论著·智能医学与预测模型

利用可解释机器学习识别关键生物标志物预测代谢功能障碍相关脂肪性肝病患者的慢性肾病风险:基于NHANES数据库的研究

徐 燕,侯文青,唐毓馨,覃中毅,王 涛,向俊宇,王 斌,文良志,陈东风 陆军军医大学(第三军医大学)大坪医院消化内科,消化系统肿瘤精准防治重庆市重点实验室,重庆

[摘要] **目的** 基于机器学习算法构建代谢功能障碍相关脂肪性肝病(metabolic dysfunction-associated steatotic liver disease,MASLD)患者发生慢性肾病(chronic kidney disease, CKD)预测模型,绘制诊断列线图以指导临床诊治。**方法** 本研究基于2007—2018年美国国家健康与营养调查(National Health and Nutrition Examination Survey, NHANES)数据库,筛选存在肝脏脂肪变性且至少伴有一项心脏代谢危险因素的MASLD患者作为分析样本。在排除合并慢性肾衰竭等患者后,最终纳入2 144例MASLD患者,收集其人口统计学、临床特征及实验室指标等34个变量。根据是否发生CKD,将患者分为CKD组(n=347)与非CKD组(n=1 797)。按7∶3比例将总样本分层抽样随机划分为训练集(n=1 501)和内部验证集(n=643),并纳入陆军军医大学大坪医院消化内科2024年1月至2025年10月期间符合既定纳入与排除标准的110例脂肪肝患者作为外部验证集。在模型构建方面,基于Lasso回归筛选变量后,分别纳入决策树(decision trees,DT)、极端梯度提升机(extreme gradient boosting, XGBoost)、K-最近邻算法(K-nearest neighbors,KNN)、朴素贝叶斯(Naive Bayes, NB)、支持向量机(support vector machine,SVM)、单隐藏层神经网络(neural networks, NNET)、梯度提升机(light gradient boosting machine,LightGBM)、随机森林(random forest,RF)共8种机器学习算法,以构建CKD风险预测模型。通过受试者工作特征(receiver operating characteristic,ROC)曲线下面积(area under curve,AUC)、敏感度、特异度等综合评价模型性能,并进一步通过校准曲线与临床决策曲线评估其临床应用价值。基于沙普利加性解释(Shapley additive explanation,SHAP)分析筛选出的重要生物标志物,构建了诊断列线图,并通过ROC曲线评估了其诊断准确性。**结果** 本研究确定了10个生物标志物,包括收缩压、年龄、糖尿病史、血尿素氮、BMI、球蛋白、高密度胆固醇、中性粒细胞计数、血尿酸、γ-谷氨酰基转移酶,将上述生物标志物通过8种机器学习算法构建预测模型,综合评价发现LightGBM模型性能最好,AUC为0.871,敏感度为0.838,特异度为0.756,准确率为0.825,F1得分为0.889,Brier得分为0.091。校准曲线显示,基于LightGBM算法构建的预测模型在训练集和验证集一致性良好;临床决策曲线显示,使用该模型预测MASLD患者相关肾损害风险有助于临床决策,具有临床实用性。基于SHAP值构建诊断列线图,并以81.7分作为风险阈值,根据列线图风险阈值可将患者分为MASLD CKD高风险和低风险人群,并且列线图显示出良好的准确性和预测性能,其AUC为0.816。**结论** LightGBM模型较其他模型效能显著,其在临床早期预测MASLD患者CKD风险中具有应用潜力。

[关键词] 代谢功能障碍相关脂肪性肝病;慢性肾病;机器学习;疾病风险预测

[中图法分类号] R195.1;R575.506;R692 [文献标志码] A

[基金项目] 国家自然科学基金面上项目(82170594)

[通信作者] 陈东风,E-mail:chendf@tmmu.edu.cn

文良志,E-mail:wenliangzhi@tmmu.edu.cn

2097-0927/Copyright © 2026, 陆军军医大学(第三军医大学)。这篇开放获取文章遵循 CC BY 许可协议(https://creativecommons.org/licenses/by/4.0/)

引用格式:徐燕,侯文青,唐毓馨,等. 利用可解释机器学习识别关键生物标志物预测代谢功能障碍相关脂肪性肝病患者的慢性肾病风险:基于NHANES数据库的研究[J]. 陆军军医大学学报,2026,48(03):354-365. DOI:10.16016/j.2097-0927.202510032.

Citation:XU Y,HOU W Q,TANG Y X,*et al*. Explainable machine learning for identifying key biomarkers in predicting chronic kidney disease risk among patients with metabolic dysfunction-associated fatty liver disease: A study based on NHANES database[J]. J Army Med Univ,2026,48(03):354-365. DOI:10.16016/j.2097-0927.202510032.

Explainable machine learning for identifying key biomarkers in predicting chronic kidney disease risk among patients with metabolic dysfunction-associated fatty liver disease: A study based on NHANES database

XU Yan, HOU Wenqing, TANG Yuxin, QIN Zhongyi, WANG Tao, XIANG Junyu, WANG Bin, WEN Liangzhi, CHEN Dongfeng Department of Gastroenterology, Chongqing Key Laboratory of Precision Prevention and Treatment for Digestive System Tumors, Chongqing, Daping Hospital, Army Medical University (Third Military Medical University), China

[**Abstract**] Objective To construct a machine learning-based prediction model for chronic kidney disease(CKD) in patients with metabolic dysfunction-associated steatotic liver disease (MASLD) and to develop a diagnostic nomogram for guiding clinical management. Methods This study was based on the National Health and Nutrition Examination Survey (NHANES) database (2007 to 2018). Patients with MASLD, defined as hepatic steatosis with at least one cardiometabolic risk factor, were selected as the analysis sample. After excluding patients with chronic renal failure and other conditions, a total of 2 144 MASLD patients were included, and 34 variables (including demographics, clinical characteristics, and laboratory indicators) were collected. Patients were stratified into a chronic kidney disease (CKD) group (n=347) and a non-CKD group (n=1 797) based on CKD status. The total participants were randomly divided into a training set (n=1 501) and an internal validation set (n=643) using stratified sampling at a 7∶3 ratio. A prospective external validation cohort comprising 110 eligible fatty liver disease patients was recruited from our department during January 2024 and October 2025, adhering strictly to predefined inclusion and exclusion criteria Variables were screened using Least Absolute Shrinkage and Selection Operator (Lasso) regression. Eight machine learning algorithms were then employed to construct the CKD risk prediction models: decision trees (DT), extreme gradient boosting (XGBoost), K-nearest neighbors (KNN), Naive Bayes (NB), support vector machine (SVM), single hidden layer neural networks (NNET), light gradient boosting machine (LightGBM), and random forest (RF). Model performance was comprehensively evaluated using receiver operating characteristic (ROC) curves area under the curve (AUC), sensitivity, specificity, calibration curves, and clinical decision curves. Based on key biomarkers identified through SHAP analysis, a diagnostic nomogram was constructed, and its diagnostic accuracy was assessed using ROC curves. Results Using interpretable machine learning algorithms, we identified 10 key biomarkers: SBP, age, diabetes, BUN, BMI, globulin, HDL-C, neutrophil count, uric acid, and gamma-glutamyl transferase. These biomarkers were utilized to construct predictive models through 8 distinct machine learning algorithms. Comprehensive evaluation revealed that the LightGBM model achieved optimal performance, with an AUC of 0.871, a sensitivity of 0.838, a specificity of 0.756, an accuracy of 0.825, an F1 score of 0.889, and a Brier score of 0.091. Calibration curve analysis demonstrated good consistency of the LightGBM-based predictive model across both training and validation sets. Clinical decision curve analysis indicated significant clinical utility of this model in predicting kidney damage risk in patients with MASLD. Based on SHAP values, we developed a diagnostic nomogram with a risk cutoff score of 81.7. This nomogram enabled categorization of patients into high-risk and low-risk subsets for MASLD-related CKD, and exhibited favorable accuracy and robust predictive reliability with an AUC value of 0.816. Conclusion The LightGBM model significantly outperforms other models and demonstrates potential for application in early clinical prediction of CKD risk among MASLD patients.

[**Key words**] metabolic dysfunction-associated steatotic liver disease; chronic kidney disease; machine learning; disease risk prediction

Supported by the General Program of National Natural Science Foundation of China (82170594). Corresponding authors: CHEN Dongfeng, E-mail: chendf@tmmu.edu.cn; WEN Liangzhi, E-mail: wenliangzhi@tmmu.edu.cn

代谢功能障碍相关脂肪性肝病(metabolic dysfunction-associated steatotic liver disease,MASLD)是一种与胰岛素抵抗(insulin resistance,IR)和遗传易感性密切相关的代谢性肝损伤[1]。目前,MASLD已成为全球慢性肝病最常见的原因,预计到2040年将影响全球高达55%的成年人口[2]。根据我国2020年流行病学数据显示,成年人MASLD患病率约为29.2%,已超过部分欧美国家[3];在过去30年中,MASLD的全球患病率增加了近50%[4]。近年来,逐渐认识到MASLD是一种多系统疾病[4-5],不仅与肝脏相关并发症[如肝硬化、肝细胞癌(hepatocellular carcinoma,HCC)、终末期肝病]的风险增加相关,且与多种肝外损害[如心血管疾病(cardiovascular disease,CVD)、慢性肾病(chronic kidney disease,CKD)和某些类型的肝外肿瘤]的风险增加相关[5-6]。目前对于MASLD相关肝外疾病虽已展开多项关联研究[4,7-8],但尚缺乏有效的生物预测标志物及良好的预测模型,这限制了临床医生早期识别高危人群并以此开展针对性干预措施。

CKD是最常见的慢性病之一,其全球疾病负担很重,但患者对此的认知较低。早期CKD的症状是非特异性的,甚至许多人没有症状,而在常规体检时才发现,或者直到他们发展为终末期CKD时才出现症状[9-10]。MASLD和CKD之间的关联主要通过共同的代谢危险因素、炎症等介导[11-12]。研究表明,在MASLD患者中,CKD的患病率明显较高(20%~50%);与单纯性脂肪变性相比,CKD的发病率与代谢功能障碍相关脂肪性肝炎(metabolic dysfunction-associated steatohepatitis,MASH)的相关性更强,特别是当MASLD患者伴有晚期肝纤维化时,其CKD患病率会进一步显著升高[6,12-13]。上述研究表明,MASLD是导致CKD发生及进展的重要原因,而由于MASLD早期症状隐匿,很多患者就医时可能已处于肝纤维化阶段,使其患CKD风险大大增加。因此,及早识别MASLD患者中的CKD高危人群是预防MASLD患者发生CKD的关键步骤。

机器学习通过算法对临床数据进行分析,为临床预测模型的开发开辟了新途径[14]。然而,机器学习模型往往难以解释模型的预测结果,通过沙普利加性解释(Shapley additive explanation,SHAP)能够对机器学习模型进行全局和局部解释,量化每个特征对整体事件预测结果的贡献度,并具有易于理解使用的可视化功能[15]。考虑到机器学习的无创性优势、MASLD患病负担的日益严峻以及临床的迫切需求,因此建立有效且性能良好的模型来预测MASLD患者合并CKD的风险十分必要且有重要的临床转化价值。本研究运用可解释的机器学习分析方法构建预测MASLD患者CKD风险的预测模型,以提供一种更为便捷且准确的方法来识别MASLD中CKD高风险个体,从而为临床决策提供具有前瞻性的支持。

1 资料与方法

1.1 数据来源

本研究数据来源于美国国家健康与营养调查(National Health and Nutrition Examination Survey,NHANES)数据库,包括2007—2018年共12年调查周期的数据。NHANES数据库由美国疾病控制与预防中心(Centers for Disease Control and Prevention,CDC)组织实施,采用多阶段分层抽样方法,系统收集美国居民的健康、营养、疾病及生活方式数据,涵盖人口学特征、体格检查、实验室检测及问卷调查等多维度信息。所有数据通过标准化流程采集,并经过严格的质量控制与伦理审查,NHANES研究已通过美国国家卫生统计中心(National Center for Health Statistics,NCHS)伦理审查委员会批准,所有参与者签署知情同意书[16]。

为评估模型的泛化能力,本研究使用独立的外部验证队列。本研究的外部验证集数据来源于陆军军医大学大坪医院消化内科的住院患者。通过回顾性收集2024年1月至2025年10月期间符合既定纳入与排除标准的脂肪肝患者病历资料,共纳入110例患者。所有病例数据从该中心的电子病历系统中提取,并进行严格数据核对与审查,以确保变量定义的准确性、测量方式的一致性以及数据整体的可靠性。本研究已通过中国人民解放军陆军特色医学中心伦理委员会审查批准[医研伦审(2025)第501号],并已获得豁免知情同意。所有数据在处理与分析过程中进行了严格的匿名化处理,以充分保护患者隐私。研究过程严格遵守《赫尔辛基宣言》的伦理准则以及我国关于病历资料用于科研的相关管理规定。

1.2 研究人群

本研究选取NHANES数据库2007—2018年的调查数据,排除不符合MASLD诊断标准以及信息缺失的患者,最终纳入2 144例参与者,采用分层抽样将数据按照7∶3比例随机分为训练集和内部验证集,训练集用于构建模型,内部验证集进行内部验证。纳入标准:①肝脏脂肪变性,脂肪肝指数(fatty liver index,FLI)指数≥60;②心脏代谢危险因素(至少满足1项):a. 身体质量指数(body mass index ,BMI)≥25 kg/m^2,男性腰围≥94 cm,女性腰围≥80 cm;b. 空腹血糖≥6.1 mmol/L,餐后2 h血糖≥7.8 mmol/L,糖化血红蛋白≥5.7%,诊断为糖尿病或接受糖尿病的降糖治疗;

c. 血压≥130/85 mmHg或使用抗高血压药物；d. 总胆固醇≥1.70 mmol/L或使用降脂疗法；e. 高密度脂蛋白胆固醇水平,男性≤1.0 mmol/L,女性≤1.3 mmol/L,或正在接受降脂治疗的个体。排除标准:①合并慢性肾衰患者;②排除过量饮酒,男性>30 g/d,女性>20 g/d;③合并病毒性肝炎(乙型肝炎、丙型肝炎);④合并药物性肝损伤;⑤计算FLI和肾小球滤过率(glomerular filtration rate,eGFR)所需的其他指标相关数据缺失、人口统计学指标缺失、实验室检查缺失、慢性病史缺失的参与者;⑥年龄<20岁;⑦怀孕或妊娠妇女。

1.3 MASLD的评估

在本研究中,主要通过FLI定义肝脏脂肪变性并结合心脏代谢标准诊断MASLD。虽然超声检查在临床实践中被认为是一线筛查技术,但超声检查在NHANES数据库大部分年份中缺失,因此,在本研究中将FLI作为替代标志物,用于评估肝脂肪变性。计算FLI[17],若FLI>60,则提示该患者存在显著的脂肪肝风险[18]。该公式在既往研究中的受试者工作特征受试者工作特征(receiver operating characteristic,ROC)曲线下面积(area under curve,AUC)为0.844,亚洲男性和女性的阳性预测值分别为83.2%和84.8%,阴性预测值分别为65.3%和87.4%[19],预测性能良好。

1.4 CKD的评估

根据KDIGO CKD评估和管理实践指南诊断标准定义CKD[20]:肾脏结构或功能异常>3个月。即存在蛋白尿或eGFR<60 mL/(min·1.73 m^2)。蛋白尿定义为尿白蛋白排泄率≥30 mg/24 h或尿白蛋白肌酐比值≥30 mg/g。eGFR由CKD-EPI公式计算。

1.5 资料收集

从NHANES数据库收集患者的人口学信息,包括性别、年龄和吸烟等流行病学信息以及高血压、糖尿病、冠心病等慢性病史;测量数据,包括BMI、腰围、上臂围、收缩压(systolic blood pressure,SBP)、舒张压(diastolic blood pressure, DBP);实验室检测结果,包括总蛋白、白蛋白、球蛋白、总胆红素(total bilirubin, TB)、丙氨酸氨基转移酶(alanine aminotransferase, ALT)、天冬氨酸氨基转移酶(aspartate aminotransferase, AST)、空腹血糖、碱性磷酸酶(alkaline phosphatase, ALP)、血清铁、血尿素氮(blood urea nitrogen, BUN)、血尿酸、γ-谷氨酰转移酶(γ-glutamyl transferase, GGT)、血红蛋白(hemoglobin, Hb)、红细胞计数、淋巴细胞计数、单核细胞计数、中性粒细胞计数、血小板计数、糖化血红蛋白(glycated hemoglobin, HbA1c)、总胆固醇(total cholesterol, TC)、低密度脂蛋白胆固醇(low-density lipoprotein cholesterol, LDL-C)、甘油三酯(triglyceride, TG)、高密度脂蛋白胆固醇(high-density lipoprotein cholesterol,HDL-C)。

1.6 模型构建与评估

使用Lasso回归筛选出最佳特征变量。具体来说,首先通过R语言"glmnet"包实施Lasso回归,并执行包含1 000次重采样的Bootstrap程序,每次重采样生成与原始训练集同等规模的Bootstrap样本;其次,对每个Bootstrap样本进行十倍交叉验证,以确定最优正则化参数λ(即均方误差最小时对应的λ_{min}),并记录在该λ下系数非零的变量;最后,在所有1 000次Bootstrap抽样完成后,计算每个变量被选入模型的频率(即选择次数百分比)。变量筛选的最终决策标准为选择频率不低于90%的变量,而非单一Bootstrap样本中λ_{min}所对应的变量集合。为确保模型的稳健性及各预测变量之间的独立性,避免多重共线性导致的模型估计失真,进一步对这些变量进行了方差膨胀因子(variance inflation factor, VIF)检验。

将最佳特征变量分别纳入决策树(decision trees, DT)、极端梯度提升机(extreme gradient boosting, XGBoost)、K-最近邻算法(K-nearest neighbors,KNN)、朴素贝叶斯(Naive Bayes, NB)、支持向量机(support vector machine,SVM)、单隐藏层神经网络(neural networks, NNET)、梯度提升机(light gradient boosting machine,LightGBM)、随机森林(random forest,RF)8种机器学习方法构建机器学习模型。采用ROC的AUC、阳性预测值、阴性预测值、准确率、敏感度、特异度、F1值、Brier评分评价模型预测性能。使用"shapviz"包绘制SHAP图以解释最优机器学习模型特征变量对预测结果的贡献度,进一步采用校准曲线与临床决策曲线分析,分别评估最优模型的校准度与临床实用性。同时通过SHAP值对生物标志物重要性排名后,选取重要的生物标志物构建诊断列线图,为客观确定风险分层的最佳截断值,基于患者的生存时间数据,使用X-tile软件(version 3.6.1)对所有患者的列线图风险评分进行系统分析。该软件通过评估所有可能的分割点,寻找能使所定义的高、低风险组之间生存差异最大化(以Log-rank检验的χ^2值为标准)的最优截断值,并通过限制性立方样条(restricted cubic spline,RCS)评估风险评分与CKD风险之间的剂量反应关系。

1.7 统计学分析

本研究的数据分析在RStudio集成开发环境中使用R语言(version 4.4.1),其中符合正态分布的连续变量以$\bar{x}\pm s$表示,组间比较采用独立样本t检验;不符合正态分布的连续变量以$M(P_{25},P_{75})$表示,组间比较采用Mann-Whitney U检验。分类变量以例(%)表示,

组间比较采用χ^2检验或Fisher精确检验。检验水准α=0.05。数据中的缺失值采用直接剔除该病例的方法。

2 结果

2.1 MASLD患者中CKD和非CKD患者的基线特征

本研究共纳入2 144例符合诊断的MASLD患者,根据有无CKD,将患者分为CKD组(n=347)和无CKD组(n=1 797),CKD组和非CKD组在总蛋白、白蛋白、球蛋白、ALT、ALP、血清铁、血尿素氮、血尿酸、BMI、腰围、收缩压、血红蛋白、白细胞、中性粒细胞计数、葡萄糖、糖化血红蛋白、高密度脂蛋白、总胆固醇、低密度脂蛋白、全身免疫炎症指数等指标差异有统计学意义(P<0.05);与非CKD组相比,诊断为CKD的患者往往女性比例更低,年龄更大,糖尿病、高血压患病率更高,如表1所示。

表1 MASLD是否合并CKD的患者基线特征[$M(P_{25},P_{75})$, 例(%)]

Table 1 Baseline characteristics of MASLD patients with or without CKD[$M(P_{25},P_{75})$, n(%)]

Variables	All(n=2 144)	No-CKD(n=1 797)	CKD(n=347)	P
Gender				<0.001
Female	501(23.4)	452(25.2)	49(14.1)	
Male	1 643(76.6)	1 345(74.8)	298(85.9)	
Age/years	43.00(33.00,56.00)	42.00(32.00,53.00)	57.00(41.00,71.00)	<0.001
Smoke				<0.001
Current smoker	422(19.7)	357(19.9)	65(18.7)	
Former smoker	620(28.9)	488(27.2)	132(38.0)	
Never smoker	1 102(51.4)	952(53.0)	150(43.2)	
Hypertension				<0.001
No	1 294(60.4)	1 164(64.8)	130(37.5)	
Yes	850(39.6)	633(35.2)	217(62.5)	
CHD				<0.001
No	2 041(95.2)	1 736(96.6)	305(87.9)	
Yes	103(4.8)	61(3.4)	42(12.1)	
Diabetes				<0.001
No	1 831(85.4)	1 606(89.4)	225(64.8)	
Yes	313(14.6)	191(10.6)	122(35.2)	
Waist circumference/cm	110.80(104.30, 119.70)	110.09(104.00, 119.00)	115.50(107.40, 127.50)	<0.001
BMI/(kg/m^2)	32.80(29.88, 36.50)	32.70(29.80, 36.26)	33.82(30.50, 39.31)	0.026
SBP/mmHg	121.33(113.33, 130.00)	120.00(112.67, 128.67)	129.33(118.80, 141.33)	<0.001
Albumin/(g/dL)	4.30(4.00, 4.40)	4.30(4.10, 4.50)	4.20(4.00, 4.40)	0.004
Globulin/(g/dL)	2.80(2.60, 3.10)	2.80(2.60, 3.10)	3.00(2.70, 3.40)	<0.001
ALT/(U/L)	26.00(20.00,36.00)	27.00(20.00, 36.00)	24.00(18.00, 32.00)	0.001
ALP/(U/L)	65.00(55.00, 79.00)	65.00(54.00, 78.00)	68.00(55.47, 88.00)	0.013
Serum iron/(μg/dL)	87.00(67.00, 111.00)	88.00(68.00, 112.00)	81.94(62.00, 104.00)	0.013
BUN/(mg/dL)	12.00(10.00, 15.00)	12.00(10.00, 15.00)	14.00(11.00, 19.00)	<0.001
Uric acid/(mg/dL)	6.20(5.40, 7.00)	6.10(5.30, 7.00)	6.40(5.60, 7.50)	0.001
Total protein/(g/dL)	7.10(6.80, 7.40)	7.10(6.80, 7.40)	7.20(6.90, 7.60)	0.002
Hb/(g/dL)	15.00(14.10, 15.90)	15.10(14.10, 15.90)	14.70(13.60, 15.80)	0.039
Lymphocyte/(10^9/L)	2.00(1.60, 2.40)	2.00(1.70, 2.40)	1.90(1.50, 2.20)	0.004
Neutrophils/(10^9/L)	4.00(3.20, 5.00)	4.00(3.20, 4.90)	4.30(3.50, 5.40)	<0.001
Glucose/(mg/dL)	103.00(96.00, 113.00)	102.00(96.00, 111.00)	111.00(100.00, 130.00)	<0.001
HbA1c/%	5.50(5.20, 5.80)	5.50(5.20, 5.80)	5.80(5.40, 6.70)	<0.001
HDL-C/(mg/dL)	43.00(37.00, 50.00)	44.00(38.00, 50.00)	41.00(36.00, 48.00)	0.008
TC/(mg/dL)	191.00(166.00, 217.00)	194.00(167.00, 218.00)	180.00(155.00, 209.49)	<0.001
LDL-C/(mg/dL)	116.00(95.00, 139.00)	117.00(96.00, 139.00)	107.00(80.58, 131.00)	<0.001
SII/(10^9/L)	61.82(46.67, 80.76)	61.22(46.22, 79.47)	66.32(51.06, 91.25)	0.006

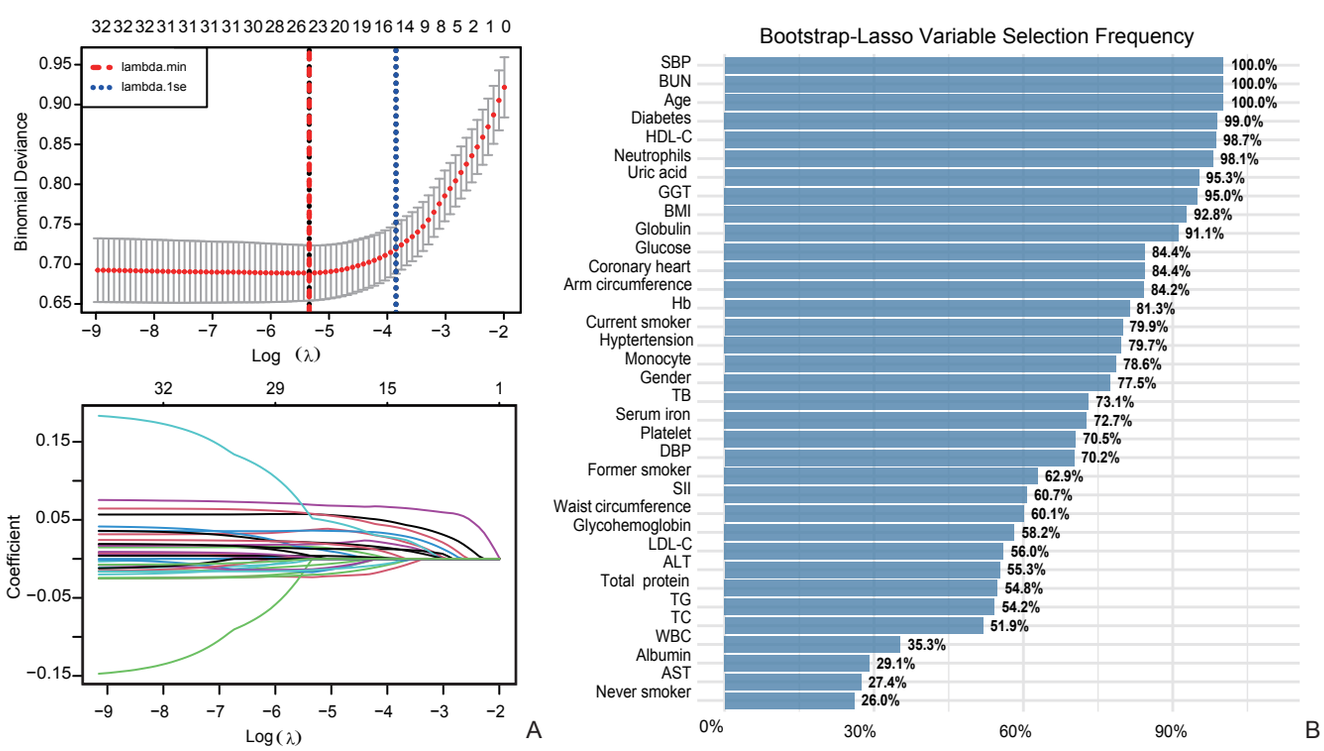
随后将2 144例患者以7∶3的比例随机分为训练集和内部验证集,训练集1 501例,内部验证集643例。训练集和内部验证集在人口学统计指标、实验室指标、慢性病史等指标中均无统计学差异,表明2个队列具有相似的人口指标和临床指标。

2.2　筛选可预测MASLD患者发生CKD的生物标志物

在训练集中,通过对34个临床变量进行基于Bootstrap的Lasso回归稳定性选择,最终识别出10个具有高度稳定性的预测变量。这些变量在1 000次Bootstrap重采样中的被选频率均不低于90%,表明其对于模型预测贡献稳健,被纳入后续机器学习建模。具体变量见图1,包括SBP、BUN、年龄、糖尿病史、HDL-C、中性粒细胞计数、血尿酸、GGT、BMI、球蛋白。VIF检验结果显示,所有变量的VIF值远低于常用的临界值(5)。具体而言,各变量的VIF值如下:年龄(1.599)、BMI(1.512)、糖尿病史(1.510)、BUN(1.411)、HDL-C(1.385)、GGT(1.317)、球蛋白(1.231)、血尿酸(1.261)、SBP(1.140)及中性粒细胞计数(1.135)。这些结果充分表明,纳入模型的变量间不存在严重的多重共线性问题,各变量所提供的信息具有较好的独立性,因此均适合纳入后续机器学习分析。

2.3　构建预测MASLD患者CKD风险的机器学习模型

基于Lasso回归筛选出的10个特征变量,进一步构建DT、XGBoost、KNN、NB、SVM、NNET、LightGBM、RF 8种预测模型。各模型在训练集上训练后,于测试集中进行评估,其性能指标包括准确率、敏感度、特异度、Brier分数、F1分数、ROC曲线下面积,结果如表2所示。ROC曲线分析显示,在训练集中,RF模型对预测MASLD患者发生CKD的效能最好(AUC=0.958),其次是LightGBM模型(AUC=0.871)、KNN模型(AUC=0.859)、XGBoost模型(AUC=0.846)、SVM模型(AUC=0.803)、NB模型(AUC=0.795)、DT模型(AUC=0.791)、NNET模型(AUC=0.781)。在验证集中,RF模型AUC降至0.783,表明该模型存在严重过拟合,不适宜选择其作为最优模型;而LightGBM模型的AUC值(AUC=0.787)在训练集与测试集下降相对合理,表明具有良好拟合性能,见图2。综合评估,LightGBM模型在训练集和测试集均具有较好的性能,测试集中其准确率为0.790,敏感度为0.821,特异度为0.624,F1得分为0.868,Brier得分为0.107,为本研究预测MASLD患者CKD危险因素的最佳机器学习算法。进一步绘制LightGBM模型分别在内部验证集及外部验



A: Ten-fold cross-validation curve of the Lasso regression model (top) and variable coefficient shrinkage path plot (bottom); B: Variable selection frequency of Lasso regression based on 1 000 bootstrap resamples

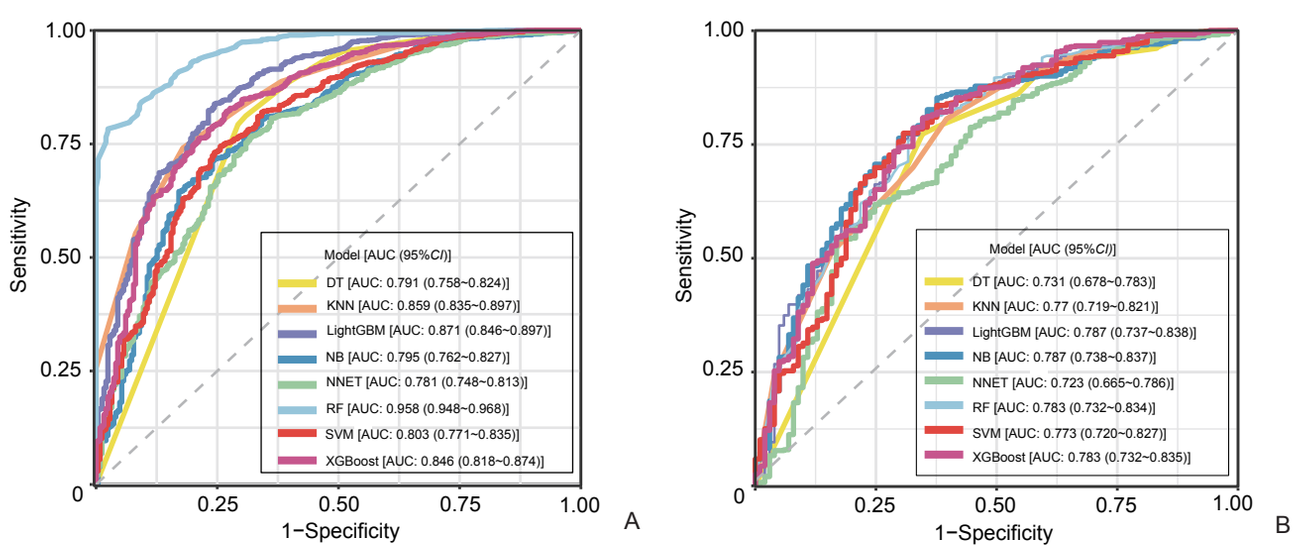
图1　Lasso回归筛选关键标志物

Figure 1　Lasso regression screening of key biomarkers

表 2 预测MASLD患者CKD风险的各机器学习模型预测性能

Table 2 Model performance of machine learning algorithms for predicting CKD risk in MASLD patients

Datasets	Models	AUC(95%*CI*)	Accuracy	Sensitivity	Specificity	F1 score	Brier score
Training set	DT	0.791(0.758~0.824)	0.792	0.812	0.691	0.867	0.096
	NB	0.795(0.762~0.827)	0.676	0.645	0.829	0.769	0.131
	RF	0.958(0.948~0.968)	0.815	0.783	0.976	0.876	0.065
	XGBoost	0.846(0.818~0.874)	0.807	0.822	0.732	0.877	0.096
	SVM	0.803(0.771~0.835)	0.745	0.745	0.744	0.830	0.116
	NNET	0.781(0.748~0.813)	0.751	0.764	0.683	0.837	0.150
	LightGBM	0.871(0.846~0.897)	0.825	0.838	0.756	0.889	0.091
	KNN	0.859(0.835~0.897)	0.754	0.741	0.821	0.834	0.099
Internal validation set	DT	0.731(0.678~0.783)	0.765	0.793	0.614	0.851	0.116
	NB	0.787(0.738~0.837)	0.816	0.852	0.624	0.887	0.133
	RF	0.783(0.732~0.834)	0.670	0.657	0.743	0.771	0.107
	XGBoost	0.783(0.732~0.835)	0.781	0.804	0.653	0.861	0.106
	SVM	0.773(0.720~0.827)	0.723	0.723	0.723	0.815	0.116
	NNET	0.723(0.665~0.786)	0.725	0.753	0.574	0.822	0.154
	LightGBM	0.787(0.737~0.838)	0.790	0.821	0.624	0.868	0.107
	KNN	0.770(0.719~0.821)	0.695	0.699	0.673	0.795	0.110
External validation set	LightGBM	0.740(0.660~0.820)	0.773	0.795	0.730	0.822	0.141



A: ROC curve of the training set; B: ROC curve of the internal validation set

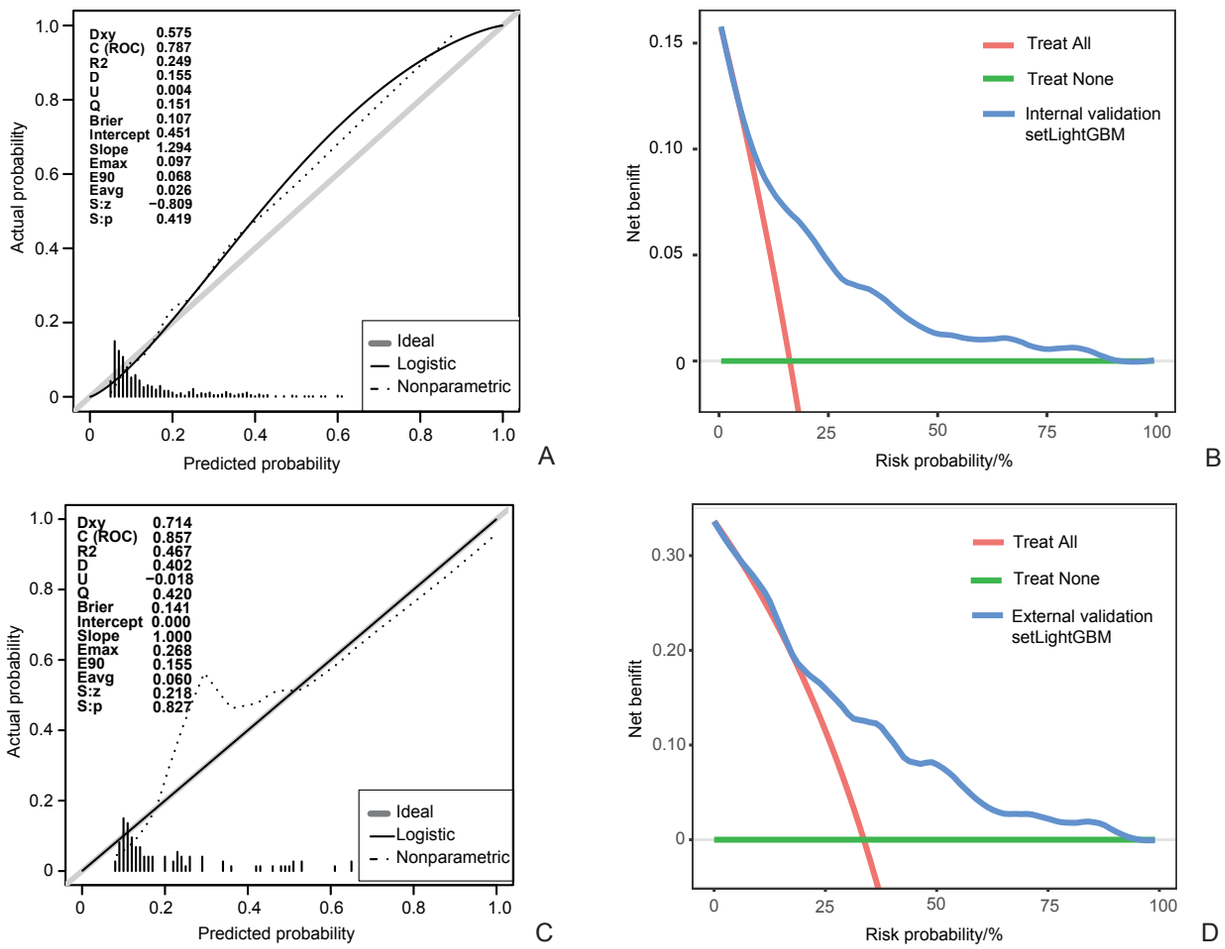
图 2 MASLD中患者CKD机器学习模型的ROC曲线

Figure 2 ROC curves of machine learning models for CKD in MASLD patients

证集的校准曲线与DCA曲线评估其模型的拟合程度及临床获益,如图3所示,校准曲线表明预测值与实际风险值一致。决策曲线分析提示,无论是在内部验证集还是外部验证集中,LightGBM模型的净获益均有20%~90%的阈值概率,超过了全治疗和不治疗策略,其在准确预测MASLD患者CKD风险的良好临床效用。因此,选择LightGBM模型作为本研究的进一步预测模型。

2.4 解释预测MASLD患者CKD风险的生物标志物

使用SHAP算法获取每个预测变量对LightGBM模型预测的结果重要性。变量重要性图显示收缩压在所有预测范围内具有最高的预测值,紧随其后的是年龄、糖尿病史、血尿素氮、BMI、球蛋白、高密度脂蛋白胆固醇、中性粒细胞计数等,而血尿酸、GGT水平对MASLD患者发生CKD的影响较弱,总的来说,在MASLD患者中预测CKD风险,其预测风险主要受8个



A: Calibration curve of the internal validation set for the LightGBM model; B: DCA clinical decision curve; C: Calibration curve of the external validation set for the LightGBM model; D: DCA clinical decision curve for the external validation set

图3 机器学习最优模型LightGBM模型内部验证集效能验证

Figure 3 Performance evaluation of the LightGBM model on internal and external validation sets

因素影响:收缩压、年龄、糖尿病史、血尿素氮、BMI、球蛋白、高密度脂蛋白胆固醇、中性粒细胞计数,见图4。

通过SHAP分析揭示了各特征变量对MASLD患者CKD风险的贡献方向与程度。本研究结果显示收缩压、年龄、糖尿病史、血尿素氮、BMI等对MASLD患者CKD风险有正向作用,即年龄越大,有糖尿病史、收缩压越高、血尿素氮越高,该MASLD患者CKD风险越高;而高密度脂蛋白胆固醇越低,MASLD患者CKD风险则越高,结果见图5。

根据SHAP汇总图,提取对结果贡献较大的前4个临床特征的SHAP依赖图,以解释临床特征影响MASLD患者发生CKD风险的内在机制,如图6所示,年龄、糖尿病史、收缩压水平、血尿素氮均与MASLD患者CKD的发生风险呈正相关。

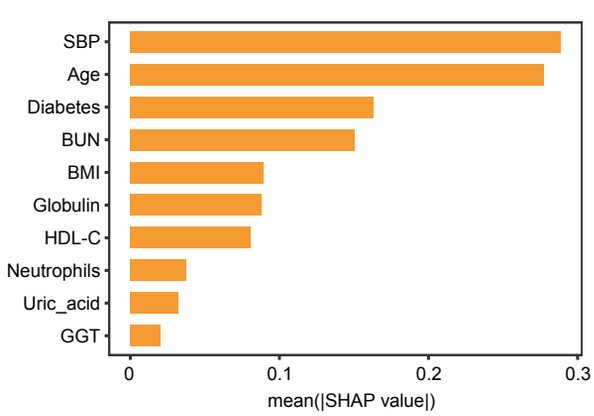


图4 最优机器学习算法LightGBM模型中的特征变量重要性排名

Figure 4 Feature importance ranking in the optimal LightGBM model for predicting chronic kidney disease risk in MASLD patients

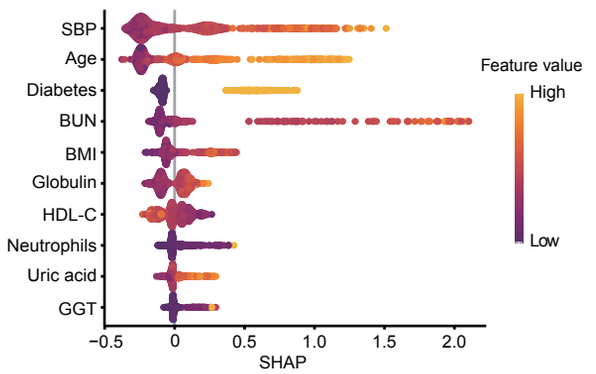
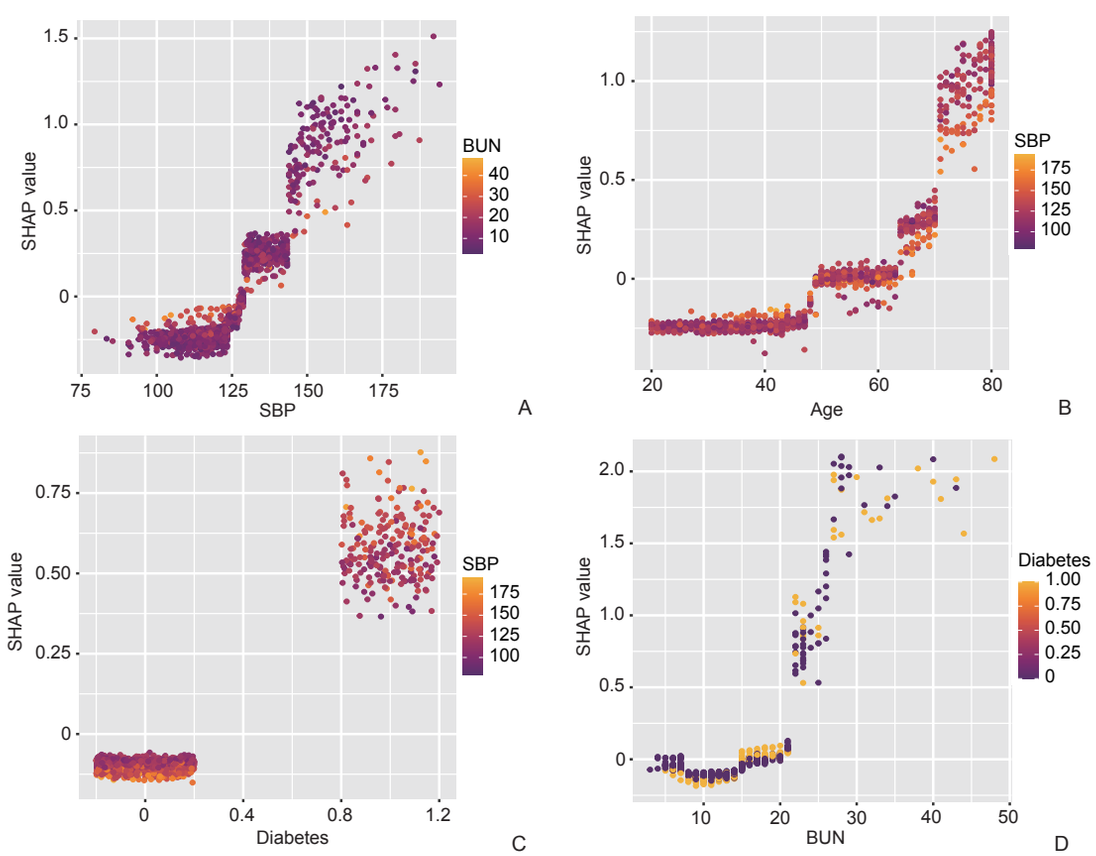


图5 LightGBM模型的10个临床特征SHAP汇总图

Figure 5 SHAP summary plot of 10 clinical features in the LightGBM model

2.5 构建诊断列线图以预测MASLD患者CKD的风险

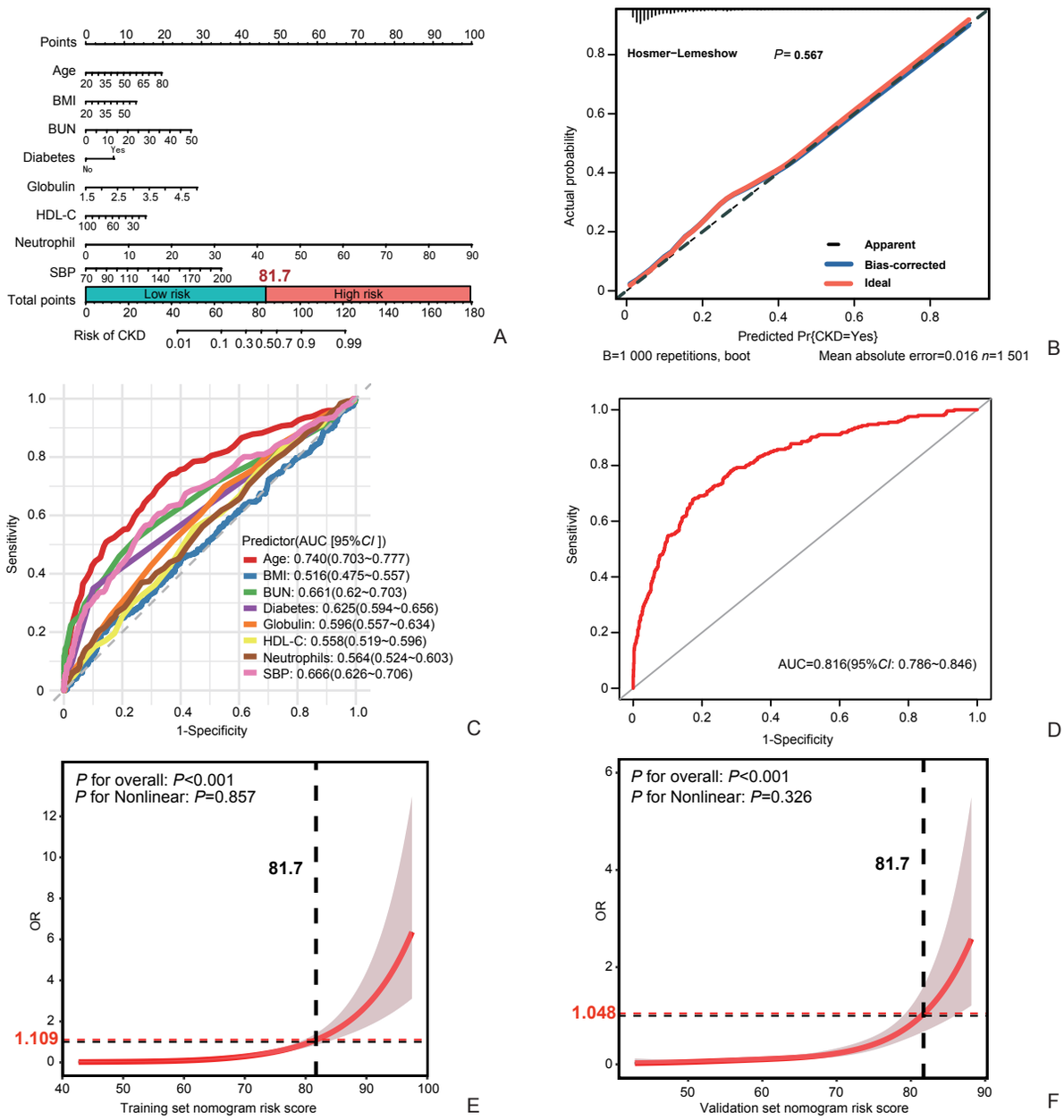
本研究基于SHAP值对预测模型的特征重要性进行量化评估,以筛选关键预测因子。图7展示了在构建临床诊断列线图以预测MASLD患者CKD风险过程中,根据SHAP值排名前8位的生物标志物及其重要性概况。软件分析结果明确显示,风险评分为81.7分时,Log-rank检验的χ^2值为268.23,P<0.000 1,表明在此分割点下2组患者的生存预后具有显著差异。因此,将81.7确定为风险分层阈值,当总分≥81.7时,可预测其MASLD患者CKD发病风险较高。通过诊断校准曲线的进一步验证,结果显示与理想线的偏差最小,证实了该模型在MASLD患者CKD风险临床预测中的精度和实用性,同时ROC曲线评估了综合预测准确性,发现当这8种生物标志物一起使用时,AUC显著改善,AUC为0.816。使用RCS分析列线图总分与MASLD患者CKD患病率的剂量反应关系,结果提示,列线图总分与CKD患病率存在线性关系(总体P<0.05,非线性P>0.05),总分越高,MASLD患者CKD发生风险越高,说明风险评分的变化能够显著预测CKD发生风险的变化。且当总评分为81.7分时,在内部和外部验证集中,MASLD患者CKD患病率OR为1.05,证明其最佳风险阈值可对MASLD患者CKD风险进行恰当分层。



A: Systolic blood pressure; B: Age; C: History of diabetes; D: Blood urea nitrogen

图6 LightGBM模型前4位SHAP值特征依赖图

Figure 6 SHAP dependence plots of the top 4 features in the LightGBM model



A: Nomogram for diagnosing CKD; B: Calibration curve of predictive accuracy; C: ROC curves of 8 factors; D: ROC curve of the diagnostic nomogram; E: Relationship between total nomogram score and CKD prevalence in MASLD patients in the training set; F: Relationship between total nomogram score and CKD prevalence in MASLD patients in the internal validation set

图7 MASLD患者CKD诊断列线图的构建

Figure 7 Construction of a nomogram for the diagnosis of CKD in MASLD patients

3 讨论

MASLD是一种典型的多系统疾病,其通过脂毒性炎症溢出效应,将肝脏局部的炎症和氧化应激反应扩散至全身,导致心血管、肌肉及肾脏等多器官损伤。目前,MASLD与心血管疾病、2型糖尿病的关联已较为明确,但其与肾损害之间的病理生理机制尚未完全阐明。同时由于肾损害早期症状隐匿,在MASLD与CKD共存时更易被忽视。因此,筛选敏感且特异的生物标志物,用于早期评估MASLD患者的CKD风险,具有重要的临床价值。

3.1 预测模型的性能与关键生物标志物的临床意义

本研究基于大规模人群数据,通过比较多种机器学习算法,构建了MASLD患者并发CKD的风险预测模型。结果表明,LightGBM模型在多项性能指标上均表现最优,显示出良好的判别能力与临床适用性。模

型中筛选出的关键生物标志物——SBP、年龄、糖尿病史、BUN、BMI、球蛋白、HDL-C和中性粒细胞计数,不仅涵盖了代谢、炎症、营养等多方面因素,也反映了MASLD与CKD共存的复杂病理生理背景,为早期识别肾损伤提供了新的视角。

3.2 关键生物标志物的病理生理学机制探讨

本研究验证了若干已知与CKD密切相关的代谢性危险因素[21-23],包括糖尿病史、高血压、血脂异常,这些因素共同指向胰岛素抵抗作为MASLD与CKD共同的核心病理机制。除代谢因素外,本研究也提示炎症与免疫反应在MASLD相关肾损害中的重要作用。球蛋白水平升高通常反映机体处于慢性炎症状态[24],而中性粒细胞作为先天免疫的关键细胞,可能通过释放中性粒细胞胞外诱捕网(neutrophil extracellular traps, NETs)等机制介导肾脏微血管内皮损伤与纤维化进程[25]。年龄在本模型中作为最具影响力的预测因子,与近年提出的器官间衰老信号传递理论相呼应。研究表明,肝脏中累积的衰老细胞可分泌TGF-β等因子[26],远程促进肾脏组织衰老及功能减退,这为理解MASLD与CKD共病机制提供了新视角。

3.3 本研究的创新性与临床价值

本研究的创新点在于:①首次系统性地将多种机器学习算法应用于MASLD患者CKD风险的预测;进一步构建的诊断列线图使临床医生可根据81.7分的风险阈值快速识别高危人群,实现早期干预。②与传统单一指标或简单评分系统相比,本模型整合多维度指标,显著提升了对CKD风险的分层能力。③模型在独立外部验证队列中仍保持良好性能,说明其具有较强的泛化能力,适用于不同人群。但本研究存在一定的局限性。在MASLD的诊断方面,本研究采用FLI指数作为肝脂肪变性的替代评估指标,而未采用更精确的影像学检查或肝穿刺活检这一金标准进行确认[27]。

综上所述,本研究构建的LightGBM模型能够有效预测MASLD患者并发CKD的风险,可作为临床辅助诊断工具帮助早期识别高危人群。未来研究可通过纳入更多风险因素,并开展多中心大样本队列验证,以进一步提升模型精度并探索更易于临床推广的生物标志物。

利益冲突 本文所有作者声明不存在利益冲突

参考文献:

[1] RINELLA M E, LAZARUS J V, RATZIU V, *et al*. A multisociety Delphi consensus statement on new fatty liver disease nomenclature[J]. J Hepatol, 2023, 79(6): 1542-1556. DOI:10.1016/j.jhep.2023.06.003.

[2] YOUNOSSI Z M, KALLIGEROS M, HENRY L. Epidemiology of metabolic dysfunction-associated steatotic liver disease[J]. Clin Mol Hepatol, 2025, 31: S32-S50. DOI:10.3350/cmh.2024.0431.

[3] WANG Z L, XU M, HU Z G, *et al*. Prevalence of nonalcoholic fatty liver disease and its metabolic risk factors in women of different ages and body mass index[J]. Menopause, 2015, 22(6): 667-673. DOI: 10.1097/GME.0000000000000352.

[4] MIAO L, TARGHER G, BYRNE C D, *et al*. Current status and future trends of the global burden of MASLD[J]. Trends Endocrinol Metab, 2024, 35(8): 697-707. DOI:10.1016/j.tem.2024.02.007.

[5] TARGHER G, TILG H, BYRNE C D. Non-alcoholic fatty liver disease: a multisystem disease requiring a multidisciplinary and holistic approach[J]. Lancet Gastroenterol Hepatol, 2021, 6(7): 578-588. DOI:10.1016/S2468-1253(21)00020-0.

[6] SUN D Q, TARGHER G, BYRNE C D, *et al*. An international Delphi consensus statement on metabolic dysfunction-associated fatty liver disease and risk of chronic kidney disease[J]. Hepatobiliary Surg Nutr, 2023, 12(3): 386-403. DOI: 10.21037/hbsn-22-421.

[7] JIANG F Y, WANG L J, YING H C, *et al*. Multisystem health comorbidity networks of metabolic dysfunction-associated steatotic liver disease[J]. Med, 2024, 5(11): 1413-1423.e3. DOI:10.1016/j.medj.2024.07.013.

[8] TARGHER G, BYRNE C D, TILG H. MASLD: a systemic metabolic disorder with cardiovascular and malignant complications[J]. Gut, 2024, 73(4): 691-702. DOI: 10.1136/gutjnl-2023-330595.

[9] WEBSTER A C, NAGLER E V, MORTON R L, *et al*. Chronic kidney disease[J]. Lancet, 2017, 389(10075): 1238-1252. DOI:10.1016/S0140-6736(16)32064-5.

[10] ROMAGNANI P, AGARWAL R, CHAN J C N, *et al*. Chronic kidney disease[J]. Nat Rev Dis Primers, 2025, 11: 8. DOI:10.1038/s41572-024-00589-9.

[11] ZHOU X D, ZHENG M H. Cardiovascular-kidney-metabolic syndrome and MASLD: integrating medical perspectives[J]. Nat Rev Cardiol, 2025, 22(11): 843. DOI: 10.1038/s41569-025-01199-y.

[12] MANTOVANI A, ZAZA G, BYRNE C D, *et al*. Nonalcoholic fatty liver disease increases risk of incident chronic kidney disease: a systematic review and meta-analysis[J]. Metabolism, 2018, 79: 64-76. DOI:10.1016/j.metabol.2017.11.003.

[13] WEI S S, SONG J, XIE Y J, *et al*. Metabolic dysfunction-

associated fatty liver disease can significantly increase the risk of chronic kidney disease in adults with type 2 diabetes [J]. Diabetes Res Clin Pract, 2023, 197: 110563. DOI: 10.1016/j.diabres.2023.110563.

[14] GREENER J G, KANDATHIL S M, MOFFAT L, *et al*. A guide to machine learning for biologists [J]. Nat Rev Mol Cell Biol, 2022, 23(1): 40-55. DOI: 10.1038/s41580-021-00407-0.

[15] CHOWDHURY S U, SAYEED S, RASHID I, *et al*. Shapley-additive-explanations-based factor analysis for dengue severity prediction using machine learning [J]. J Imaging, 2022, 8 (9): 229. DOI: 10.3390/jimaging8090229.

[16] CURTIN L R, MOHADJER L K, DOHRMANN S M, *et al*. National health and nutrition examination survey: sample design, 2007-2010 [J]. Vital Health Stat 2, 2013(160): 1-23.

[17] BEDOGNI G, BELLENTANI S, MIGLIOLI L, *et al*. The fatty liver index: a simple and accurate predictor of hepatic steatosis in the general population [J]. BMC Gastroenterol, 2006, 6(1): 33. DOI: 10.1186/1471-230X-6-33.

[18] CARLI F, SABATINI S, GAGGINI M, *et al*. Fatty liver index (FLI) identifies not only individuals with liver steatosis but also at high cardiometabolic risk [J]. Int J Mol Sci, 2023, 24(19): 14651. DOI: 10.3390/ijms241914651.

[19] NOMURA T, ONO M, KOBAYASHI K, *et al*. Validation of fatty liver index as a predictor of hepatic steatosis in Asian populations: Impact of alcohol consumption and sex [J]. Hepatol Res, 2023, 53 (10): 968-977. DOI: 10.1111/hepr.13935.

[20] WANG Y Y, YU Y T, ZHANG H J, *et al*. Cardiovascular and renal burdens among patients with MAFLD and NAFLD in China [J]. Front Endocrinol, 2022, 13: 968766. DOI: 10.3389/fendo.2022.968766.

[21] NDUMELE C E, NEELAND I J, TUTTLE K R, *et al*. A synopsis of the evidence for the science and clinical management of cardiovascular-kidney-metabolic (CKM) syndrome: a scientific statement from the American Heart Association [J]. Circulation, 2023, 148(20): 1636-1664. DOI: 10.1161/CIR.0000000000001186.

[22] THEOFILIS P, VORDONI A, KALAITZIDIS R G. Interplay between metabolic dysfunction-associated fatty liver disease and chronic kidney disease: Epidemiology, pathophysiologic mechanisms, and treatment considerations [J]. World J Gastroenterol, 2022, 28 (39): 5691-5706. DOI: 10.3748/wjg.v28.i39.5691.

[23] BYRNE C D, TARGHER G. NAFLD as a driver of chronic kidney disease [J]. J Hepatol, 2020, 72(4): 785-801. DOI: 10.1016/j.jhep.2020.01.013.

[24] PARK J, KIM H J, KIM J, *et al*. Predictive value of serum albumin-to-globulin ratio for incident chronic kidney disease: a 12-year community-based prospective study [J]. PLoS One, 2020, 15(9): e0238421. DOI: 10.1371/journal.pone.0238421.

[25] JORCH S K, KUBES P. An emerging role for neutrophil extracellular traps in noninfectious disease [J]. Nat Med, 2017, 23(3): 279-287. DOI: 10.1038/nm.4294.

[26] KIOURTIS C, TERRADAS-TERRADAS M, GEE L M, *et al*. Hepatocellular senescence induces multi-organ senescence and dysfunction *via* TGFβ [J]. Nat Cell Biol, 2024, 26 (12): 2075-2083. DOI: 10.1038/s41556-024-01543-3.

[27] CHUNG G E, JEONG S M, CHO E J, *et al*. The association of fatty liver index and BARD score with all-cause and cause-specific mortality in patients with type 2 diabetes mellitus: a nationwide population-based study [J]. Cardiovasc Diabetol, 2022, 21 (1): 273. DOI: 10.1186/s12933-022-01691-6.

(**收稿**:2025-10-15;**修回**:2026-01-12)

(编辑 邓强庭)