DOI:10.16781/j.0258-879x.2021.10.1115

·论　著·

基于机器学习算法的新型冠状病毒肺炎患者院内结局预测

彭　驰[1]，齐戈尧[1]，张晨旭[1]，郭玉峰[2]，金志超[1*]
1. 海军军医大学（第二军医大学）卫生勤务学系卫生统计学教研室，上海 200433
2. 海军军医大学（第二军医大学）长征医院医务处，上海 200003

［摘要］ **目的**　利用机器学习算法构建新型冠状病毒肺炎（COVID-19）患者临床结局的预测模型，并探索结局相关因子。**方法**　收集2020年2月5日至4月15日武汉市火神山医院及华中科技大学同济医学院附属同济医院光谷院区收治的COVID-19患者的临床指标与结局（院内死亡和院内接受气管插管治疗），利用人工神经网络（ANN）、朴素贝叶斯、logistic回归、随机森林4种机器学习算法构建患者临床结局的预测模型。**结果**　共纳入4 804例COVID-19患者，其中发生院内死亡100例（2.08%）、接受气管插管治疗87例（1.81%）。与院内死亡相关性最强的变量为白细胞计数、白蛋白、钙离子、血尿素氮、心肌型肌酸激酶同工酶和年龄，与院内接受气管插管治疗相关性最强的变量为白细胞计数、淋巴细胞绝对值、超敏CRP、总胆红素、钙离子和年龄，分别利用以上变量、基于4种机器学习算法构建院内死亡和院内接受气管插管治疗预测模型。4种预测模型中，相较于基于ANN、logistic回归、随机森林算法构建的模型［预测院内死亡的AUC值（95% *CI*）分别为0.938（0.882～0.993）、0.926（0.865～0.987）、0.867（0.780～0.954），预测院内接受气管插管治疗的AUC值（95% *CI*）分别为0.932（0.814～0.980）、0.935（0.817～0.981）、0.936（0.921～0.972）］，基于朴素贝叶斯算法构建的模型在预测COVID-19患者院内死亡（*AUC*=0.952，95% *CI* 0.925～0.979）和接受气管插管治疗（*AUC*=0.948，95% *CI* 0.896～0.965）方面性能均最佳。**结论**　4种机器学习算法在预测COVID-19患者临床结局方面性能良好，其中以基于朴素贝叶斯算法构建的预测模型最佳。白细胞计数、白蛋白、钙离子、血尿素氮、心肌型肌酸激酶同工酶和年龄可以用来预测COVID-19患者院内死亡，白细胞计数、淋巴细胞绝对值、超敏CRP、总胆红素、钙离子和年龄可以用来预测患者院内是否接受气管插管治疗。

［关键词］ 机器学习；算法；新型冠状病毒肺炎；医院死亡率；气管内插管
［中图分类号］ R 511；R 563.12　［文献标志码］ A　［文章编号］ 0258-879X(2021)10-1115-09

Prediction of in-hospital clinical outcomes of coronavirus disease 2019 patients based on machine learning algorithms

PENG Chi[1], QI Ge-yao[1], ZHANG Chen-xu[1], GUO Yu-feng[2], JIN Zhi-chao[1*]
1. Department of Health Statistics, Faculty of Health Services, Naval Medical University (Second Military Medical University), Shanghai 200433, China
2. Medical Affair Office, Changzheng Hospital, Naval Medical University (Second Military Medical University), Shanghai 200003, China

［**Abstract**］ **Objective**　To construct prediction models for the clinical outcomes of coronavirus disease 2019 (COVID-19) patients using machine learning algorithms, and explore the outcome-related factors. **Methods**　The clinical indexes and outcomes (in-hospital mortality and receiving tracheal intubation) of COVID-19 patients who were admitted to Wuhan Huoshenshan Hospital or Guanggu Branch of Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology from Feb. 5 to Apr. 15, 2020 were collected. The prediction models for the clinical outcomes were constructed using artificial neural network (ANN), naive Bayes, logistic regression and random forest algorithms. **Results**　A total of 4 804 COVID-19 patients were included, of whom 100 (2.08%) patients died and 87 (1.81%) patients received tracheal intubation during the hospitalization. White blood cell (WBC), albumin, calcium, blood urea nitrogen, creatine kinase-myocardial band (CK-MB) and age were the most correlated variables with in-hospital mortality. WBC, lymphocyte, hypersensitivity C reaction protein (hs-CRP), total bilirubin, calcium and age were the most correlated variables

［收稿日期］ 2021-05-07　［接受日期］ 2021-07-07
［基金项目］ 上海市公共卫生体系建设三年行动计划学科建设项目（GWV-10.1-XK05），海军军医大学（第二军医大学）“三航”计划. Supported by Discipline Construction Program of Shanghai 3-Year Action Plan for Public Health System Construction (GWV-10.1-XK05) and “San Hang” Program of Naval Medical University (Second Military Medical University).
［作者简介］ 彭　驰，硕士生. E-mail: pengchi2019@126.com
* 通信作者（Corresponding author）. Tel: 021-81871442, E-mail: jinzhichao@smmu.edu.cn

with in-hospital tracheal intubation. With the above variables and based on the 4 machine learning algorithms, the prediction models for in-hospital mortality and tracheal intubation were constructed. In the 4 prediction models, the model constructed based on naive Bayes algorithm had the best performance in predicting in-hospital mortality (area under curve［*AUC*］=0.952, 95% confidence interval［*CI*］0.925-0.979) and tracheal intubation (*AUC*=0.948, 95% *CI* 0.896-0.965) versus the models constructed based on ANN, logistic regression and random forest algorithms (the *AUC*［95% *CI*］values for predicting in-hospital mortality were 0.938［0.882-0.993］, 0.926［0.865-0.987］and 0.867［0.780-0.954］, and the *AUC*［95% *CI*］values for predicting in-hospital tracheal intubation were 0.932［0.814-0.980］, 0.935［0.817-0.981］and 0.936［0.921-0.972］, respectively). **Conclusion** The 4 machine learning algorithms have good performance in predicting the clinical outcomes of COVID-19 patients. WBC, albumin, calcium, blood urea nitrogen, CK-MB and age can be used to predict the in-hospital mortality of COVID-19 patients; while WBC, lymphocyte count, hs-CRP, total bilirubin, calcium and age can be used to predict the in-hospital tracheal intubation.

［**Key words**］ machine learning; algorithms; coronavirus disease 2019; hospital mortality; intertracheal intubation

［Acad J Sec Mil Med Univ, 2021, 42(10): 1115-1123］

由严重急性呼吸综合征冠状病毒 2（severe acute respiratory syndrome coronavirus 2，SARS-CoV-2）引起的新型冠状病毒肺炎（coronavirus disease 2019，COVID-19）在全球范围内蔓延，截至 2021 年 3 月 14 日，已造成 11 921 万人感染，264 万人死亡[1]。不断增加的感染人数给全球医疗卫生服务造成了巨大的压力，如何利用最简单的指标识别有死亡风险或需要气管插管机械通气的 COVID-19 患者并给予早期干预，成为当前亟须解决的问题。

目前已有多项预测 COVID-19 患者院内死亡的研究[2-8]，这些研究的纳入指标和方法各异（表 1），如 Kaufmann 等[2] 利用 COVID-19 患者的入院时心房利钠肽构建预测模型预测患者死亡率，ROC 曲线的 AUC 值为 0.832；Zhang 等[3] 利用 COVID-19 患者的入院时临床指标和胸部 CT 表现评估患者预后。但现有的预测模型主要存在以下问题：（1）部分研究用于构建预测模型的样本量不大；（2）部分研究确定的预测因子在临床中无法快速获取。

表 1　COVID-19 患者院内死亡预测的文献回顾

Tab 1　Literature review of predicting in-hospital mortality in patients with COVID-19

Study	Predictive factor	Method	*n*	AUC
Kaufmann, et al[2]	Mid-regional pro-ANP	Cox regression	213	0.832
Zhang, et al[3]	CT features, age, LDH, diarrhea	XGBoost	198	0.924
Zhang, et al[4]	D-dimer	Cox proportional hazard model	343	0.890
Bertsimas, et al[5]	Age, oxygen saturation, CRP, BUN, creatinine	XGBoost	3 927	0.920
Zhao, et al[6]	Heart failure, procalcitonin, LDH, COPD, oxygen saturation, heart rate, age	Logistic regression model	641	0.830
Utrero-Rico, et al[7]	NLR, LDH, IL-6, age, PaO_2/FiO_2	Logistic regression model	1 477	0.910
Li, et al[8]	Age, severity at admission, dyspnea, cardiovascular disease, LDH, TBil, glucose, urea	Logistic regression model	4 086	0.920

COVID-19: Coronavirus disease 2019; AUC: Area under curve; pro-ANP: Pro-atrial natriuretic peptide; CT: Computed tomography; LDH: Lactate dehydrogenase; CRP: C reactive protein; BUN: Blood urea nitrogen; COPD: Chronic obstructive pulmonary disease; NLR: Neutrophil-to-lymphocyte ratio; IL-6: Interleukin 6; PaO_2: Arterial partial pressure of oxygen; FiO_2: Fraction of inspired oxygen; TBil: Total bilirubin.

本研究基于 4 种不同的机器学习算法，主要利用大样本 COVID-19 患者入院时的常规检查指标构建模型用于预测患者的临床结局，以帮助临床医师用常规检查指标快速评估COVID-19患者的预后并及时干预。

1　资料和方法

1.1　*数据来源*　选择 2020 年 2 月 5 日至 4 月 15 日收治的 4 804 例 COVID-19 患者为研究对象，其中 3 040 例来自武汉市火神山医院，1 764 例来自华中科技大学同济医学院附属同济医院光谷院区。所有入选患者均依据国家卫生健康委员会发布的《新型冠状病毒感染的肺炎诊疗方案（试行第五版）》[9] 诊断为 COVID-19。从患者的电子病历中提取人口学信息、入院相关实验室检查结果、合并症及临床结局等，并将信息缺失患者比例＞20% 的指标排除。

本研究获得海军军医大学（第二军医大学）伦理委员会审批。为了保护患者隐私，本研究在从医院病历系统中收集数据时隐去患者姓名、住址等识别信息。

1.2　纳入指标　本研究纳入了患者人口学信息，如年龄、性别；既往疾病史，如高血压、糖尿病、冠心病、慢性阻塞性肺疾病、肾脏疾病、恶性肿瘤；首次入院时基本检查指标，如血压、呼吸频率、脉率；临床症状，如咳嗽、乏力、发热、咳痰、喘息；首次入院时血常规检查，如白细胞计数、淋巴细胞绝对值、单核细胞绝对值、中性粒细胞绝对值、嗜酸性粒细胞绝对值、嗜碱性粒细胞绝对值、红细胞计数、血红蛋白、血细胞比容、平均红细胞体积、平均红细胞血红蛋白量、血小板计数、血小板平均体积、超敏 CRP；入院肝功能检查，如丙氨酸转氨酶、天冬氨酸转氨酶、碱性磷酸酶、总蛋白、白蛋白、尿素、总胆红素、直接胆红素、二氧化碳结合力、总胆汁酸；入院肾功能检查，如血肌酐、血尿素氮、尿酸、钠离子、钾离子、钙离子、氯离子等。

1.3　临床结局　不良临床结局指标包括院内死亡和院内接受气管插管治疗，均从临床病历中获取。

1.4　统计学处理　采用人工神经网络（artificial neural network，ANN）、朴素贝叶斯（naive Bayes）、logistic 回归、随机森林 4 种机器学习算法进行 COVID-19 患者临床结局预测建模。将火神山医院患者数据作为训练集，华中科技大学同济医学院附属同济医院光谷院区患者数据作为测试集。由于训练集中结局变量的阳性与阴性值不平衡（2 972 例存活、68 例院内死亡；2 982 例未接受气管插管治疗、58 例接受气管插管治疗），因此采用过采样法平衡阳性与阴性的比例[10-11]。最终，将是否发生院内死亡的比例由 43.7 ∶ 1 转化为 50 ∶ 50，过采样后样本量为 5 331 例；将是否接受气管插管治疗的比例由 51.4 ∶ 1 转化为 50 ∶ 50，过采样后样本量为 5 922 例，主要过程如图 1 所示。

不均衡分类是一种有监督的学习[10-11]，其处理的数据中一种分类的占比远远大于其他分类，这在二分类数据中尤为常见。在处理不均衡分类数据时，由于无法从样本量小的分类中获取足够的信息，可能会造成算法不稳定，导致预测结果出现偏倚。过采样法是一种处理不均衡数据的方法，采用过采样法根据分类较少的样本的规律生成过多的该分类的样本，使数据趋于均衡。

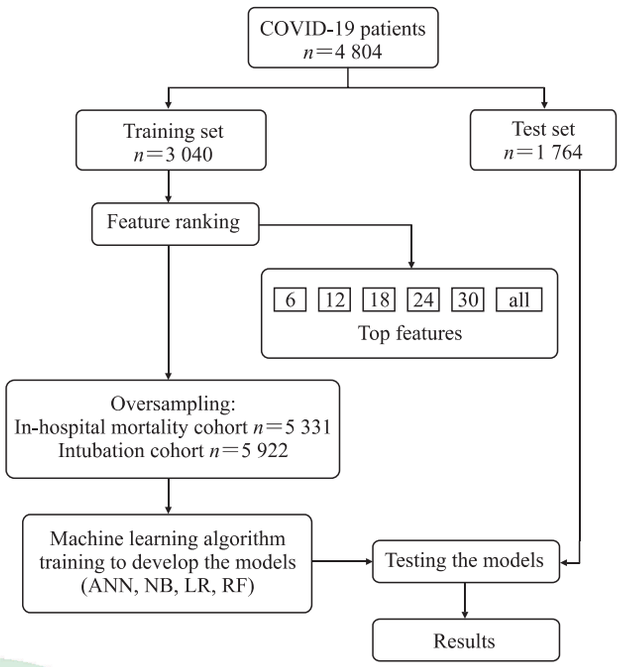

图 1　数据训练与测试流程图

Fig 1　Flow chart of data training and testing

COVID-19: Coronavirus disease 2019; ANN: Artificial neural network; NB: Naive Bayes; LR: Logistic regression; RF: Random forest.

ANN 算法类似生物神经系统，依赖于节点和连接工作[12]。ANN 算法通过反向传播算法调整权值来减小损失函数的误差值，这是监督学习算法的标准过程。朴素贝叶斯算法是一种基于贝叶斯定理的简单分类器[13]，该算法首先从训练集中估计出每个变量的条件概率，然后根据实际结果不断修正该概率，最终找到后验概率最大的输出。logistic 回归算法是以某一事件发生的概率为因变量、因变量的影响因素为自变量建立的回归模型，该算法通过 sigmoid 函数预测各类别的概率。随机森林算法是使用多棵树训练和预测样本的分类器，该算法由几棵相互独立的决策树组成[14]。

本研究采用特征排序法确定对临床结局影响最显著的变量。以 6/12/18/24/30/all 的分组将特征值进行排序，利用递归特征消除法选择最优的变量组合。4 种机器学习算法在训练数据集时，参数设置均为十折重复交叉验证，重复 3 次。所有这些机器学习算法均通过灵敏度、特异度、精确度、AUC 值进行评估。

本研究使用 R 4.0.2 和 SAS 9.4 软件进行统计学分析。呈正态分布的计量资料以 $\bar{x}\pm s$ 表示，偏态分布的计量资料以中位数（下四分位数，上四

分位数）表示，分别采用独立样本 t 检验和 Mann-Whitney U 检验进行组间数据的比较。计数资料以例数和百分数表示，采用 χ^2 检验或 Fisher 确切概率法进行组间数据的比较。对于数据中存在的缺失值，采用多重插补的方法处理（数据缺失情况见表 2）。检验水准（α）为 0.05。

表 2　全组病例的数据缺失情况

Tab 2　Missing data of whole group of cases

N=4 804, n (%)

Variable	Missing	Variable	Missing
WBC	54 (1.12)	AST	140 (2.91)
Lymphocyte	54 (1.12)	ALP	146 (3.04)
Monocyte	54 (1.12)	TP	144 (3.00)
Neutrophil	54 (1.12)	Albumin	145 (3.02)
Eosinophil	54 (1.12)	TBil	145 (3.02)
Basophil	55 (1.14)	DBil	158 (3.29)
RBC	54 (1.12)	CO_2	278 (5.79)
Hemoglobin	54 (1.12)	TBA	177 (3.68)
Hematocrit	54 (1.12)	Na^+	281 (5.85)
MCV	54 (1.12)	K^+	286 (5.95)
MCH	54 (1.12)	Ca^{2+}	286 (5.95)
Platelet	55 (1.14)	Cl^-	277 (5.77)
MPV	54 (1.12)	Creatinine	205 (4.27)
hs-CRP	397 (8.26)	UA	209 (4.35)
ALT	150 (3.12)	CK-MB	824 (17.15)

WBC: White blood cell; RBC: Red blood cell; MCV: Mean corpuscular volume; MCH: Mean corpuscular hemoglobin; MPV: Mean platelet volume; hs-CRP: Hypersensitivity C reactive protein; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; ALP: Alkaline phosphatase; TP: Total protein; TBil: Total bilirubin; DBil: Direct bilirubin; TBA: Total bile acid; UA: Uric acid; CK-MB: Creatine kinase-myocardial band.

2　结　果

2.1　COVID-19 患者基线特征　4 804 例 COVID-19 患者中 87 例（1.81%）于住院期间接受气管插管治疗，4 717 例住院期间未行气管插管治疗；100 例（2.08%）于出院前死亡，4 704 例出院时存活。分析 COVID-19 患者的基线特征结果（表 3）显示，与未发生不良临床结局的患者相比，院内死亡与接受气管插管治疗的患者年龄较大，男性占比较高，入院时呼吸频率较快，舒张压较低，较多有糖尿病、冠心病、慢性阻塞性肺疾病、肾脏疾病史，白细胞计数、中性粒细胞绝对值、血小板平均体积、超敏 CRP、天冬氨酸转氨酶、碱性磷酸酶、尿素、总胆红素、直接胆红素、血肌酐、血尿素氮、心肌型肌酸激酶同工酶水平均较高，而淋巴细胞绝对值、嗜酸性粒细胞绝对值、嗜碱性粒细胞绝对值、红细胞计数、血红蛋白、血细胞比容、血小板计数、总蛋白、白蛋白、尿酸、钙离子水平均较低，差异均有统计学意义（P 均＜0.05）；此外，与未发生不良临床结局的患者相比，院内死亡患者出现咳嗽症状较少、单核细胞绝对值和 CO_2 水平均较低，院内接受气管插管治疗的患者丙氨酸转氨酶水平较高、血清氯离子水平较低、有恶性肿瘤病史者较多，差异均有统计学意义（P 均＜0.05）。

表 3　不同临床结局 COVID-19 患者的基线特征

Tab 3　Baseline characteristics of COVID-19 patients with different outcomes

Characteristic	In-hospital mortality		In-hospital tracheal intubation	
	Alive N=4 704	Died N=100	Non-intubation N=4 717	Intubation N=87
General condition				
Age, n (%)				
＜18 years	21 (0.45)	0	21 (0.45)	0
18-44 years	900 (19.13)	1 (1.00)	901 (19.10)	0
45-59 years	1 507 (32.04)	12 (12.00)	1507 (31.95)	12 (13.79)
≥60 years	2 276 (48.38)	87 (87.00)	2 288 (48.51)	75 (86.21)
Male, n (%)	2 220 (47.19)	66 (66.00)	2 226 (47.19)	60 (68.97)
RR/min^{-1}, M (Q_L, Q_U)	20.0 (19.0, 20.0)	21.0 (20.0, 24.0)	20.0 (19.0, 20.0)	21.0 (19.0, 23.0)
Pulse/min^{-1}, M (Q_L, Q_U)	84.0 (78.0, 94.0)	88.0 (78.0, 98.0)	84.0 (78.0, 94.0)	88.0 (78.0, 98.0)
SBP/mmHg, M (Q_L, Q_U)	130 (120, 140)	130 (120, 144)	130 (120, 140)	131 (120, 143)
DBP/mmHg, M (Q_L, Q_U)	80 (75, 89)	76 (67, 86)	80 (75, 89)	78 (68, 86)
Clinical symptom, n (%)				
Cough	2 717 (57.76)	45 (45.00)	2 717 (57.60)	45 (51.72)
Fatigue	1 651 (35.10)	38 (38.00)	1 657 (35.13)	32 (36.78)
Fever	2 917 (62.01)	61 (61.00)	2 924 (61.99)	54 (62.07)
Sputum	243 (5.17)	4 (4.00)	243 (5.15)	4 (4.60)
Gasp	719 (15.28)	20 (20.00)	722 (15.31)	17 (19.54)

续表 3

Characteristic	In-hospital mortality		In-hospital tracheal intubation	
	Alive N=4 704	Died N=100	Non-intubation N=4 717	Intubation N=87
Laboratory test, M (Q_L, Q_U)				
WBC/(L^{-1}, $\times10^9$)	5.70 (4.70, 6.90)	8.50 (6.45, 12.10)	5.70 (4.70, 6.90)	8.80 (6.65, 12.80)
Lymphocyte/(L^{-1}, $\times10^9$)	1.53 (1.16, 1.90)	0.64 (0.42, 0.93)	1.53 (1.16, 1.90)	0.74 (0.42, 1.02)
Monocyte/(L^{-1}, $\times10^9$)	0.41 (0.32, 0.52)	0.36 (0.24, 0.56)	0.41 (0.32, 0.52)	0.36 (0.24, 0.60)
Neutrophil/(L^{-1}, $\times10^9$)	3.44 (2.67, 4.47)	7.25 (5.30, 10.7)	3.44 (2.67, 4.47)	7.25 (5.35, 11.90)
Eosinophil/(L^{-1}, $\times10^9$)	0.11 (0.07, 0.19)	0.02 (0.01, 0.06)	0.11 (0.07, 0.19)	0.02 (0.01, 0.06)
Basophil/(L^{-1}, $\times10^9$)	0.02 (0.01, 0.03)	0.01 (0.01, 0.01)	0.02 (0.01, 0.03)	0.01 (0.01, 0.02)
RBC/(L^{-1}, $\times10^{12}$)	4.07 (3.73, 4.41)	3.90 (3.39, 4.23)	4.07 (3.73, 4.41)	3.96 (3.37, 4.25)
Hemoglobin/($g•L^{-1}$)	126.0 (116.0, 137.0)	117.0 (102.0, 132.0)	126.0 (116.0, 137.0)	121.0 (104.0, 134.0)
Hematocrit/%	37.30 (34.50, 40.30)	35.20 (31.10, 38.50)	37.30 (34.50, 40.30)	36.00 (31.40, 39.30)
MCV/fL	92.30 (89.60, 94.90)	92.30 (88.00, 96.70)	92.30 (89.60, 94.90)	92.50 (89.00, 97.20)
MCH/pg	31.20 (30.20, 32.20)	31.00 (29.70, 32.60)	31.20 (30.20, 32.20)	31.40 (29.80, 32.50)
Platelet/(L^{-1}, $\times10^9$)	221.0 (181.0, 270.0)	152.0 (88.2, 230.0)	221.0 (181.0, 270.0)	159.0 (96.2, 249.0)
MPV/fL	9.80 (9.20, 10.60)	10.50 (9.75, 11.40)	9.90 (9.20, 10.60)	10.30 (9.65, 11.40)
hs-CRP/($mg•L^{-1}$)	1.59 (0.63, 3.77)	12.10 (6.26, 84.60)	1.59 (0.63, 3.77)	16.10 (6.66, 84.00)
ALT/($U•L^{-1}$)	21.60 (14.00, 36.10)	24.00 (15.70, 44.60)	21.60 (14.00, 36.00)	26.00 (16.90, 49.80)
AST/($U•L^{-1}$)	18.40 (14.40, 25.10)	30.60 (21.50, 47.40)	18.40 (14.40, 25.10)	27.10 (19.80, 42.20)
ALP/($U•L^{-1}$)	69.80 (58.00, 84.00)	89.60 (68.80, 117.00)	69.80 (58.00, 83.90)	91.20 (69.00, 116.00)
TP/($g•L^{-1}$)	67.00 (62.70, 71.50)	60.20 (55.40, 65.40)	67.00 (62.60, 71.40)	61.00 (56.10, 66.10)
Albumin/($g•L^{-1}$)	38.30 (35.40, 40.70)	30.50 (27.10, 33.40)	38.30 (35.40, 40.70)	31.40 (27.50, 34.90)
Creatinine/($\mu mol•L^{-1}$)	64.10 (55.00, 75.60)	76.40 (62.40, 105.00)	64.10 (55.10, 75.80)	73.80 (58.90, 90.90)
TBil/($\mu mol•L^{-1}$)	9.40 (7.30, 12.40)	13.00 (9.55, 19.50)	9.40 (7.30, 12.40)	12.60 (10.10, 18.90)
DBil/($\mu mol•L^{-1}$)	3.50 (2.60, 4.70)	6.90 (4.47, 10.50)	3.50 (2.60, 4.70)	6.45 (4.43, 9.52)
CO_2/($mmol•L^{-1}$)	24.10 (22.70, 25.60)	22.10 (19.60, 25.90)	24.10 (22.70, 25.60)	23.60 (20.60, 26.60)
TBA/($\mu mol•L^{-1}$)	3.80 (2.40, 6.10)	4.30 (2.58, 6.82)	3.80 (2.40, 6.10)	4.00 (2.70, 6.88)
BUN/($mmol•L^{-1}$)	4.41 (3.62, 5.46)	8.46 (5.29, 12.40)	4.41 (3.62, 5.48)	6.66 (4.97, 11.20)
UA/($\mu mol•L^{-1}$)	284.0 (231.0, 347.0)	262.0 (192.0, 337.0)	284.0 (231.0, 348.0)	230.0 (176.0, 319.0)
Na^+/($mmol•L^{-1}$)	141.0 (139.0, 143.0)	140.0 (137.0, 145.0)	141.0 (139.0, 143.0)	140.0 (136.0, 144.0)
K^+/($mmol•L^{-1}$)	4.18 (3.90, 4.49)	4.30 (3.80, 4.60)	4.18 (3.90, 4.49)	4.30 (3.74, 4.60)
Ca^{2+}/($mmol•L^{-1}$)	2.17 (2.10, 2.24)	1.97 (1.87, 2.03)	2.17 (2.10, 2.23)	1.97 (1.88, 2.05)
Cl^-/($mmol•L^{-1}$)	106.0 (104.0, 108.0)	104.0 (99.0, 110.0)	106.0 (104.0, 108.0)	104.0 (99.4, 108.0)
CK-MB/($U•L^{-1}$)	7.30 (1.20, 9.70)	12.80 (8.10, 22.50)	7.30 (1.20, 9.80)	10.30 (5.62, 17.30)
Medical history, n (%)				
Hypertension	1 429 (30.38)	28 (28.00)	1 429 (30.29)	28 (32.18)
Diabetes	630 (13.39)	25 (25.00)	634 (13.44)	21 (24.14)
CHD	269 (5.72)	22 (22.00)	276 (5.85)	15 (17.24)
COPD	43 (0.91)	6 (6.00)	42 (0.89)	7 (8.05)
Kidney disease	124 (2.64)	16 (16.00)	126 (2.67)	14 (16.09)
Cancer	42 (0.89)	3 (3.00)	41 (0.87)	4 (4.60)

1 mmHg＝0.133 kPa. COVID-19: Coronavirus disease 2019; RR: Respiratory rate; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; WBC: White blood cell; RBC: Red blood cell; MCV: Mean corpuscular volume; MCH: Mean corpuscular hemoglobin; MPV: Mean platelet volume; hs-CRP: Hypersensitivity C reactive protein; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; ALP: Alkaline phosphatase; TP: Total protein; TBil: Total bilirubin; DBil: Direct bilirubin; TBA: Total bile acid; BUN: Blood urea nitrogen; UA: Uric acid; CK-MB: Creatine kinase-myocardial band; CHD: Coronary heart disease; COPD: Chronic obstructive pulmonary disease; M (Q_L, Q_U): Median (lower quartile, upper quartile).

2.2　机器学习算法预测临床结局　（1）院内死亡预测。与院内死亡相关性最强的变量为白细胞计数、白蛋白、钙离子、血尿素氮、心肌型肌酸激酶同工酶和年龄。将这 6 个变量纳入 4 种机器学习算法构建模型，在测试集中 4 种预测模型均表现出良好的效果（图 2A，表 4），ANN 算法的 AUC 值为 0.938（95% *CI* 0.882～0.993）、朴素贝叶斯算法的 AUC 值为 0.952（95% *CI* 0.925～0.979）、logistic 回归算法的 AUC 值为 0.926（95% *CI* 0.865～0.987）、随机森林算法的 AUC 值为 0.867（95% *CI* 0.780～0.954），其中用朴素贝叶斯算法构建的预测模型效能最佳。

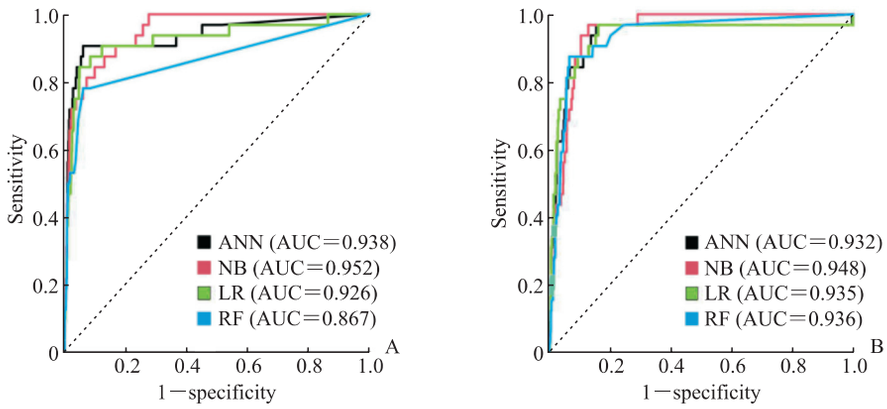

图 2　基于机器学习算法构建的 COVID-19 患者临床结局预测模型的 ROC 曲线

Fig 2　ROC curves of prediction models for clinical outcomes of COVID-19 patients based on machine learning algorithms

A: ROC curves for in-hospital mortality; B: ROC curves for receiving in-hospital tracheal intubation. COVID-19: Coronavirus disease 2019; ROC: Receiver operating characteristic; ANN: Artificial neural network; NB: Naive Bayes; LR: Logistic regression; RF: Random forest; AUC: Area under curve.

表 4　基于 4 种机器学习算法构建的 COVID-19 患者临床结局预测模型在测试集中的预测效能

Tab 4　Predictive performance of prediction models for clinical outcomes of COVID-19 patients based on 4 machine learning algorithms in test set

Outcome	Sensitivity	Specificity	Accuracy
In-hospital mortality			
ANN	0.812	0.954	0.952
NB	0.750	0.967	0.963
LR	0.750	0.954	0.951
RF	0.000	1.000	0.982
In-hospital tracheal intubation			
ANN	0.844	0.897	0.896
NB	0.938	0.887	0.888
LR	0.844	0.915	0.914
RF	0.406	0.976	0.966

COVID-19: Coronavirus disease 2019; ANN: Artificial neural network; NB: Naive Bayes; LR: Logistic regression; RF: Random forest.

（2）院内接受气管插管治疗预测。与院内接受气管插管治疗相关性最强的变量为白细胞计数、淋巴细胞绝对值、超敏 CRP、总胆红素、钙离子和年龄。将这 6 个变量纳入 4 种机器学习算法构建模型，在测试集中 4 种预测模型均表现出良好的效果（图 2B，表 4），ANN 算法的 AUC 值为 0.932（95% *CI* 0.814～0.980）、朴素贝叶斯算法的 AUC 值为 0.948（95% *CI* 0.896～0.965）、logistic 回归算法的 AUC 值为 0.935（95% *CI* 0.817～0.981）、随机森林算法的 AUC 值为 0.936（95% *CI* 0.921～0.972），其中用朴素贝叶斯算法构建的预测模型效能最佳。

3　讨　论

本研究采用 4 种机器学习算法构建 COVID-19 患者临床结局预测模型，以来自武汉市火神山医院的 COVID-19 患者数据作为训练集，以华中科技大学同济医学院附属同济医院光谷院区的患者数据作为测试集，结果显示构建的模型在预测患者院内死亡和是否接受气管插管治疗方面表现良好。本研究结果还显示白细胞计数、白蛋白、钙离子、血尿素氮、心肌型肌酸激酶同工酶和年龄这 6 个变量组合是预测 COVID-19 患者院内死亡的最佳变量组合，白细胞计数、淋巴细胞绝对值、超敏 CRP、总胆红素、钙离子和年龄这 6 个变量组合是预测 COVID-19 患者是否接受气管插管治疗的最佳变量组合。

Li 等[8] 采用 logistic 回归对与本研究同批的 COVID-19 患者进行了院内死亡预测，纳入变量包括年龄、入院时病情严重程度、呼吸困难、心血管疾病、乳酸脱氢酶、总胆红素、血糖和尿素，构建的预测模型在外部测试集中表现良好（AUC 值为 0.920，95% *CI* 0.86～0.98）。本研究将 COVID-19 患者院内死亡作为主要临床结局，接受气管插管治疗作为次要临床结局，同时侧重 COVID-19 患者入院首次实验室检查指标，采用多种机器学习算法进行对比。

血清白蛋白是血浆中最重要的蛋白质，在人体内维持稳定的营养和血浆胶体渗透压。白蛋白水平

与炎症状态关系密切，人体内的炎症反应可增加毛细血管通透性，导致血清白蛋白逸出[15]。此外，随着组织间隙的扩大，白蛋白的分布体积增大，导致白蛋白半衰期缩短。COVID-19 患者病情加重可能引发严重炎症。Li 等[16]在对 523 例 COVID-19 患者的研究中发现，白蛋白水平降低的患者住院生存时间明显缩短。此前有研究报道白蛋白与 COVID-19 患者的死亡率独立相关[17]。本研究也发现血清白蛋白水平与 COVID-19 患者院内死亡有很强的相关性。

研究表明，钙离子可以通过损害线粒体功能引起细胞炎症反应[18]。而细胞炎症反应被认为与 COVID-19 患者死亡有关[19]。细胞实验结果显示，钙通道阻滞剂能够阻止 SARS-CoV-2 的复制[20]。一项由 39 家医院联合开展的多中心研究表明，钙通道阻滞剂可以有效降低 COVID-19 患者的死亡率[21]。本研究结果表明，钙离子与 COVID-19 患者的院内死亡和气管插管治疗有很强的相关性。

年龄被认为是 COVID-19 患者病情为重型和死亡的独立危险因素[19,22]。老年患者多伴有心脑血管疾病，这些疾病可能会加重 COVID-19 的症状。本研究结果显示，年龄与 COVID-19 患者的院内死亡和接受气管插管治疗均有关。

急性心脏损伤是 COVID-19 患者报告最多的心血管疾病，发生率为 8%～12%；由病毒侵袭心肌细胞引起的直接心肌损伤和全身性炎症似乎是造成心脏损伤的最常见机制[23]。当心肌组织损伤严重时，心肌型肌酸激酶同工酶被释放入血，血清中的心肌型肌酸激酶同工酶升高即成为诊断急性心脏损伤的重要标准。Han 等[24]开展的一项基于 273 例 COVID-19 患者的研究发现，心肌型肌酸激酶同工酶浓度升高与 COVID-19 严重程度有关。本研究也证实了这一结论。

白细胞是人体内非常重要的一类血细胞，具有吞噬异物并产生抗体的能力。白细胞计数升高多见于炎症、感染等。与非重型患者相比，重型和死亡 COVID-19 患者的白细胞计数显著升高[25]。研究表明，白细胞计数升高是 COVID-19 患者死亡的独立危险因素[26]，本研究结果与此一致。

血尿素氮是蛋白质代谢的主要终末产物。既往研究发现，血尿素氮可以作为预测患者入院 48 h 后发生器官功能衰竭的指标[27]，它在评估肾功能方面也有作用。血尿素氮水平增高也是心力衰竭患者预后较差的一个预测因子[28]。Cheng 等[29]研究发现，COVID-19 患者入院时的血尿素氮水平与患者的死亡率有关，可以作为重型 COVID-19 患者的一个风险评估指标。本研究也发现血尿素氮水平与 COVID-19 患者死亡呈强相关性。

淋巴细胞是白细胞的一种，是机体发挥免疫应答功能的重要细胞。研究表明淋巴细胞绝对值与 COVID-19 患者的病情进展密切相关，重型患者的淋巴细胞绝对值显著降低[30]。本研究发现淋巴细胞绝对值与 COVID-19 患者院内接受气管插管治疗关系密切。

超敏 CRP 是一种由肝脏合成的炎症反应急性期的非特异性标志物，临床上可以作为预测心血管疾病的因子。现有证据表明，SARS-CoV-2 可以破坏心肌细胞，造成心脏损伤[23]。因此，超敏 CRP 可以在预测 COVID-19 患者炎症程度方面发挥作用。Guan 等[31]研究也发现超敏 CRP 可以作为 COVID-19 患者死亡风险的预测指标。在本研究中，超敏 CRP 水平与 COVID-19 患者院内接受气管插管治疗密切相关。

总胆红素包括直接胆红素和间接胆红素，主要作为诊断肝脏疾病的指标。研究表明部分危重型 COVID-19 患者可发生严重的肝功能障碍，表现为总胆红素升高，因而总胆红素可能与 COVID-19 的病情相关[32]。Zhan 等[33]的研究也发现总胆红素水平可能是重型 COVID-19 的危险因素。本研究表明总胆红素水平与 COVID-19 患者院内接受气管插管治疗相关。

本研究存在如下局限性：（1）由于 COVID-19 患者入院情况的限制，造成部分重要指标缺失过多（如 BMI、血氧浓度等），这可能会对研究结果造成影响，如纳入这些变量，模型预测效果可能更好。（2）本研究为回顾性研究，尚需进一步开展大样本前瞻性研究进行验证。

综上所述，本研究分析了与 COVID-19 患者临床结局相关性较强的预测因子，借助机器学习算法能够较好地预测其院内临床结局，从而有助于指导 COVID-19 患者的后续治疗和掌握疾病转归情况。

［参 考 文 献］

［1］　World Health Organization. Weekly epidemiological

update on COVID-19—16 March 2021[EB/OL]. (2021-03-16)[2021-03-16]. *https://www.who.int/publications/m/item/weekly-epidemiological-update---16-march-2021.*

[2] KAUFMANN C C, AHMED A, KASSEM M, FREYNHOFER M K, JÄGER B, AICHER G, et al. Mid-regional pro-atrial natriuretic peptide independently predicts short-term mortality in COVID-19[J/OL]. Eur J Clin Invest, 2021, 51: e13531. DOI: 10.1111/eci.13531.

[3] ZHANG R, OUYANG H, FU L, WANG S, HAN J, HUANG K, et al. CT features of SARS-CoV-2 pneumonia according to clinical presentation: a retrospective analysis of 120 consecutive patients from Wuhan city[J]. Eur Radiol, 2020, 30: 4417-4426.

[4] ZHANG L, YAN X, FAN Q, LIU H, LIU X, LIU Z, et al. D-dimer levels on admission to predict in-hospital mortality in patients with COVID-19[J]. J Thromb Haemost, 2020, 18: 1324-1329.

[5] BERTSIMAS D, LUKIN G, MINGARDI L, NOHADANI O, ORFANOUDAKI A, STELLATO B, et al. COVID-19 mortality risk assessment: an international multi-center study[J/OL]. PLoS One, 2020, 15: e0243262. DOI: 10.1371/journal.pone.0243262.

[6] ZHAO Z, CHEN A, HOU W, GRAHAM J M, LI H, RICHMAN P S, et al. Prediction model and risk scores of ICU admission and mortality in COVID-19[J/OL]. PLoS One, 2020, 15: e0236618. DOI: 10.1371/journal.pone.0236618.

[7] UTRERO-RICO A, RUIZ-HORNILLOS J, GONZÁLEZ-CUADRADO C, RITA C G, ALMOGUERA B, MINGUEZ P, et al. IL-6-based mortality prediction model for COVID-19: validation and update in multicenter and second wave cohorts[J/OL]. J Allergy Clin Immunol, 2021, 147: 1652-1661.e1. DOI: 10.1016/j.jaci.2021.02.021.

[8] LI L, FANG X, CHENG L, WANG P, LI S, YU H, et al. Development and validation of a prognostic nomogram for predicting in-hospital mortality of COVID-19: a multicenter retrospective cohort study of 4 086 cases in China[J]. Aging (Albany NY), 2021, 13: 3176-3189.

[9] 中华人民共和国国家卫生健康委员会. 新型冠状病毒感染的肺炎诊疗方案(试行第五版)[EB/OL].(2020-02-05)[2021-02-05]. *http://www.nhc.gov.cn/yzygj/s7653p/202002/3b09b894ac9b4204a79db5b8912d4440/files/7260301a393845fc87fcf6dd52965ecb.pdf.*

[10] CHAWLA N V. Data mining for imbalanced datasets: an overview[M]//MAIMON O, ROKACH L. Data mining and knowledge discovery handbook. Boston, MA: Springer US, 2009: 875-886.

[11] LING C X, LI C. Data mining for direct marketing: problems and solutions[C]//Proceedings of the fourth international conference on knowledge discovery and data mining. New York, NY: AAAI Press, 1998: 73-79.

[12] OJHA V K, ABRAHAM A, SNÁŠEL V. Metaheuristic design of feedforward neural networks: a review of two decades of research[J]. Eng Appl Artif Intell, 2017, 60: 97-116.

[13] JOHN G H, LANGLEY P. Estimating continuous distributions in Bayesian classifiers[C]//Proceedings of the eleventh conference on uncertainty in artificial intelligence. Montréal, Qué, Canada: Morgan Kaufmann Publishers Inc., 1995: 338-345.

[14] BREIMAN L. Random forests[J]. Mach Learn, 2001, 45: 5-32.

[15] SOETERS P B, WOLFE R R, SHENKIN A. Hypoalbuminemia: pathogenesis and clinical significance[J]. JPEN J Parenter Enteral Nutr, 2019, 43: 181-193.

[16] LI G, ZHOU C L, BA Y M, WANG Y M, SONG B, CHENG X B, et al. Nutritional risk and therapy for severe and critical COVID-19 patients: a multicenter retrospective observational study[J]. Clin Nutr, 2021, 40: 2154-2161.

[17] VIOLI F, CANGEMI R, ROMITI G F, CECCARELLI G, OLIVA A, ALESSANDRI F, et al. Is albumin predictor of mortality in COVID-19?[J]. Antioxid Redox Signal, 2021, 35: 139-142.

[18] HORNG T. Calcium signaling and mitochondrial destabilization in the triggering of the NLRP3 inflammasome[J]. Trends Immunol, 2014, 35: 253-261.

[19] ZHOU F, YU T, DU R, FAN G, LIU Y, LIU Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study[J]. Lancet, 2020, 395: 1054-1062.

[20] ZHANG L K, SUN Y, ZENG H, WANG Q, JIANG X, SHANG W J, et al. Calcium channel blocker amlodipine besylate therapy is associated with reduced case fatality rate of COVID-19 patients with hypertension[J/OL]. Cell Discov, 2020, 6: 96. DOI: 10.1038/s41421-020-00235-0.

[21] NEURAZ A, LERNER I, DIGAN W, PARIS N, TSOPRA R, ROGIER A, et al. Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the COVID-19 pandemic[J/OL]. J Med Internet Res, 2020, 22: e20773. DOI: 10.2196/20773.

[22] DU R H, LIANG L R, YANG C Q, WANG W, CAO T Z, LI M, et al. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study[J/OL]. Eur Respir J, 2020, 55. DOI: 10.1183/13993003.00524-2020.

[23] BANSAL M. Cardiovascular disease and COVID-19[J].

Diabetes Metab Syndr: Clin Res Rev, 2020, 14: 247-250.

［24］ HAN H, XIE L L, LIU R, YANG J, LIU F, WU K L, et al. Analysis of heart injury laboratory parameters in 273 COVID-19 patients in one hospital in Wuhan, China[J]. J Med Virol, 2020, 92: 819-823.

［25］ HENRY B M, DE OLIVEIRA M H S, BENOIT S, PLEBANI M, LIPPI G. Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis[J]. Clin Chem Lab Med CCLM, 2020, 58: 1021-1028.

［26］ LI T, WANG X, ZHUANG X, WANG H, LI A, HUANG L, et al. Baseline characteristics and changes of biomarkers in disease course predict prognosis of patients with COVID-19[J]. Intern Emerg Med, 2021, 16: 1165-1172.

［27］ KAZORY A. Emergence of blood urea nitrogen as a biomarker of neurohormonal activation in heart failure[J]. Am J Cardiol, 2010, 106: 694-700.

［28］ ARONSON D, MITTLEMAN M A, BURGER A J. Elevated blood urea nitrogen level as a predictor of mortality in patients admitted for decompensated heart failure[J]. Am J Med, 2004, 116: 466-473.

［29］ CHENG A, HU L, WANG Y, HUANG L, ZHAO L, ZHANG C, et al. Diagnostic performance of initial blood urea nitrogen combined with D-dimer levels for predicting in-hospital mortality in COVID-19 patients[J/OL]. Int J Antimicrob Agents, 2020, 56: 106110. DOI: 10.1016/j.ijantimicag.2020.106110.

［30］ DENG R, WANG C, YE Y, GOU L, FU Z, YE B, et al. Clinical manifestations of blood cell parameters and inflammatory factors in 92 patients with COVID-19[J/OL]. Ann Transl Med, 2021, 9: 62. DOI: 10.21037/atm-20-7765.

［31］ GUAN X, ZHANG B, FU M, LI M Y, YUAN X, ZHU Y W, et al. Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study[J]. Ann Med, 2021, 53: 257-266.

［32］ FAN H, CAI J, TIAN A, LI Y, YUAN H, JIANG Z, et al. Comparison of liver biomarkers in 288 COVID-19 patients: a mono-centric study in the early phase of pandemic[J/OL]. Front Med (Lausanne), 2020, 7: 584888. DOI: 10.3389/fmed.2020.584888.

［33］ ZHAN N, GUO Y, TIAN S, HUANG B, TIAN X, ZOU J, et al. Clinical characteristics of COVID-19 complicated with pleural effusion[J/OL]. BMC Infect Dis, 2021, 21: 176. DOI: 10.1186/s12879-021-05856-8.

［本文编辑］　商素芳