·智慧医疗· doi:10.3969/j.issn.1671-8348.2022.24.030
网络首发 http://kns.cnki.net/kcms/detail/50.1097.R.20221014.1452.006.html(2022-10-17)

基于随机森林模型法的 AMI 患者并发 AKI 预测模型的建立

李 龙,刘真义,李浩然,药永红△
(中国人民解放军联勤保障部队第九四五医院急诊科,四川雅安 625000)

[摘要] **目的** 建立基于随机森林模型法的预测模型对急性心肌梗死(AMI)患者并发急性肾损伤(AKI)进行预测,找出相关重要指标。**方法** 选取 2014 年 1 月至 2021 年 1 月该院急诊科收治的 AMI 患者 1 362 例作为研究对象,将合并 AKI 患者设为观察组(270 例),未合并 AKI 设为对照组(1 092 例)。在确定 30 个变量后进行数据的相关统计和分析,随机选取 75%的病例进行训练数据库的建立,25%的病例作为测试数据库,采用 R 语言进行数据的筛选和模型的建立,对其进行相关评估,并与其余 3 种机器学习模型进行对比。**结果** 1 362 例患者中合并 AKI 270 例(19.82%)。两组患者除血小板、球蛋白、入院时体温、血钠、天门冬氨酸氨基转移酶、丙氨酸氨基转移酶比较差异无统计学意义($P>0.05$)外;其余各指标比较,差异均有统计学意义($P<0.05$);随机森林模型受试者工作曲线下面积为 0.894,均高于其余 3 种模型,灵敏度为 0.792,特异度为 0.867;模型中变量重要性依次为首次肌酐、尿素值,机械通气、年龄和 D-二聚体。**结论** 基于随机森林模型对 AMI 患者是否发生 AKI 进行预测具有较好的预测效能,在实际临床工作中具有一定参考价值。

[关键词] 急性心肌梗死;急性肾损伤;机器学习;预测模型;受试者工作曲线
[中图法分类号] R540.4 [文献标识码] A [文章编号] 1671-8348(2022)24-4304-04

Establishment of prediction model in patients with AMI complicating AKI based on random forest model method

LI Long, LIU Zhenyi, LI Haoran, YAO Yonghong△
(*Department of Emergency, 945 Hospital of PLA Joint Logistics Support Force, Ya'an, Sichuan* 625000, *China*)

[**Abstract**] **Objective** To establish a prediction model based on the random forest model method to predict the patients with acute myocardial infarction (AMI) complicating acute kidney injury (AKI) for finding relevant important indicators. **Methods** A total of 1 362 patients with AMI admitted and treated in the emergency department of this hospital from January 2014 to January 2021 were selected as the research subjects. The patients with AMI complicating AKI served as the observation group(270 cases) and those without complicating AKI as the control group(1 092 cases). After determining 30 variables, the relevant statistics and analysis of the data were performed, and 75% of the cases were randomly selected to conduct the training database establishment, 25% of the cases as the test database, the R language was used to conduct the data screening and model establishment, the related evaluation on them was conducted, then which was compared with the other three kinds of machine learning models. **Results** Among 1 362 cases, there were 270 cases (19.82%) complicating AKI. The comparison of platelets, globulin, admission body temperature, blood sodium, glutamic oxalacetic transaminase and alanine aminotransferase between the two groups showed no statistical difference ($P>0.05$). The other indicators showed statistically significant differences between the two groups ($P<0.05$); the area under the receiver operating curve of the random forest model was 0.894, which was higher than those in the other three models, the sensitivity was 0.792 and the specificity was 0.867; the importance of variables in the model was as follows: first time creatinine, urea value, mechanical ventilation, age and D-dimer. **Conclusion** Based on the random forest model, predicting the occurrence of AKI in the patients with AMI has good predictive power, and has a certain reference value in actual clinical work.

[**Key words**] acute myocardial infarction; acute kidney injury; machine learning; pedictive model; receiver operating curve

作者简介:李龙(1989-),主治医师,硕士,主要从事急危重症患者临床急救的研究。 △ **通信作者**,E-mail:1463668948@qq.com。

急性心肌梗死(AMI)是心脏内科最为常见的急危重症疾病类型之一,尽管随着医疗技术和水平的不断提升及治疗手段的更新,但 AMI 的救治率仍不高[1]。急性肾损伤(AKI)是指人体肾脏功能在短时间内出现进行性下降的一类疾病综合征,诱发 AKI 的因素较多,在发生 AMI 后出现肾功能障碍称为心肾综合征[2]。AMI 患者最为常见的并发症之一为 AKI,其主要原因包括冠状动脉造影剂对肾脏的损伤和 AMI 所致的肾脏血流灌注不足。一旦 AMI 患者出现 AKI 病情将会急剧加重,严重影响患者预后,故而在实际临床工作中对 AMI 患者是否发生 AKI 进行一定的预测显得尤为重要[3]。既往诸多研究对 AMI 患者并发 AKI 预测模型多采用 logistic 回归模型进行分析,随着人工智能的不断更新,机器学习得到广泛应用,预测模型的建立方法也逐渐增多[4]。本研究探讨基于随机森林模型法对 AMI 患者并发 AKI 进行预测的效果,旨在为临床诊治提供理论依据。

1 资料与方法

1.1 一般资料

选取 2014 年 1 月至 2021 年 1 月本院急诊科收治的 AMI 患者 1 362 例作为研究对象。将合并 AKI 患者设为观察组(270 例),未合并 AKI 设为对照组(1 092 例)。AMI 诊断标准[5]:具有典型临床表现、特征性的心电图改变,以及血液学指标变化;AKI 诊断标准[6]:48 h 内血清肌酐水平升高大于或等于 0.3 mg/dL 或已超过基础肌酐值的 1.5 倍以上或 6 h 尿量持续少于 0.5 mL·kg^{-1}·h^{-1}。纳入标准:(1)主要诊断符合 AMI 诊断标准;(2)发生 AMI 后 24 h 内入院。排除标准:(1)血液学检查指标中尿素或肌酐检查缺失;(2)既往已有终末期肾病或透析治疗。本研究符合《世界医学协会赫尔辛基宣言》相关要求。

1.2 方法

1.2.1 自变量的选择

根据实际临床情况结合文献[7-13]初步确定 140 个与 AMI 和 AKI 具有相关性的变量,因收集和整理变量数较大,需进行一定的降维处理。(1)剔除缺失程度达 15%以上的变量,而后采用 R 语言中的 CARET 程序对所收集的数据进行相关预处理。剔除与其他自变量存在较强相关性的变量(程序语句:find correlation),因自变量中还存在一定的多重线性问题,继续采用相关程序(程序语句:find Liner Combos)进行查找及剔除。(2)对数据缺失未达 15%的数据进行相关处理和补充,对符合正态分布的数据采用平均数给予补充,对偏态分布的数据采用中位数给予补充。数据经过初步处理后再进行相关筛选,采用 rfFuncs 建立随机森林模型,而后采用相关命令进行自变量的选择,经过程序的选择后最终获得并确定 30 个变量,主要包括人口学资料、疾病危险因素、生命体征、实验室检查等。

1.2.2 模型的建立

经过自变量的确定后,在 1 362 例患者中随机选取 75%的病例进行训练数据库的建立,25%的病例作为测试数据库。随机森林算法采用相关程序包——随机森林进行,抽样方式选取 Bootstrap。随机森林模型当中具有 2 个重要的数据参数,包括决策树棵树——ntree、分裂节点预估变量数目——mtry。首先进行 mtry 节点值的选取,此节点值即为二叉数的变量数目,此模型所对应的最小值为 24,而后进行测试阶段,进行最佳 ntree 的匹配,当 ntree=900 时此模型表现最佳。模型中变量的重要性采用 importance 函数进行计算,数值越大表示重要性越强。

1.2.3 模型的评估及对比

建立模型后进行一定程度评估,看其是否适合对疾病进行预测。采用 R 语言对测试数据库中的匹配数据进行计算,得出其准确率、灵敏度和特异度,再采用 R 语言中的程序包——pROC 计算受试者工作特征曲线下面积(AUC),评估建立的随机森林模型。同时进行朴素贝叶斯、支持向量机及人工神经网络等其他较为常用的机器学习方法的计算,并将所有结果与随机森林模型进行相关对比。

1.3 统计学处理

采用 Rv3.6.0 软件进行数据分析,计量资料以 $\bar{x}\pm s$ 表示,组间比较采用独立样本 t 检验;计数资料以率表示,组间比较采用 χ^2 检验。以 $P<0.05$ 为差异有统计学意义。

2 结 果

2.1 一般资料

1 362 例患者中合并 AKI 270 例(19.82%),119 例(8.74%)患者给予机械通气。两组患者血小板、球蛋白、入院时体温、血钠、丙氨酸氨基转移酶、天门冬氨酸氨基转移酶比较,差异无统计学意义($P>0.05$);其余各指标比较,差异均有统计学意义($P<0.05$),见表 1。

表 1 两组患者一般资料比较

指标	对照组(n=1 092)	观察组(n=270)	t/χ^2	P
年龄($\bar{x}\pm s$,岁)	64.65±12.18	70.24±12.19	−6.755	<0.001
性别[n(%)]			20.592	<0.001
男	790(72.34)	157(58.15)		

续表1　两组患者一般资料比较

指标	对照组(n=1 092)	观察组(n=270)	t/χ^2	P
女	302(27.66)	113(41.85)		
有饮酒史[n(%)]	410(37.55)	70(25.93)	12.807	<0.001
有吸烟史[n(%)]	598(54.76)	90(33.33)	39.765	<0.001
有高血压[n(%)]	622(56.96)	181(67.04)	9.085	0.003
有糖尿病[n(%)]	217(19.87)	90(33.33)	22.468	<0.001
给予机械通气[n(%)]	41(3.75)	78(28.89)	171.506	<0.001
Killip分级[n(%)]			208.750	<0.001
1级	797(72.99)	87(32.22)		
2级	192(17.58)	69(25.56)		
3级	42(3.85)	48(17.78)		
4级	61(5.59)	66(24.44)		
未治愈[n(%)]	47(4.30)	65(24.07)	112.112	<0.001
发生恶性事件[n(%)]	48(4.40)	37(13.70)	32.054	<0.001
白细胞($\bar{x}\pm s$,$\times10^9$/L)	9.16±2.33	10.37±3.15	−7.083	<0.001
中性粒细胞($\bar{x}\pm s$,$\times10^9$/L)	6.51±2.89	7.85±3.13	−6.708	<0.001
淋巴细胞($\bar{x}\pm s$,$\times10^9$/L)	1.73±0.42	1.21±0.31	19.096	<0.001
红细胞($\bar{x}\pm s$,$\times10^{12}$/L)	4.73±1.82	4.09±1.79	5.191	<0.001
红细胞分布宽度($\bar{x}\pm s$)	0.11±0.06	0.14±0.05	−7.589	<0.001
血小板($\bar{x}\pm s$,$\times10^9$/L)	208.48±38.95	206.83±40.22	0.619	0.536
平均血小板体积($\bar{x}\pm s$,fL)	9.46±2.48	9.82±2.82	−2.076	0.038
D-二聚体($\bar{x}\pm s$,μg/L)	0.62±0.28	1.23±0.42	−28.701	<0.001
白蛋白($\bar{x}\pm s$,g/L)	37.89±4.65	35.25±5.17	8.165	<0.001
球蛋白($\bar{x}\pm s$,g/L)	25.87±4.78	26.15±5.33	−0.842	0.400
首次肌酐($\bar{x}\pm s$,μmol/L)	70.89±21.27	107.35±33.59	−22.158	<0.001
首次尿素($\bar{x}\pm s$,mmol/L)	5.48±1.22	9.37±2.64	−35.684	<0.001
尿红细胞($\bar{x}\pm s$,$\times10^{12}$/L)	6.49±3.16	10.93±4.85	−18.358	<0.001
舒张压($\bar{x}\pm s$,mm Hg)	81.58±9.22	75.16±10.05	10.059	<0.001
急诊入院时心率($\bar{x}\pm s$,次/分)	75.85±11.29	86.63±13.17	−13.573	<0.001
入院时体温($\bar{x}\pm s$,℃)	36.83±1.86	36.81±1.73	0.160	0.873
呼吸频率($\bar{x}\pm s$,次/分)	17.24±2.75	20.56±3.15	−17.239	<0.001
血磷($\bar{x}\pm s$,mmol/L)	1.01±0.34	1.25±0.42	−9.884	<0.001
血钾($\bar{x}\pm s$,mmol/L)	3.62±0.75	3.97±1.03	−6.334	<0.001
血钠($\bar{x}\pm s$,mmol/L)	142.37±25.48	143.29±27.16	−0.524	0.600
凝血酶原时间($\bar{x}\pm s$,s)	1.01±0.54	1.12±0.39	−3.150	0.002
C反应蛋白($\bar{x}\pm s$,mg/L)	3.89±1.43	12.82±6.17	−43.388	<0.001
丙氨酸氨基转移酶($\bar{x}\pm s$,U/L)	102.37±33.54	105.29±45.27	−1.188	0.235
天门冬氨酸氨基转移酶($\bar{x}\pm s$,U/L)	34.83±11.42	36.24±13.51	−1.749	0.081
乳酸脱氢酶($\bar{x}\pm s$,U/L)	231.79±83.27	287.58±97.28	−9.520	<0.001
三酰甘油($\bar{x}\pm s$,mmol/L)	1.55±0.43	1.21±0.38	11.894	<0.001
总胆固醇($\bar{x}\pm s$,mmol/L)	4.47±1.38	4.13±1.24	3.696	0.038
脑钠肽($\bar{x}\pm s$,ng/L)	583.29±208.16	2 814.57±742.94	−86.532	<0.001

2.2　模型测试　共341例测试数据库中患者进行预测，其中290

例患者预测正确，正确率为 85.04%。见表 2。

2.3 预测效能

随机森林模型 AUC 为 0.894，均高于其余 3 种模型，灵敏度为 0.792，特异度为 0.867。见表 3。首次肌酐、尿素值、机械通气、年龄、D-二聚体为其前五重要变量。

表 2 随机森林模型测试结果

项目	预测		准确率(%)
	发生	不发生	
实际			85.04
发生	265	42	
未发生	9	25	

表 3 各模型预测效能比较

模型类型	AUC	灵敏度	特异度	准确度
随机森林模型	0.894	0.792	0.867	0.850
支持向量机	0.868	0.792	0.822	0.798
朴素贝叶斯	0.890	0.851	0.807	0.842
人工神经网络	0.820	0.924	0.613	0.863

3 讨 论

随着大数据时代的到来，更多的人工智能算法也同样用于医学各领域中，随机森林法就是其中之一，其在医学大数据的处理中表现出了极高的效能，特别是在基因、药物、疾病等领域中展现出了其独有的特点，既往对 AMI 并发 AKI 患者多采用多因素 logistic 回归模型进行预测，应用随机森林模型的研究仍较少见[7]。本研究通过建立随机森林模型进行疾病的预测，最终对测试数据库评估的结果显示，该模型预测准确率为 85.04%，AUC 值为 0.894，均高于其余 3 种模型，灵敏度为 0.792，特异度为 0.867，提示该预测模型预测能力较好，且高于其余 3 种常用的机器学习模型。

本研究对所选取的自变量进行了重要性排序，结果显示，首次肌酐、尿素值、机械通气、年龄、D-二聚体为其前五重要变量。肌酐及尿素值代表了患者肾功能情况的基线水平，既往研究表明，AMI 合并 AKI 的预测中肾功能为其重要的影响因素之一[14]。AMI 患者全身各器官均会在一定程度上出现灌注不足的情况，对肾功能基线水平较差的患者而言，病情将会进一步加重，故在临床工作中应对 AMI 患者进行常规肾功能基线水平测定，实时掌握患者病情的进展情况。D-二聚体作为一种血栓标志物，常用于如肺栓塞的诊断中，同时也可用于 AMI 的早期诊断及预后预测[8]。既往对糖尿病肾病早期肾损伤的研究表明，D-二聚体同样与肾脏功能密切相关[15]。因此推测，D-二聚体在 AMI 患者合并 AKI 的预测中具有一定的价值。

本研究模型中年龄因素同样占据一定的重要性，合并 AKI 患者年龄明显大于未合并 AKI 者，与既往研究结果相似[9]。另一方面，本研究中是否给予机械通气也为模型中的重要因素之一，分析其原因在于 AMI 患者一般情况下病情较为危重，发生心、肺功能障碍或衰竭的风险加大，本研究 1 362 例患者中 8.74%使用了机械通气治疗。既往研究表明，机械通气是发生 AKI 的独立危险因素，对患者血流动力学、炎性反应等多方面造成一定程度的影响，同时，发生 AKI 后又会反作用于机械通气的治疗及预后，故而对此类患者而言，机械通气模式及参数的设定显得尤为重要，将会直接影响患者预后[10]。

既往研究将随机森林模型用于预测造影剂所致的 AKI 中，同时还与传统 Logistic 回归模型进行了相关对比，最终结果显示，机器学习模式明显优于传统模式[11]。本研究将随机森林预测模型与其他 3 种机器学习模型进行了相关对比，结果显示，随机森林模型预测效能均优于其余 3 种，分析其中原因在于，本研究 1 362 例患者并非全部进行了冠状动脉造影，因此，未能对造影剂所致的 AKI 进行相关区分；另一方面本研究人群与国外研究人群存在一定的差异，纳入的自变量也存在一定的不同。与国外研究对比发现，肾脏功能基线水平、年龄均被纳入模型中，由此可见，此两项对 AMI 患者是否发生 AKI 的预测具有非常重要的参考价值[12]。但本研究仍存在一些不足和局限：(1)因自变量选取较多，在实际临床工作中可能实用性方面有所制约；(2)仅为单中心研究，样本来源受限，特别是对一些缺失值较多的变量被迫进行了剔除，导致结果可能存在一定的偏倚。

综上所述，基于随机森林模型对 AMI 患者是否发生 AKI 具有较好的预测效能，在实际临床工作中具有一定的参考价值，特别是对变量数据不存在缺失的患者建议积极使用。

参考文献

[1] REED G W，Rossi J E，Cannon C P. Acute myocardial infarction[J]. Lancet，2017，389(10065)：197-210.

[2] RONCO C，BELLASI A，DI LULLO L. Cardiorenal syndrome：an overview[J]. Adv Chronic Kidney Dis，2018，25(5)：382-390.

[3] 吴淡森，石松菁. “肾”事风云：再谈急性肾损伤的基础与临床研究[J]. 中华急诊医学杂志，2019，28(9)：1066-1070.

[4] SUN L，ZHU W，CHEN X，et al. Machine learning to predict contrastinduced acute （下转第 4312 页）

[7] 谢志勇,李志莲,董伟,等.慢性肾小球疾病谱演变和膜性肾病流行病学特点[J].临床肾脏病杂志,2019,19(7):471-476.
[8] ZENG C,NAN Y,XU F,et al. Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning[J]. J Pathol,2020,252(1):53-64.
[9] SALVI M,ACHARYA U R,MOLINARI F,et al. The impact of pre-and post-image processing techniques on deep learning frameworks:a comprehensive review for digital pathology image analysis[J]. Comput Biol Med,2021,128:104129.
[10] BOUTELDJA N,KLINKHAMMER B M,BÜL OW R D,et al. Deep learning-based segmentation and quantification in experimental kidney histopathology[J]. J Am Soc Nephrol,2021,32(1):52-68.
[11] JAYAPANDIAN C P,CHEN Y,JANOWCZYK A R,et al. Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains[J]. Kidney Int,2021,99(1):86-101.
[12] CHAGAS P,SOUZA L,Araújo I,et al. Classification of glomerular hypercellularity using convolutional features and support vector machine[J]. Artif Intell Med,2020,103:101808.
[13] TAO Z,BINGQIANG H,HUILING L,et al. NSCR-based densenet for lung tumor recognition using chest CT image[J]. Biomed Res In,2020,2020:6636321.
[14] HUANG G,LIU Z,PLEISS G,et al. Convolutional networks with dense connectivity[J]. IEEE Trans Pattern Anal Mach Intell,2022,44(12):8704-8716.
[15] NORMAN B,PEDOIA V,NOWOROLSKI A,et al. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs[J]. J Digit Imaging,2019,32(3):471-477.

(收稿日期:2022-04-23 修回日期:2022-09-11)

(上接第 4307 页)

kidney injury in patients with acute myocardial infarction[J]. Front Med (Lausanne),2020,7:592007.
[5] 中华医学会心血管病学分会,中华心血管病杂志编辑委员会.急性 ST 段抬高型心肌梗死诊断和治疗指南[J].中华心血管病杂志,2010,38(8):675-690.
[6] LEVEY A S,JAMES M T. Acute kidney injury[J]. Ann Intern Med,2017,167(9):66-80.
[7] MODY P,WANG T,MCNAMARA R,et al. Association of acute kidney injury and chronic kidney disease with processes of care and long-term outcomes in patients with acute myocardial infarction[J]. Eur Heart J Qual Care Clin Outcomes,2018,4(1):43-50.
[8] WANG C,PEI Y Y,MA Y H,et al. Risk factors for acute kidney injury in patients with acute myocardial infarction[J]. Chin Med J (Engl),2019,132(14):1660-1665.
[9] 肖辉,郝元涛,徐晓,等.基于随机森林算法和 Logistic 回归模型的糖尿病风险因素研究[J].中国数字医学,2018,13(1):33-35.
[10] XU F B,CHENG H,YUE T,et al. Derivation and validation of a prediction score for acute kidney injury secondary to acute myocardial infarction in Chinese patients[J]. BMC Nephrol,2019,20(1):195.
[11] GUPTA S,KO D T,AZIZI P,et al. Evaluation of machine learning algorithms for predicting readmission after acute myocardial infarction using routinely collected clinical data[J]. Can J Cardiol,2020,36(6):878-885.
[12] 蓝潞杭,蒋炫东,王茂峰,等.随机森林模型预测急性心肌梗死后急性肾损伤[J].中华急诊医学杂志,2021,30(4):491-495.
[13] 蒋远霞,唐艳,易扬,等.高尿酸血症是脓毒症患者发生急性肾损伤的独立危险因素[J].中华急诊医学杂志,2020,29(9):1178-1183.
[14] VIGNOLI A,TENORI L,GIUSTI B,et al. NMR-based metabolomics identifies patients at high risk of death within two years after acute myocardial infarction in the AMI-Florence Ⅱ cohort[J]. BMC Med,2019,17(1):3.
[15] 赵梦蝶,孙九爱.机器学习在心血管疾病诊断中的研究进展[J].北京生物医学工程,2020,39(2):208-214.

(收稿日期:2022-02-18 修回日期:2022-06-23)