC	73.64	87.14	73.30
	(95%CI)	(72.05,75.23)	(86.06,88.22)	(71.10,74.91)
	灵敏性	61.73	81.99	64.59
	特异性	69.37	70.85	68.42
	约登指数	31.10	52.84	33.01
测试集	AUC	76.10	70.96	76.08
	(95%CI)	(70.69,81.50)	(65.61,76.31)	(70.58,81.59)
	灵敏性	70.51	65.39	73.08
	特异性	71.31	66.19	69.52
	约登指数	41.82	31.58	42.60

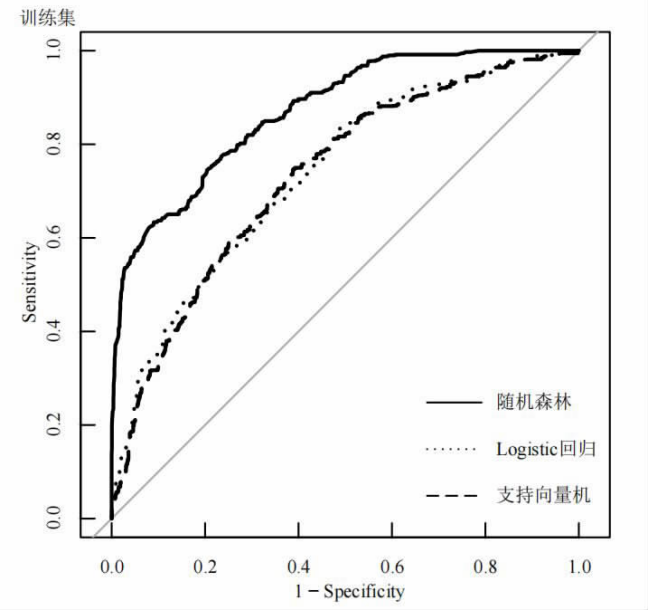

图 2　三种模型训练集的 ROC 曲线图

Figure 2　ROC curves of the training set of three models

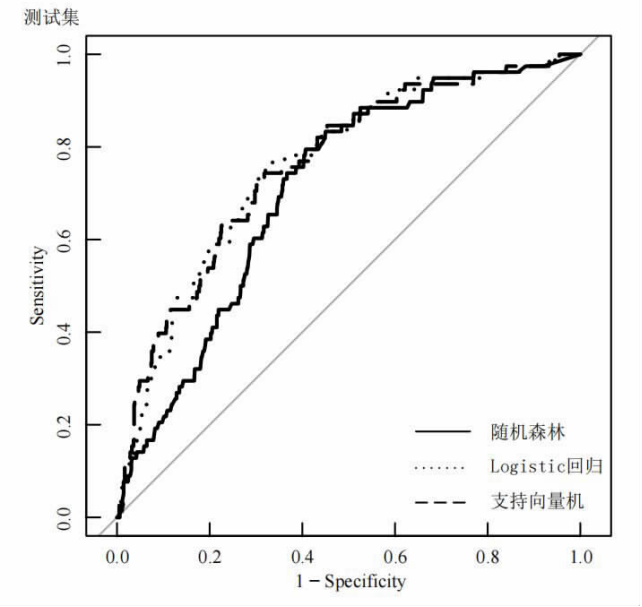

图 3　三种模型测试集的 ROC 曲线图

Figure 3　ROC curves of the test set of three models

3　讨　论

Wang 等[8]研究显示我国成年人中 COPD 患者接近1 亿，但仅 12%的 COPD 患者做过肺功能检测，多数患者并不知晓自己患病。目前 COPD 确诊主要依靠肺功能检测，而作为金标准的肺功能检测成本高、操作难，大规模应用于人群筛查具有一定局限性。机器学习算法已被广泛应用于疾病筛检，国内研究使用 LR、SVM、RF、决策树、神经网络等多种方法建立高血压、糖尿病及动脉粥样硬化等[9-14]筛检模型。多个地区均在探索 COPD 早筛工具，周家为等[15]使用多种问卷模型筛检 COPD 风险人群，其中《慢阻肺人群筛查问卷》效果最佳，该问卷包括年龄、吸烟、呼吸、气短及咳痰5 个变量，其特异度、约登指数均低于本研究，分别为58.25%、37.00%。李章龙[16]通过 BMI、家族呼吸系统疾病史及生物燃料暴露等 16 种变量构建 SVM、决策树等五种筛检模型，其中 RF 效果最佳，AUC=96.35%。Wang 等人[4]通过烹饪燃料类型、烹饪排风及支气管扩张实验后 FEV1 等指标列线图来筛检 CODP 患者，模型 AUC=81%。虽然上述研究筛检效果较好，但其纳入变量较多、且包含体格检查与肺功能检查指标，收集较麻烦，不便于基层医疗机构使用。本研究建立的三种筛检模型仅纳入年龄及哮喘等 6 个简单变量，相较其他地区的多个筛检变量、体检指标或肺功能检测指标[4,16-17]，变量相对简单、易于收集。且本研究结果显示测试集 LR、RF、SVM 的 AUC 分别为 76.10%、70.96%、76.08%，均具有较好的筛检效果，相较上述提到的其他模型，更适宜在基层医疗机构进行推广。

本研究使用 delong 法比较三种模型的筛检效果，训练集结果显示 LR 模型和 SVM 模型在 AUC 性能方面表现相当，RF 模型性能最好。测试集结果显示 LR 模型和 SVM 模型在 AUC 性能方面表现相当，RF 模型性能最差。RF 是集成学习方法，可同时纳入定性和定量变量，是一个包含许多随机生成的决策树的集成分类器[18]。李章龙[16]构建多种 COPD 筛检模型，发现 RF 效果最佳，但王娇娇[9]发现 RF 预测钢铁工人动脉粥样硬化风险的效果低于 SVM。本研究训练集结果显示 RF 筛检效果强于 SVM 与 LR，但在测试集中弱于 LR 与 SVM，可能是 RF 模型在训练集上的学习能力过强，发生过拟合[9]，或是因为测试集数据有限，限制了 RF 抽样随机的优势。SVM 是二分类的监督学习模型，可将高维大数据分类为少量的数据点，从而在短时间内实现类别的区分[19]，在疾病筛检方面得到广泛应用[20]。本次研究显示测试集 SVM 模型的 AUC、约登指数分别为 76.08%、42.60%，灵敏度、特异度分别为 73.08%、69.52%，表明该模型筛检效果较好，准确判定调查对象 COPD 患者的筛检价值较高。LR 是经典方法，适用于二分类结局资料，本次研究结果显示测试集 LR 模型的 AUC（76.10%）、灵敏度（70.51%）与特异度（71.31%）均较高，显示该模型在调查对象中准确判断 COPD 患者的筛检价值较高。

综上所述，基于 LR、RF、SVM 构建的 COPD 患者筛检模型，测试集中 SVM 模型与 LR 模型性能相当，可为 COPD 筛检提供参考。此外，构建模型所纳入的 6 个变量都较易获得，由此构建的筛检工具使用方便、操作简单，基层医疗机构可用其筛检 COPD 患者，提醒筛查患者进行确诊与治疗，做到 COPD 的“三早预防”，提升患者的生活质量。由于本研究仅进行内

部验证，缺乏外部验证，所以存在一定局限性；且本研究基于横断面研究数据，下一步需要进行前瞻性研究来验证该模型的可靠性。

利益冲突声明　本研究不存在任何利益冲突

参考文献

[1] 刘涛，李凌，周婕，等. 贵州省常住居民死亡原因研究报告-2017[M]. 贵阳：贵州科技出版社，2021.
Liu T, Li L, Zhou J, et al. A Study Report on the Causes of Death among Permanent Residents in Guizhou Province (2017) [M]. Guiyang: Guizhou Science and Technology Publishing House, 2021.

[2] 吕学莉，丛舒，樊静，等. 2014-2015 年中国 40 岁及以上慢性阻塞性肺疾病患者肺功能检查率及其影响因素分析[J]. 中华流行病学杂志，2020，41(5)：672-677.
Lv XL, Cong S, Fan J, et al. Analyses of the rate of spirometry examination and its related factors in chronic obstructive pulmonary disease patients aged 40 years or older in China, 2014-2015 [J]. Chinese Journal of Epidemiology, 2020, 41(5): 672-677.

[3] Lee SC, An C, Yoo J, et al. Development and validation of a nomogram to predict pulmonary function and the presence of chronic obstructive pulmonary disease in a Korean population [J]. BMC Pulmonary Medicine, 2021, 21(1): 32.

[4] Wang YD, Li Z, Li FS. Development and assessment of prediction models for the development of COPD in a typical rural area in northwest China [J]. International Journal of Chronic Obstructive Pulmonary Disease, 2021, 16: 477-486.

[5] Zhang BY, Sun D, Niu HT, et al. Development of a prediction model to identify undiagnosed chronic obstructive pulmonary disease patients in primary care settings in China [J]. Chin Med J (Engl), 2023, 136(6): 676-682.

[6] Lin A, Mao C, Rao BQ, et al. Development and validation of nomogram including high altitude as a risk factor for COPD: A cross-sectional study based on Gansu population [J]. Front Public Health, 2023, 11: 1127566.

[7] 方利文，包鹤龄，王宝华，等. 中国居民慢性阻塞性肺疾病监测内容与方法概述 [J]. 中华流行病学杂志，2018，39(5)：546-550.
Fang LW, Bao HL, Wang BH, et al. A summary of item and method of National chronic obstructive pulmonary disease surveillance in China[J]. Chinese Journal of Epidemiology, 2018, 39(5): 546-550.

[8] Wang C, Xu JY, 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[J]. The Lancet, 2018, 391(10131): 1706-1717.

[9] 王娇娇，陈圆煜，郑子薇，等. 三种风险预测模型预测钢铁工人颈动脉粥样硬化的效能比较[J]. 中国全科医学，2022，25(11)：1334-1339.
Wang JJ, Chen YY, Zheng ZW, et al. Comparison of three risk prediction models for carotid atherosclerosis in steelworkers [J]. Chinese General Practice, 2022, 25(11): 1334-1339.

[10] 李禄伟，黄倩，施佳成，等. 基于三种统计学方法构建的超重及肥胖人群高血压发病预测模型的分析比较[J]. 现代预防医学，2021，48(11)：2061-2066.
Li LW, Huang Q, Shi JC, et al. Screening risk factors and interaction analysis of hypertension in overweight and obesity population based on three statistical models[J]. Modern Preventive Medicine, 2021, 48 (11): 2061-2066.

[11] 侯强，王英，李慕帆，等. 太行山食管癌高发区农村居民高血压风险预测模型构建[J]. 现代预防医学，2023，50(6)：1133-1138.
Hou Q, Wang Y, Li MF, et al. Construction of hypertension risk prediction model for rural residents with high esophageal cancer in Taihang Mountain [J]. Modern Preventive Medicine, 2023, 50 (6): 1133-1138.

[12] Khan A, Khan A, Khan MM, et al. Cardiovascular and diabetes diseases classification using ensemble stacking classifiers with SVM as a Meta classifier[J]. Diagnostics (Basel), 2022, 12(11): 2595.

[13] Nuryani N, Pambudi Utomo T, Wiyono N, et al. Cuffless hypertension detection using swarm support vector machine utilizing photoplethysmogram and electrocardiogram [J]. J Biomed Phys Eng, 2023, 13(5): 477-488.

[14] Silva GFS, Fagundes TP, Teixeira BC, et al. Machine learning for hypertension prediction: a systematic review [J]. Current Hypertension Reports, 2022, 24(11): 523-533.

[15] 周家为，王玮. 不同问卷对慢阻肺的筛查价值[J]. 中华健康管理学杂志，2022，16(7)：444-449.
Zhou JW, Wang W. Comparative study on the accuracy of different questionnaires in COPD screening [J]. Chinese Journal of Health Management, 2022, 16(7): 444-449.

[16] 李章龙. 广东省社区 COPD 患者知识知晓、检查、治疗状况及 COPD 诊断模型[D]. 广州：广东药科大学，2021.
Li ZL. Knowledge awareness, examination and treatment status of COPD parientsand COPD diagnosis model in the community of Guangdong province [D]. Guangzhou: Guangdong Pharmaceutical University, 2021.

[17] 董秋月，高丛丛，刘才睿，等. 山东省≥40 岁人群慢性阻塞性肺疾病患病风险列线图预测模型建立 [J]. 中国公共卫生，2023，39(5)：604-611.
Dong QY, Gao CC, Liu CR, et al. Establishment of a nomogram-based risk prediction model for chronic obstructive pulmonary disease in residents aged 40 years and over in Shandong province [J]. Chinese Journal of Public Health, 2023, 39 (5): 604-611.

[18] Hong WD, Lu YJ, Zhou XY, et al. Usefulness of random forest algorithm in predicting severe acute pancreatitis [J]. Frontiers in Cellular and Infection Microbiology, 2022, 12: 893294.

[19] Nordin NI, Mustafa WA, Lola MS, et al. Enhancing COVID-19 classification accuracy with a hybrid SVM-LR model [J]. BIOENGINEERING-BASEL, 2023, 10(11): 1318.

[20] Pokorny T, Vrba J, Fiser O, et al. On the role of training data for SVM-Based microwave brain stroke detection and classification [J]. Sensors, 2023, 23(4): 2031.

收稿日期：2023-12-01