DOI：10.16781/j.CN31-2187/R.20211053

·论　著·

基于机器学习的原发性高血压并发脑梗死的风险预测模型

刘　婷[1,2]，朱　琴[1]，徐　琳[1]，杜志银[1,2*]
1. 重庆医科大学医学信息学院卫生信息管理与决策教研室，重庆 400016
2. 重庆医科大学医学数据研究院，重庆 400016

［摘要］ **目的**　利用机器学习算法构建原发性高血压并发脑梗死的风险预测模型，并探索原发性高血压患者并发脑梗死的危险因素。**方法**　收集重庆市 7 家医院 2015 年 1 月 1 日至 2019 年 12 月 31 日确诊的 1 478 例原发性高血压并发脑梗死患者及 2 826 例无脑梗死的原发性高血压患者的 42 项临床指标资料。采用单因素分析筛选输入指标，将 4 304 名患者按照 7 ∶ 3 随机分为训练集（n=3 012）和测试集（n=1 292），训练集的数据用于构建 logistic 回归、决策树、随机森林、XGBoost 模型，测试集中的数据用于内部验证。计算各输入指标在 4 个模型中的相对重要性评分，使用阳性预测值、阴性预测值、准确度、F1 值、ROC 曲线的 AUC 值及 Delong 检验等评价 4 个模型对原发性高血压并发脑梗死的预测价值。**结果**　单因素分析筛选出 29 项差异有统计学意义的指标，基于此构建的 logistic 回归、决策树、随机森林和 XGBoost 模型预测原发性高血压并发脑梗死的 AUC 值均较高。Delong 检验结果显示，随机森林和 XGBoost 模型的预测性能均优于 logistic 回归和决策树模型，其中 XGBoost 模型的阴性预测值、准确度、F1 值、AUC 值均最高，分别为 0.780（95% *CI* 0.778～0.782）、0.766（95% *CI* 0.764～0.768）、0.603（95% *CI* 0.599～0.607）、0.808（95% *CI* 0.804～0.811）。相对重要性评分结果显示，logistic 回归、决策树、随机森林、XGBoost 模型均提示血细胞比容、白蛋白、就诊年龄、白细胞计数、胆碱酯酶和载脂蛋白 A1 是原发性高血压并发脑梗死的重要影响因素。**结论**　基于机器学习的预测原发性高血压并发脑梗死风险的 logistic 回归、决策树、随机森林和 XGBoost 模型均有较高的诊断价值，其中 XGBoost 模型的综合诊断效能最佳。血细胞比容、白蛋白、就诊年龄、白细胞计数、胆碱酯酶和载脂蛋白 A1 可用于预测原发性高血压患者的脑梗死患病风险。

［关键词］ 原发性高血压；脑梗死；机器学习；危险因素；预测模型
［中图分类号］ R 743.33；R 544.1　　［文献标志码］ A　　［文章编号］ 2097-1338(2022)03-0258-08

Risk prediction models of essential hypertension complicated with cerebral infarction based on machine learning algorithm

LIU Ting[1,2], ZHU Qin[1], XU Lin[1], DU Zhi-yin[1,2*]
1. Department of Health Information Management and Decision Making, School of Medical Informatics, Chongqing Medical University, Chongqing 400016, China
2. Medical Data Science Academy, Chongqing Medical University, Chongqing 400016, China

［**Abstract**］ **Objective**　To construct a risk prediction model of essential hypertension complicated with cerebral infarction based on machine learning algorithm, and explore the risk factors. **Methods**　The data of 42 clinical indexes of 1 478 patients with essential hypertension complicated with cerebral infarction and 2 826 patients with essential hypertension without cerebral infarction in 7 hospitals of Chongqing from Jan. 1, 2015 to Dec. 31, 2019 were collected. Univariate analysis was used to screen the input indexes. The 4 304 patients were randomly divided into training set (n=3 012) and test set (n=1 292) with a ratio of 7 ∶ 3. The data of the training set was used to construct logistic regression, decision tree, random forest and XGBoost models, and the data of the test set was used for internal verification. The relative importance scores of each input index in the 4 models were calculated. The positive predictive value, negative predictive value, accuracy, F1 value, area under curve (AUC) value of receiver operating characteristic (ROC) curve and Delong test were used to

［收稿日期］ 2021-10-20　　［接受日期］ 2021-11-16
［基金项目］ 重庆医科大学校级哲学社会科学专项科研项目（201725），重庆医科大学智慧医学研究项目（YJSZHYX202002）. Supported by Special Research Project for Philosophy and Social Science of Chongqing Medical University (201725) and Intelligent Medicine Research Project of Chongqing Medical University (YJSZHYX202002).
［作者简介］ 刘　婷，硕士生. E-mail: tingliu@stu.cqmu.edu.cn
*通信作者（Corresponding author）. Tel: 023-68480060, E-mail: duzhiyin@cqmu.edu.cn

evaluate the predictive diagnostic value of the 4 models for essential hypertension complicated with cerebral infarction. **Results**　A total of 29 statistically significant indexes were selected by univariate analysis. The AUC values of essential hypertension complicated with cerebral infarction predicted by logistic regression, decision tree, random forest and XGBoost models were higher. The results of Delong test showed that the prediction performance of random forest and XGBoost models was better than that of logistic regression and decision tree models. The negative predictive value, accuracy, F1 value and AUC value of XGBoost model were the highest, being 0.780 (95% confidence interval［*CI*］ 0.778-0.782), 0.766 (95% *CI* 0.764-0.768), 0.603 (95% *CI* 0.599-0.607) and 0.808 (95% *CI* 0.804-0.811), respectively. The results of relative importance scores showed that logistic regression, decision tree, random forest and XGBoost models all suggested that hematocrit, albumin, age, white blood cell count, choline esterase and apolipoprotein A1 were important influencing factors of essential hypertension complicated with cerebral infarction. **Conclusion**　The risk prediction models of essential hypertension complicated with cerebral infarction based on machine learning, such as logistic regression, decision tree, random forest and XGBoost models, have high diagnostic value, among which XGBoost model has the best comprehensive diagnostic efficiency. Hematocrit, albumin, age, white blood cell count, choline esterase and apolipoprotein A1 can be used to predict the risk of cerebral infarction in patients with essential hypertension.

［**Key words**］ essential hypertension; cerebral infarction; machine learning; risk factors; prediction model

［Acad J Naval Med Univ, 2022, 43(3): 258-265］

《中国心血管健康与疾病报告 2020 概要》显示，心血管病是我国城乡居民死亡的首位原因，且其患病率处于持续上升趋势[1]。原发性高血压起病隐匿，进展缓慢，是最常见的心血管疾病之一，可导致多种靶器官损伤，其中脑梗死是其最常见的并发症[2]。脑梗死的病程较长，多在安静休息或睡眠状态下急性发病，主要表现为肢体无力、口角歪斜、语言或意识障碍等，具有高发病率、高复发率、高致残率等特点[3]，严重影响了患者的生活质量，增加了家庭和社会的负担，已成为我国重大公共卫生问题之一。

既往关于原发性高血压并发脑梗死的研究多对其病因构成、诊断方法、治疗及预后等进行临床分析[4-5]，仅有少数基于危险因素建立预测模型的研究[6-7]，如 Ma 等[6]基于缺血修饰白蛋白、脂蛋白相关磷脂酶 A2 等构建了高血压并发脑梗死的预后预测模型，AUC 值为 0.969。但目前国内尚无研究人员基于真实世界的临床数据通过机器学习方法对原发性高血压导致脑梗死的风险进行综合分析。

本研究结合患者的一般资料、病因和实验室检查指标构建了一种基于机器学习算法的原发性高血压并发脑梗死的个体风险预测模型，以期辅助临床医师尽早识别原发性高血压患者中并发脑梗死的患者，为临床防治工作提供参考依据。

1　资料和方法

1.1　研究对象　本研究数据来源于 7 家大型医院的电子病历数据库，从 2015 年 1 月 1 日至 2019 年 12 月 31 日诊断为原发性高血压的 33 532 例患者资料中根据纳入和排除标准获得 4 304 例原发性高血压患者的资料，来源于重庆市东南医院 319 例、重庆市铜梁区人民医院 89 例、重庆医科大学附属第二医院 2 881 例、重庆医科大学附属第三医院 54 例、重庆医科大学附属大学城医院 947 例、重庆医科大学附属儿童医院 4 例、重庆医科大学附属永川医院 10 例，其中 1 478 例原发性高血压并发脑梗死患者为研究组，2 826 例无脑梗死的原发性高血压患者为对照组。

研究组纳入标准：（1）首次诊断为脑梗死，既往史中无脑梗死病史；（2）头颅 CT 或 MRI 检查证实脑梗死，且出院诊断为脑梗死；（3）既往史或现病史中有明确的高血压患病年数，且病案首页中有明确的原发性高血压诊断。研究组排除标准：（1）由动脉炎、烟雾病、血液系统疾病（如红细胞增多症）、结缔组织病等引起的脑梗死；（2）合并严重感染、内分泌疾病、肿瘤等。对照组纳入标准：（1）病案首页有原发性高血压的明确诊断；（2）既往史中有明确原发性高血压病史及患病年数。对照组排除标准：（1）有肾动脉狭窄、主动脉狭窄等心、脑、肾血管病变；（2）合并继发性高血压、结核、肿瘤等。本研究通过重庆医科大学医学研究伦理委员会审批。

1.2　观察指标选取　基于文献检索和临床诊疗指南[3,8]，选取患者的一般资料、血常规、凝血功

能、尿常规、血生物化学指标等基线资料作为原发性高血压并发脑梗死的可能影响因素。所有实验室指标均取住院患者的首次检查结果，排除缺失率≥30%的指标，最终共纳入42项指标进行研究。

1.3　统计学处理　采用Excel 2013进行数据预处理，采用SPSS 25.0软件进行单因素分析以筛选重要指标，采用Python 3.6.9的scikit-learn 0.21.3及XGBoost 1.3.1工具实现机器学习模型构建与评估。将性别、尿胆红素等分类变量编码为二进制值，将缺失率＜30%的指标进行均值填补。采用Kolmogorov-Smirnov检验进行正态性检验，经检验发现研究组和对照组所有计量资料均为偏态分布，因此本研究计量资料以中位数（下四分位数，上四分位数）表示，两组间比较采用Mann-Whitney U检验。计数资料以例数和百分数表示，两组间比较用χ^2检验。将单因素分析差异有统计学意义（P＜0.05）的指标进行逐步向前logistic回归分析（$\alpha_{入}$=0.05，$\alpha_{出}$=0.1），筛选原发性高血压并发脑梗死的独立危险因素。

将筛选后的变量作为输入变量，以是否发生脑梗死作为结局变量，使用scikit-learn 0.21.3中train_test_split模块将全部样本按照7∶3随机分为训练集（n=3 012）和测试集（n=1 292），设置其参数stratify=y以保持划分后的数据集中阳性和阴性病例之间的平衡，通过random_state参数将数据进行100次随机拆分，在训练集中分别使用LogisticRegression、DecisionTreeClassifier、RandomForestClassifier、XGBClassifier模块构建logistic回归、决策树、随机森林、XGBoost模型；采用GridSearchCV模块（网格搜索算法）对每个模型进行参数调优，将ROC曲线的AUC值作为评价指标。在测试集中采用阳性预测值、阴性预测值、准确度、F1值、AUC值及Delong检验评估4种模型的预测性能，所有指标的结果均以$\bar{x}$及其95% CI表示。采用各模型的feature_importances_计算各个指标的相对重要性评分，评分绝对值越高的指标对模型的影响越大。

2　结　果

2.1　研究组与对照组基线资料的单因素分析　研究组患者的就诊年龄、男性患者占比、糖尿病患者占比、动脉粥样硬化患者占比、吸烟患者占比、饮酒患者占比、白细胞计数、红细胞分布宽度变异系数、碱性磷酸酶、直接胆红素均高于对照组，血细胞比容、大血小板比率、血小板计数、血小板比容、血小板体积分布宽度、前白蛋白、白蛋白、总胆固醇、甘油三酯、胆碱酯酶、载脂蛋白A1、载脂蛋白B、氯离子、钠离子、钾离子、丙氨酸转氨酶、天冬氨酸转氨酶、高密度脂蛋白胆固醇、低密度脂蛋白胆固醇均低于对照组（P均＜0.05，表1）。

2.2　原发性高血压并发脑梗死影响因素的logistic回归分析　以原发性高血压患者是否并发脑梗死为因变量（是=1，否=0），单因素分析差异有统计学意义的指标为自变量，进行logistic回归分析。结果显示患者就诊年龄、白细胞计数、糖尿病、动脉粥样硬化、性别（男）是原发性高血压并发脑梗死的独立危险因素，血细胞比容、白蛋白、胆碱酯酶、氯离子、钾离子、丙氨酸转氨酶、高密度脂蛋白胆固醇是原发性高血压并发脑梗死的独立保护因素（表2）。

2.3　构建机器学习模型　将表1中两组间比较差异有统计学意义的29项指标纳入logistic回归、决策树、随机森林及XGBoost模型4个机器学习模型，在训练集中通过GridSearchCV模块确定每个模型的最优参数，以ROC曲线的AUC值作为评价指标，得到各个模型的最优参数分别为：logistic回归：penalty='L1'，C=0.3，solver='liblinear'；决策树criterion='gini'，splitter='best'，max_depth=5，max_features=16，min_samples_leaf=2，random_state=1，min_impurity_split=0.001；随机森林模型n_estimators=120，max_features='auto'，criterion='entropy'，max_depth=30，min_samples_leaf=10；XGBoost模型learning_rate=0.01，max_depth=7，objective='binary:logistic'，min_child_weight=10，alpha=0，subsample=0.85，colsample_bytree=0.7，n_estimators=1 000。

表 1　研究组与对照组基线资料的单因素分析

Tab 1　Univariate analysis of baseline data of study and control groups

Index	Control group N=2 826	Study group N=1 478	Statistic	P value
General information				
Age/year, M (Q_L, Q_U)	64.00 (54.00, 73.00)	70.50 (62.00, 78.00)	Z=−14.246	<0.001
Male, n (%)	1 147 (40.6)	759 (51.4)	χ^2=45.587	<0.001
Hypertension duration/year, M (Q_L, Q_U)	7.00 (2.00, 10.00)	9.08 (3.00, 10.00)	Z=−1.391	0.164
Diabetes mellitus, n (%)	661 (23.4)	466 (31.5)	χ^2=33.261	<0.001
Atherosclerosis, n (%)	1 427 (50.5)	994 (67.3)	χ^2=110.739	<0.001
Smoking, n (%)	693 (24.5)	450 (30.4)	χ^2=17.463	<0.001
Drinking, n (%)	596 (21.1)	375 (25.4)	χ^2=10.186	0.001
Blood routine, M (Q_L, Q_U)				
White blood cell/(L^{-1}, $\times10^9$)	6.30 (5.20, 7.48)	6.86 (5.68, 8.39)	Z=−9.731	<0.001
Mean corpusular volume/fL	90.00 (87.40, 93.00)	89.80 (87.10, 93.00)	Z=−0.595	0.552
RDW-CV/%	13.30 (12.80, 13.90)	13.50 (12.90, 14.00)	Z=−5.118	<0.001
RDW-SD/fL	43.60 (41.90, 44.40)	43.60 (41.90, 44.90)	Z=−1.472	0.141
Hematocrit/%	37.85 (21.55, 41.60)	26.41 (0.40, 39.50)	Z=−14.295	<0.001
MCH/pg	30.30 (29.40, 31.30)	30.20 (29.28, 31.30)	Z=−0.086	0.932
MCHC/(g•L^{-1})	335.34 (329.00, 342.00)	335.34 (328.00, 342.00)	Z=−0.222	0.824
Platelet-large cell rate/%	34.85 (30.00, 40.50)	34.85 (28.90, 39.71)	Z=−2.865	0.004
Platelet/(L^{-1}, $\times10^9$)	194.56 (158.00, 228.00)	191.00 (153.00, 222.00)	Z=−2.140	0.032
Platelet hematocrit/%	0.22 (0.19, 0.25)	0.22 (0.18, 0.25)	Z=−2.710	0.007
PDW/%	14.32 (12.50, 16.00)	14.32 (12.20, 15.70)	Z=−2.506	0.012
Mean platelet volume/fL	11.18 (10.60, 12.00)	11.18 (10.50, 11.90)	Z=−1.818	0.069
Coagulation function, M (Q_L, Q_U)				
APTT/s	33.53 (31.30, 36.50)	33.75 (30.90, 36.80)	Z=−1.362	0.173
Fibrinogen concentration/(g•L^{-1})	3.31 (2.83, 3.52)	3.30 (2.85, 3.66)	Z=−1.425	0.154
Urine routine, n (%)				
Urinary bilirubin positive	9 (0.3)	10 (0.7)	χ^2=2.832	0.092
Blood biochemistry, M (Q_L, Q_U)				
Prealbumin/(mg•L^{-1})	247.15 (221.00, 285.00)	242.00 (206.00, 269.00)	Z=−8.125	<0.001
Albumin/(g•L^{-1})	42.40 (39.90, 44.90)	40.89 (38.10, 43.20)	Z=−12.460	<0.001
Total cholesterol/(mmol•L^{-1})	4.67 (4.05, 5.27)	4.61 (3.86, 5.20)	Z=−4.034	<0.001
Total bilirubin/(μmol•L^{-1})	11.50 (8.80, 14.27)	11.40 (8.90, 14.30)	Z=−0.208	0.836
Triglyceride/(mmol•L^{-1})	1.48 (1.06, 2.01)	1.44 (1.01, 1.90)	Z=−2.261	0.024
Alkaline phosphatase/(U•L^{-1})	79.00 (66.00, 89.38)	80.06 (66.00, 93.00)	Z=−2.052	0.040
Creatinine/(μmol•L^{-1})	69.70 (57.27, 83.57)	70.80 (58.87, 84.20)	Z=−1.637	0.102
Choline esterase/(U•L^{-1})	8 063.17 (7 410.13, 9 030.00)	8 063.17 (6 646.68, 8 742.50)	Z=−7.262	<0.001
Apolipoprotein A1/(g•L^{-1})	1.55 (1.42, 1.71)	1.55 (1.31, 1.66)	Z=−6.987	<0.001
Apolipoprotein B/(g•L^{-1})	0.98 (0.84, 1.09)	0.98 (0.80, 1.10)	Z=−2.341	0.019
Chloride ion/(mmol•L^{-1})	104.00 (102.10, 106.50)	103.84 (101.30, 105.91)	Z=−5.242	<0.001
Sodium ion/(mmol•L^{-1})	141.90 (140.30, 143.90)	141.63 (139.60, 143.10)	Z=−6.826	<0.001
Potassium ion/(mmol•L^{-1})	3.92 (3.69, 4.17)	3.90 (3.60, 4.12)	Z=−4.733	<0.001
Magnesium ion/(mmol•L^{-1})	0.87 (0.82, 0.90)	0.87 (0.81, 0.91)	Z=−1.246	0.213
Alanine aminotransferase/(U•L^{-1})	19.96 (14.00, 27.24)	18.00 (12.61, 23.98)	Z=−6.932	<0.001
Aspartate aminotransferase/(U•L^{-1})	21.00 (17.88, 25.80)	21.00 (17.00, 25.00)	Z=−2.807	0.005
Direct bilirubin/(μmol•L^{-1})	3.93 (3.00, 4.90)	4.10 (3.10, 5.20)	Z=−2.977	0.003
Indirect bilirubin/(μmol•L^{-1})	7.60 (5.60, 9.60)	7.40 (5.50, 9.30)	Z=−1.528	0.126
HDL-C/(mmol•L^{-1})	1.22 (1.03, 1.41)	1.17 (0.97, 1.35)	Z=−5.141	<0.001
LDL-C/(mmol•L^{-1})	2.67 (2.15, 3.17)	2.66 (2.07, 3.11)	Z=−2.100	0.036

Control group: The essential hypertension patients without cerebral infarction; Study group: The essential hypertension patients with cerebral infarction. RDW-CV: Coefficient of variation of red cell volume distribution width; RDW-SD: Standard deviation of red cell volume distribution width; MCH: Mean corpuscular hemoglobin content; MCHC: Mean corpuscular hemoglobin concentration; PDW: Platelet distribution width; APTT: Activated partial thromboplastin time; HDL-C: High density lipoprotein-cholesterol; LDL-C: Low density lipoprotein-cholesterol; M (Q_L, Q_U): Median (lower quartile, upper quartile).

表 2　原发性高血压并发脑梗死影响因素的 logistic 回归分析

Tab 2　Logistic regression analysis of influencing factors of essential hypertension complicated with cerebral infarction

Index	*b*	*SE*	*OR* (95% *CI*)	*P* value
Age	0.032	0.003	1.033 (1.026, 1.039)	＜0.001
White blood cell	0.126	0.016	1.135 (1.101, 1.170)	＜0.001
Hematocrit	−0.033	0.002	0.967 (0.963, 0.971)	＜0.001
Albumin	−0.077	0.011	0.926 (0.907, 0.946)	＜0.001
Choline esterase	0.000	0.000	1.000 (1.000, 1.000)	0.012
Chloride ion	−0.042	0.009	0.959 (0.942, 0.977)	＜0.001
Potassium ion	−0.383	0.083	0.682 (0.579, 0.803)	＜0.001
ALT	−0.007	0.002	0.993 (0.989, 0.997)	0.001
HDL-C	−0.257	0.118	0.774 (0.614, 0.975)	0.030
Diabetes mellitus	0.190	0.081	1.209 (1.031, 1.418)	0.019
Atherosclerosis	0.395	0.074	1.485 (1.284, 1.717)	＜0.001
Gender (male)	0.540	0.076	1.716 (1.480, 1.990)	＜0.001

ALT: Alanine aminotransferase; HDL-C: High density lipoprotein-cholesterol; *b*: Regression coefficient; *SE*: Standard error; *OR*: Odds ratio; *CI*: Confidence interval.

2.4　*各模型对原发性高血压并发脑梗死的预测性能比较*　将构建的 logistic 回归、决策树、随机森林和 XGBoost 模型在测试集中进行内部验证，结果显示各模型的 AUC 值均较高。Delong 检验结果显示，随机森林模型和 XGBoost 模型的预测性能均优于 logistic 回归模型和决策树模型，其中 XGBoost 模型的阴性预测值、准确度、F1 值、AUC 值均最高。见图 1、表 3。

2.5　*原发性高血压并发脑梗死的影响因素分析*　4 种机器学习模型的相对重要性评分结果显示，在单因素分析差异有统计学意义的 29 个指标中，血细胞比容、白蛋白、就诊年龄、白细胞计数在 4 个模型中的相对重要性评分绝对值均较大，胆碱酯酶、载脂蛋白 A1 在随机森林和 XGBoost 模型中的相对重要性评分绝对值均较大，表明这几个指标可能是原发性高血压并发脑梗死的重要影响因素（表 4）。

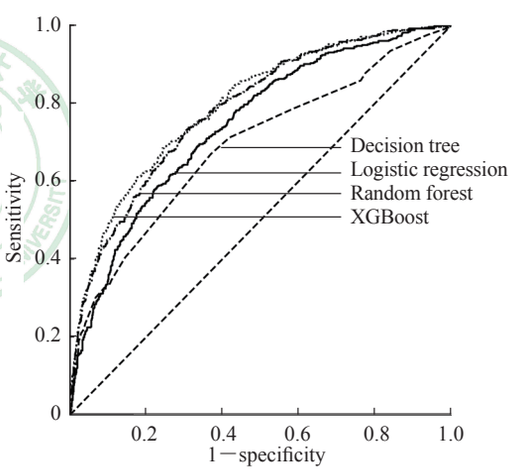

图 1　各模型在测试集中预测原发性高血压并发脑梗死的 ROC 曲线

Fig 1　ROC curves of each model in test set for prediction of essential hypertension complicated with cerebral infarction

ROC: Receiver operating characteristic.

表 3　各模型在测试集中对原发性高血压并发脑梗死的预测性能

Tab 3　Predictive performance of each model for essential hypertension complicated with cerebral infarction in test set

$\bar{x}$ (95% *CI*)

Model	PPV	NPV	Accuracy	F1 value	AUC
Logistic	0.655 (0.646, 0.665)	0.746 (0.744, 0.749)	0.726 (0.722, 0.729)	0.517 (0.511, 0.524)	0.755 (0.750, 0.760)
Decision tree	0.668 (0.649, 0.686)	0.730 (0.726, 0.734)	0.717 (0.713, 0.722)	0.466 (0.451, 0.481)	0.708 (0.698, 0.717)
Random forest	0.792 (0.783, 0.801)	0.741 (0.738, 0.743)	0.749 (0.746, 0.751)	0.499 (0.491, 0.506)	0.796 (0.792, 0.800)
XGBoost	0.724 (0.716, 0.732)	0.780 (0.778, 0.782)	0.766 (0.764, 0.768)	0.603 (0.599, 0.607)	0.808 (0.804, 0.811)

PPV: Positive predictive rate; NPV: Negative predictive rate; AUC: Area under curve; *CI*: Confidence interval.

表 4　4 种机器学习模型特征相对重要性排序

Tab 4　Ranking of relative importance of features of 4 machine learning models

Rank	Logistic regression		Decision tree		Random forest		XGBoost	
	Indicator	RI	Indicator	RI	Indicator	RI	Indicator	RI
1	Hematocrit	−0.559	Hematocrit	0.242	Hematocrit	0.121	Hematocrit	0.071
2	Age	0.437	Albumin	0.125	Age	0.084	Choline esterase	0.062
3	White blood cell	0.350	White blood cell	0.104	White blood cell	0.074	Age	0.057
4	Albumin	−0.273	Chloride ion	0.103	Albumin	0.060	Albumin	0.049
5	ALT	−0.240	Apolipoprotein A1	0.101	Choline esterase	0.053	Atherosclerosis	0.048
6	Platelet count	−0.228	Age	0.072	Apolipoprotein A1	0.046	White blood cell	0.046
7	Male	0.212	HDL-C	0.050	Chloride ion	0.036	Apolipoprotein A1	0.044
8	P-LCR	−0.201	Atherosclerosis	0.042	HDL-C	0.035	Apolipoprotein B	0.038
9	Platelet hematocrit	0.192	ALT	0.035	Sodium ion	0.346	Male	0.036
10	Atherosclerosis	0.169	Alkaline phosphatase	0.032	Prealbumin	0.034	Potassium ion	0.033
11	Potassium ion	−0.141	PDW	0.030	Potassium ion	0.033	Chloride ion	0.032
12	Choline esterase	0.126	Prealbumin	0.017	Alkaline phosphatase	0.030	Diabetes mellitus	0.030
13	Chloride ion	−0.121	Platelet hematocrit	0.016	ALT	0.030	Drinking	0.030
14	PDW	0.095	Sodium ion	0.016	Total cholesterol	0.029	Prealbumin	0.029
15	HDL-C	−0.081	LDL-C	0.008	Apolipoprotein B	0.029	Sodium ion	0.029
16	Diabetes mellitus	0.059	Triglyceride	0.006	Platelet count	0.028	PDW	0.029
17	Alkaline phosphatase	0.050	Smoking	0.000	P-LCR	0.028	Smoking	0.028
18	Smoking	0.048	Apolipoprotein B	0.000	LDL-C	0.028	ALT	0.028
19	Prealbumin	−0.046	RDW-CV	0.000	Direct bilirubin	0.027	HDL-C	0.028
20	Sodium ion	−0.024	Platelet count	0.000	RDW-CV	0.026	P-LCR	0.027
21	Apolipoprotein A1	−0.023	Potassium ion	0.000	Triglyceride	0.026	LDL-C	0.026
22	LDL-C	0.017	AST	0.000	AST	0.025	RDW-CV	0.026
23	Direct bilirubin	−0.013	Direct bilirubin	0.000	PDW	0.024	Total cholesterol	0.026
24	Apolipoprotein B	−0.011	Total cholesterol	0.000	Atherosclerosis	0.020	Platelet count	0.025
25	Drinking	0.011	Diabetes mellitus	0.000	Platelet hematocrit	0.019	AST	0.025
26	Triglyceride	−0.006	P-LCR	0.000	Male	0.009	Alkaline phosphatase	0.025
27	AST	0.000	Male	0.000	Smoking	0.005	Platelet hematocrit	0.025
28	Total cholesterol	0.000	Choline esterase	0.000	Diabetes mellitus	0.004	Triglyceride	0.025
29	RDW-CV	0.000	Drinking	0.000	Drinking	0.003	Direct bilirubin	0.023

RI: Relative importance; ALT: Alanine aminotransferase; P-LCR: Platelet-large cell rate; PDW: Platelet distribution width; HDL-C: High density lipoprotein-cholesterol; LDL-C: Low density lipoprotein-cholesterol; AST: Aspartate aminotransferase; RDW-CV: Coefficient of variation of red cell volume distribution width.

3　讨　论

原发性高血压并发脑梗死是一种常见且危害性极大的疾病，其起病急骤，发病机制复杂，病情发展快，易引发死亡[9]。头颅 CT 或 MRI 检查虽可以很好地排除出血性脑卒中，但 CT 对病灶的分辨率差，而 MRI 虽然可以清晰地显示病灶，但费用较高、检查时间较长，两者均可能使患者错过最佳治疗时间[10]。MRI 弥散加权成像可显示发病 2 h 内缺血病变，但是弥散加权成像对磁场的匀场要求较高，且靠近骨组织的脑内病变会出现伪影[11]。数字减影血管造影检查在发现血管狭窄、闭塞及其他血管病变方面准确性高，是脑血管病变检查的金标准，但存在费用高、有创、技术条件要求高等缺点[3]，患者在患病早期常常拒绝该检查，因而多在临床上出现明显脑梗死症状后才能确诊脑梗死，无法在早期筛查原发性高血压患者的脑梗死患病风险。因此，建立预测模型尽早对原发性高血压并发脑梗死进行风险评估对于该病的预防具有积极意义。

本研究结合患者的一般资料、病因、实验室

检查结果，采用logistic回归、决策树、随机森林及XGBoost算法构建了4个预测模型。经Delong检验显示，随机森林和XGBoost模型的预测性能优于logistic回归和决策树模型，其中综合预测效能最优的是XGBoost模型，其AUC值最高，为0.808（95% *CI* 0.804～0.811）。XGBoost算法是经过优化的分布式梯度提升库，具有高效、高灵活性、可移植性等优点，因此能得到更好的预测精度。

众所周知，血细胞比容是血液黏度的主要决定因素，在调节脑血流量中起着重要的作用。本研究中二元logistic回归结果表明，较低水平的血细胞比容提示原发性高血压患者发生脑梗死的可能性更高。有研究表明，当原发性高血压患者的血细胞比容＞50%时，会导致血液黏度和外周阻力增加，脑血流量减少，梗死范围增加；当血细胞比容降低到30%以下时，脑血流量增加，葡萄糖和能量代谢严重受损，脑供氧不足或贫血性缺氧与缺血叠加，会导致更严重的脑代谢损害[12-13]。白蛋白作为一种独特的多功能蛋白，具有抗氧化和抗炎特性，可以抑制血小板聚集，对神经元具有保护作用。本研究发现白蛋白是原发性高血压患者并发脑梗死的一个独立保护因素。低白蛋白血症不仅会导致血管腔内液体流失，还会导致血小板聚集[14]。一项前瞻性研究发现，相较于血清白蛋白水平较低的高血压患者，有较高血清白蛋白的高血压患者发生脑梗死的风险更低、脑梗死预后效果更好[15]。还有研究发现，血清白蛋白被证明是首次和复发性缺血性脑卒中临床结局的预测因子[16]。梁明月和赵会民[17]研究发现，高龄是原发性高血压患者发生脑梗死的重要影响因素。炎症反应参与脑梗死的所有阶段，可导致缺血性损伤的发展和神经功能的恶化。白细胞计数作为反映系统性炎症的标志物，其升高提示脑梗死患者的缺血性损伤加重和梗死体积增大，常可作为脑梗死预后不良的危险因素[18]。Möller等[19]研究发现当原发性高血压并发脑梗死时，循环髓系白细胞数量和活化状态增加，吸引白细胞的趋化因子水平增高，从而导致白细胞计数增加。这表明原发性高血压患者白细胞计数增加提示其可能有更高的脑梗死患病风险。本研究中随机森林和XGBoost模型均提示，胆碱酯酶、载脂蛋白A1是原发性高血压并发脑梗死的重要影响因素。胆碱酯酶是一种可催化酰基胆碱水解的糖蛋白，能反映肝脏合成能力，也是全身性炎症和营养不良的影响因子[20]。当原发性高血压患者并发脑梗死或出现缺血缺氧、全身炎症反应时，会导致肝脏合成功能下降，发生应激反应而诱发高分解代谢，从而使胆碱酯酶消耗增多；同时，炎症反应能激活胆碱能抗炎通路，导致乙酰胆碱代偿性增加，从而抑制胆碱酯酶的活性，最终导致脑梗死进展[21]。载脂蛋白A1主要存在高密度脂蛋白胆固醇中，能调节炎症反应，其检测结果不受甘油三酯水平影响，可反映高密度脂蛋白胆固醇的水平[22]。较低的载脂蛋白A1浓度提示可能有更高的颈动脉内膜中层厚度，表明原发性高血压患者有更高的脑梗死患病风险[23]。

本研究存在以下局限性：（1）本研究在指标选取时删除了缺失率≥30%的指标（如血清同型半胱氨酸、尿酸），这些指标可能存在潜在价值，有待数据量扩大后进一步分析。（2）本研究虽然基于真实世界的数据，但仅仅是对部分医院的数据进行了验证，还需要后期结合临床对更多的临床病例进行观察验证等。

综上所述，本研究筛选出原发性高血压患者并发脑梗死的危险因素中相关性较强的预测因子，构建了基于logistic回归、决策树、随机森林和XGBoost算法的原发性高血压并发脑梗死预测模型，通过准确度、F1值、AUC值等评价指标评估了4个模型的预测性能，发现XGBoost算法能够较好地预测原发性高血压患者的脑梗死患病风险，有助于临床医师尽早识别潜在的脑梗死患者，最终达到早发现、早诊断、早治疗的目的。

［参 考 文 献］

［1］ 中国心血管健康与疾病报告编写组.中国心血管健康与疾病报告2020概要［J］.中国循环杂志，2021，36：521-545.

［2］ 国家卫生健康委员会疾病预防控制局，国家心血管病中心，中国医学科学院阜外医院，中国疾病预防控制中心，中华医学会心血管病学分会，中国医师协会高血压专业委员会，等.中国高血压健康管理规范（2019）［J］.中华心血管病杂志，2020，48：10-46.

［3］ 中华医学会神经病学分会，中华医学会神经病学分会脑血管病学组.中国急性缺血性脑卒中诊治指南2018［J］.中华神经科杂志，2018，51：666-682.

［4］ 李齐光，陈培松，梁雅茹，钟国权.原发性高血压并发脑梗死的危险因素分析［J］.中国当代医药，2020，27：12-15，20.

[5] CANTONE M, LANZA G, PUGLISI V, VINCIGUERRA L, MANDELLI J, FISICARO F, et al. Hypertensive crisis in acute cerebrovascular diseases presenting at the emergency department: a narrative review[J/OL]. Brain Sci, 2021, 11: 70. DOI: 10.3390/brainsci11010070.

[6] MA J, SHEN L K, BAO L, YUAN H, WANG Y X, LIU H, et al. A novel prognosis prediction model, including cytotoxic T lymphocyte-associated antigen-4, ischemia-modified albumin, lipoprotein-associated phospholipase A2, glial fibrillary acidic protein, and homocysteine, for ischemic stroke in the Chinese hypertensive population[J/OL]. J Clin Lab Anal, 2021, 35: e23756. DOI: 10.1002/jcla.23756.

[7] YANG Y J, ZHENG J, DU Z Z, LI Y, CAI Y P. Accurate prediction of stroke for hypertensive patients based on medical big data and machine learning algorithms: retrospective study[J/OL]. JMIR Med Inform, 2021, 9: e30277. DOI: 10.2196/30277.

[8] 《中国高血压防治指南》修订委员会. 中国高血压防治指南 2018 年修订版[J]. 心脑血管病防治, 2019, 19: 1-44.

[9] 杨翠, 樊凡, 王庆松. 缺血性脑卒中患者急性高血压反应与脑卒中后认知功能障碍的相关性研究[J]. 中华老年心脑血管病杂志, 2018, 20: 1023-1026.

[10] 牛雁军. CT 和MRI 对急性缺血性脑卒中的诊断效果对比分析[J]. 临床医药文献电子杂志, 2020, 7: 136, 145.

[11] 张文博. DWI 与MRA 联合检测在急性脑梗死患者血管病变评估及临床意义[J]. 中国CT 和MRI 杂志, 2018, 16: 1-4.

[12] MELLER A, GOLAB-JANOWSKA M, PACZKOWSKA E, MACHALINSKI B, PAWLUKOWSKA W, NOWACKI P. Reduced hemoglobin levels combined with an increased plasma concentration of vasoconstrictive endothelin-1 are strongly associated with poor outcome during acute ischemic stroke[J]. Curr Neurovascular Res, 2018, 15: 193-203.

[13] STAVROPOULOS K, IMPRIALOS K P, BOULOUKOU S, BOUTARI C, DOUMAS M. Hematocrit and stroke: a forgotten and neglected link?[J]. Semin Thromb Hemost, 2017, 43: 591-598.

[14] 刘明苏, 汤也, 付庆, 李光勤. 急性脑梗死非溶栓患者出血转化的影响因素分析[J]. 中国当代医药, 2020, 27: 16-20.

[15] WANG C Y, DENG L H, QIU S, BIAN H Y, WANG L, LI Y X, et al. Serum albumin is negatively associated with hemorrhagic transformation in acute ischemic stroke patients[J]. Cerebrovasc Dis Basel Switz, 2019, 47: 88-94.

[16] KIBOSHI R, SATOH S, MIKAMI K, KITAJIMA M, URUSHIZAKA M, METOKI N, et al. Serum albumin, body mass index, and preceding Ⅹa and P2Y12 inhibitors predict prognosis of recurrent ischemic stroke[J/OL]. J Stroke Cerebrovasc Dis, 2021, 30: 105681. DOI: 10.1016/j.jstrokecerebrovasdis.2021.105681.

[17] 梁明月, 赵会民. 单核细胞/高密度脂蛋白胆固醇比值诊断高血压并发无症状脑梗死的价值[J]. 实用医学杂志, 2019, 35: 2645-2648.

[18] QUAN K H, WANG A X, ZHANG X L, WANG Y J. Leukocyte count and adverse clinical outcomes in acute ischemic stroke patients[J/OL]. Front Neurol, 2019, 10: 1240. DOI: 10.3389/fneur.2019.01240.

[19] MÖLLER K, PÖSEL C, KRANZ A, SCHULZ I, SCHEIBE J, DIDWISCHUS N, et al. Arterial hypertension aggravates innate immune responses after experimental stroke[J/OL]. Front Cell Neurosci, 2015, 9: 461. DOI: 10.3389/fncel.2015.00461.

[20] LI M Q, CHEN Y, ZHANG Y L, LIU X Y, XIE T T, YIN J J, et al. Admission serum cholinesterase concentration for prediction of in-hospital mortality in very elderly patients with acute ischemic stroke: a retrospective study[J]. Aging Clin Exp Res, 2020, 32: 2667-2675.

[21] 张峰, 王卫国, 谢燕, 黄芬. 血清胆碱酯酶水平与急性脑梗死患者病情严重程度及预后的关系研究[J]. 实用心脑肺血管病杂志, 2017, 25: 24-27.

[22] 王旭颖, 王永红. 血清脂蛋白(a)载脂蛋白A1 载脂蛋白B 与脑梗死相关性研究[J]. 现代医药卫生, 2017, 33: 988-991.

[23] ZIVANOVIC Z, DIVJAK I, JOVICEVIC M, RABI-ZIKIC T, RADOVANOVIC B, RUZICKA-KALOCI S, et al. Association between apolipoproteins AⅠ and B and ultrasound indicators of carotid atherosclerosis[J]. Curr Vasc Pharmacol, 2018, 16: 376-384.

[本文编辑]　杨亚红