· 流行病与统计方法 ·

基于机器学习的肝硬化发病风险预测模型研究

马永华[1],杨欣妤[1],王艳霞[1],宋旺辰[1],郭贵雅[1],王爱民[1], 石福艳[1],于兆锋[2]
1. 山东第二医科大学公共卫生学院卫生统计学教研室,山东 潍坊 261053;2. 山东第二医科大学人事处

摘要:目的 开发并验证适用于一般人群的肝硬化发病风险预测模型,揭示常见生物标志物与肝硬化发病风险之间的关联,为临床和公共卫生实践中肝硬化发病风险评估提供科学依据。**方法** 本研究使用来自英国生物样本库(UK Biobank)的 28 种生物标志物和 6 种协变量数据。参与者于 2006 年至 2010 年间入组,随访至 2022 年 10 月 31 日,主要结果是新发肝硬化病例。将所有满足纳入排除要求的参与者,按 7∶3比例随机分为训练集和测试集。基于训练集,采用 XGBoost 算法、Lasso – Cox 回归和随机生存森林三种机器学习算法来筛选影响肝硬化发生的关键预测变量,并利用筛选出的关键特征变量构建基于 Cox 比例风险模型的肝硬化发病风险预测模型。通过与传统评分 FIB – 4 评分、APRI 评分进行比较,评估预测模型的性能,并在验证集中进行验证。**结果** 本研究共筛选了 237 965 名符合条件的研究对象进行研究,其中肝硬化患者 1 064 名,非肝硬化患者 236 801 名。结合三种机器学习方法和临床应用情况共筛选出 5 个影响肝硬化发生的关键变量,分别为 γ – 谷氨酰转移酶、天冬氨酸氨基转移酶、胱抑素 C、体质量指数、白蛋白。基于这 5 种特征变量构建的肝硬化发病风险预测模型,预测性能显著超过传统 FIB – 4 评分模型和 APRI 评分模型。本研究构建的肝硬化发病风险预测模型、FIB – 4 评分模型和 APRI 评分模型,在训练集中 AUC 分别为 0.827(0.806 ~ 0.848)、0.665(0.637 ~ 0.693)、0.740(0.713 ~ 0.766);在验证集中,AUC 分别 0.811(0.776 ~ 0.845)、0.664(0.621 ~ 0.708)、0.711(0.668 ~ 0.755)。本研究建立的预测模型的 AUC 均优于两种传统评分预测模型,差异均具有统计学意义($P < 0.05$)。本研究构建的肝硬化发病风险预测模型,通过与传统的 FIB – 4 评分模型、APRI 评分模型对比,在净重新分类指数(NRI)和综合判别改善指数(IDI)两个关键指标上均表现出显著优势($P < 0.001$),且在训练集和测试集中的结果一致性较高。肝硬化发病风险预测模型结果显示,胱抑素 C(CysC)在肝纤维化疾病的预测中具有较高的特异性和预测价值[$HR = 2.51$,95% CI: 2.22 ~ 2.83]。**结论** γ – 谷氨酰转移酶、天冬氨酸氨基转移酶、胱抑素 C、体质量指数和白蛋白是影响肝硬化发病的重要因素。利用机器学习多重筛选得到的特征变量构建的肝硬化发病风险的预测模型,具有较强的预测能力,可为肝硬化发病风险的早期筛查提供有效工具。

关键词:肝硬化;生物标志物;机器学习;预测模型
中图分类号:R575.2 **文献标志码:**A **文章编号:**1003 – 8507(2025)22 – 4038 – 09
DOI:10.20043/j.cnki.MPM.202506471

Research on the prediction model of Liver Cirrhosis incidence risk based on machine learning

MA Yong – hua*, YANG Xin – yu, WANG Yan – xia, SONG Wang – chen, GUO Gui – ya, WANG Ai – min, SHI Fu – yan, YU Zhao – feng
* *Department of Health Statistics, School of Public Health, Shandong Second Medical University, Weifang, Shandong* 261053, *China*

Abstract: Objective To develop and validate a risk prediction model for liver cirrhosis applicable to the general population, to reveal the association between common biomarkers and the risk of liver cirrhosis, and to provide a scientific basis for the assessment of liver cirrhosis risk in clinical and public health practice. **Methods** This study utilized data on 28 biomarkers and 6 covariates from the UK Biobank. Participants were enrolled between 2006 and 2010 and followed up until October 31, 2022. The primary outcome was new cases of liver cirrhosis. All participants meeting the inclusion and exclusion criteria were randomly divided into a training set and a test set at a ratio of 7∶3. Based on the training set, three machine learning algorithms, namely XGBoost, Lasso – Cox regression, and random survival forest, were used to screen for key predictive variables influencing the occurrence of liver cirrhosis. Key feature variables were then used to construct a liver cirrhosis risk

基金项目:国家自然科学基金项目(81803337,81872719);山东省自然科学基金项目(ZR2023MH313);山东省高等学校青创人才引育计划项目(2019 – 6 – 156,Lu – Jiao);山东省本科高校教学改革研究项目(M2024137)
作者简介:马永华(2001—),男,硕士在读,研究方向:卫生统计学
通信作者:石福艳和于兆锋为共同通信作者。石福艳,E – mail:shifuyan@sdmu.edu.cn;于兆锋,E – mail:yzhf@sdsmu.edu.cn

prediction model based on the Cox proportional hazards model. The performance of the prediction model was evaluated by comparing it with traditional scores, FIB - 4 score and APRI score, and validated in the validation set. **Results** A total of 237 965 eligible subjects were screened for this study, including 1 064 patients with liver cirrhosis and 236 801 non - liver cirrhosis patients. Five key variables influencing the occurrence of liver cirrhosis were identified by combining three machine learning methods and clinical application scenarios, namely γ - glutamyl transferase, aspartate aminotransferase, cystatin C, body mass index, and albumin. The liver cirrhosis risk prediction model constructed based on these five characteristic variables significantly outperformed the traditional FIB - 4 score model and APRI score model in terms of predictive performance. The AUC values of the liver cirrhosis risk prediction model, FIB - 4 score model, and APRI score model in the training set were 0. 827 (0. 806 - 0. 848), 0. 665 (0. 637 - 0. 693), and 0. 740 (0. 713 - 0. 766), respectively; in the validation set, they were 0. 811 (0. 776 - 0. 845), 0. 664 (0. 621 - 0. 708), and 0. 711 (0. 668 - 0. 755), respectively. The AUC values of the prediction model established in this study were superior to those of the two traditional scoring prediction models, and the differences were statistically significant ($P < 0.05$). Compared with the traditional FIB - 4 score model and APRI score model, the liver cirrhosis risk prediction model constructed in this study demonstrated significant advantages in the two key indicators of net reclassification index (NRI) and integrated discrimination improvement index (IDI) ($P < 0.001$), and the results were highly consistent in both the training set and the test set. The results of the liver cirrhosis risk prediction model showed that cystatin C had high specificity and predictive value in the prediction of liver fibrosis diseases [$HR = 2.51$, 95% CI: 2. 22 - 2. 83]. **Conclusion** γ - glutamyl transferase, aspartate aminotransferase, cystatin C, body mass index and albumin are important factors influencing the onset of liver cirrhosis. The prediction model for the risk of liver cirrhosis constructed by using the feature variables obtained through multiple screening by machine learning has strong predictive ability and provides an effective tool for the early screening of the risk of liver cirrhosis.

Keywords: Liver cirrhosis; Biomarkers; Machine learning; Predictive model

肝硬化是全球第 11 大常见的死亡病因,每年约 200 万人死于该疾病[1]。近年来,肝硬化所带来的疾病负担逐渐加重,成为全球范围内日益突出的公共卫生问题[2-3]。肝硬化病因复杂多样,许多患者同时存在多种致病因素[4]。另外,肝硬化的发生及发展通常需要数年时间,并且由于慢性肝病患者间的异质性,预测肝硬化的进展及其并发症目前具有一定挑战性[5]。相关研究表明,多种肝脏相关生物标志物与肝硬化发生风险之间密切相关。也有研究表明肝硬化发生还与一些非肝脏生物标志物(如代谢物和炎症因子)等存在一定关联[6]。原因可能与许多生物标志物的循环水平会受到昼夜节律的影响,并且其容易受到如饮酒等多种因素的干扰[7]。因此,基于多种生物标志物来评估肝硬化发病风险具有重要临床意义。

目前有关肝硬化的研究主要集中在肝硬化的诊断上,而肝硬化发病风险预测的研究相对较少。迄今为止,已有几种复合生物标志物和评分系统被开发用于肝硬化的预测和早期检测。例如,FIB - 4 评分(Fibrosis - 4 Index)、APRI 评分(Aspartate Aminotransferase to Platelet Ratio Index)。然而,这些临床风险模型最初是为慢性肝病患者设计的,而对于一般人群而言,其模型的敏感性和特异性较差[8]。鉴于此,本研究拟使用英国生物样本库(UK Biobank)的相关数据,采用 XGBoost 算法、Lasso - Cox 回归和随机生存森林(Random Survival Forests, RSF)三种机器学习方法,筛选影响肝硬化发生的关键生物标志物。并基于这些生物标志物开发并验证适用于一般人群的肝硬化发病风险预测模型。该研究为肝硬化发病风险预测提供了科学依据。

1　材料与方法

1.1　*研究对象*　本研究从英国生物银行(UK Biobank)数据库共导出 502 366 人的相关数据。参与者于 2006 年至 2010 年间入组,随访至 2022 年 10 月 31 日。主要结果是新发肝硬化病例。按照以下排除标准筛选研究对象:(1)排除基线通过国际疾病分类(ICD - 10)代码确定的患任何类型的肝病(酒精性肝病、中毒性肝病、未分类的慢性肝炎、未分类的肝衰竭疾病)的研究对象共 1 101 人。(2)排除缺乏血液生化指标的研究对象共 263 083 人。(3)排除肝硬化发病日期早于入组开始随访时间的研究对象共 217 人。将符合上述标准的 237 965 名研究对象纳入研究,其中肝硬化患者 1 064 人,未患肝硬化者 236 801 人。

1.2　*主要研究变量*　临床生物标志物:UK Biobank 开展了一个项目,测量几乎所有参与者基线时收集的生物样本中的血液生化指标[9]。共有 30 种生物标志物可用,其中 28 种在超过 75% 的参与者中进行了测量,将这 28 种血液生化指标纳入研究。排除了缺乏血液生化指标的所有研究对象。

协变量:采用了年龄、性别、体质量指数、吸烟状况、饮酒频率和身体活动水平。吸烟状况分为三组:从不吸烟者、以前吸烟者和现在吸烟者。饮酒频率分

为三组:从不或仅在特殊场合饮酒,≤2 次/周,≥3 次/周。使用每周持续≥10 分钟的适度体育活动天数作为替代指标来测量身体活动水平,分为三组:0 ~ 1 d/w、2 ~ 4 d/w 和 5 ~ 7 d/w。这些协变量的缺失比例均 <0.5%,对于协变量的缺失值,采用随机森林法进行插补。该插补策略保持了数据的原始分布特征,确保后续分析的可靠性。

1.3 XGBoost 算法 XGBoost 算法将决策树作为基学习器,通过集成算法,在模型训练过程中利用所有中央处理器核心并行计算特征分裂的增益。输入自变量后,计算每棵树的预测值,并对上一棵树的预测残差进行泰勒展开式的一阶和二阶导数,以实现模型损失函数的快速优化[10]。XGBoost 算法在变量筛选方面,采用了变量重要性法,能够通过评估各个特征对模型输出的影响来识别和选择重要的特征。常见的评估指标包括增益重要性(Gain)、Gini 指数、Shannon 熵等,能够有效地衡量每个特征在模型训练过程中的贡献。

1.4 Lasso – Cox 回归模型 Lasso – Cox 回归模型是一种基于 Cox 回归模型的半参数、正则化方法,广泛应用于生存分析领域。该模型通过结合 Lasso 回归的 L1 正则化与 Cox 比例风险模型的优势,能够在高维数据中有效地进行变量选择和系数估计。通过最小化 Cox 模型的似然函数并加入 L1 正则化项,以此对生存数据进行建模。在该过程中,Lasso 正则化项控制回归系数的大小,使得部分系数趋于零,从而在高维数据下选择出最具预测意义的变量。

1.5 随机生存森林模型 随机生存森林模型(RSF)是一种基于生存树的非参数、非线性的集成机器学习方法[11],是对随机森林在右删失生存数据上的扩展。该模型通过应用自助采样和随机节点分裂的方法,生长独立的二元生存树,并将所有树集成形成 RSF 模型。RSF 模型通过变量筛选,可以了解各个变量在建模过程中的作用大小。常用的变量筛选方法包括变量重要性(Variable Importance, VIMP)法和最小深度(minimal depth, MD)法。此外,RSF 能够自动拟合交互作用并控制过度拟合[12],有助于识别疾病预后的共同影响因素。

1.6 统计学分析 本研究所有统计分析均使用 R 4.1.1 软件进行。使用 mice 包(v 3.16.0)进行缺失值插补;随机生存森林、Lasso – Cox 回归和 XGBoost 模型分别采用 randomForestSRC 包、glmnet 包和 xgboost 包构建。使用 nricens(v 1.6)包进行预测模型比较。本研究采用双侧检验水准 $\alpha = 0.05$。

2 结 果

2.1 研究对象的基线特征 本研究共纳入 237 965 名研究对象,其中肝硬化患者 1 064 名,未患肝硬化者 236 801 名。肝硬化患者组研究对象的平均年龄为(58.7 ± 7.5)岁;男性 706 人(66.4%),女性 358 人(33.6%)。未患肝硬化组研究对象的平均年龄为(56.2 ± 8.3)岁;男性 129 861 人(54.8%),女性 107 040人(45.2%)。除脂蛋白 A 外,所有协变量及生物标志物在肝硬化患者组和未患肝硬化组之间具有差异性($P < 0.05$)。见表 1。

表 1 纳入研究对象的基线特征[$\bar{x} \pm s$,n(%)]

Table 1 Baseline characteristics of included subjects[$\bar{x} \pm s$,n(%)]

变量	肝硬化患者(n=1 064)	未患肝硬化者(n=236 801)	t/χ^2 值	P 值
年龄(Age,岁)	58.7 ± 7.5	56.2 ± 8.3	−10.359	<0.001
性别(SEX)			56.933	<0.001
男	706(54.8)	129 861(66.4)		
女	358(45.2)	107 040(33.6)		
体质指数(BMI)	30.7 ± 6.1	27.2 ± 4.6	−22.697	<0.001
饮酒频率(Alcohol intake frequency, AIF)			283.359	0.004
从不或仅限特殊场合	205(29.0)	487(17.2)		
≤2 次/周	233(33.0)	1 059(37.4)		
≥3 次/周	269(38.0)	1 280(45.3)		
不愿回答	0(0.0)	2(0.1)		
吸烟状况(Smoking status, SS)			347.524	<0.001
从不	288(40.7)	1 559(55.1)		
以前	303(42.9)	993(35.1)		
当前	109(15.4)	265(9.4)		
不愿回答	7(1.0)	11(0.4)		
身体活动水平(Physical activity level, PAL, d/w)			291.798	0.001
0 ~ 1	166(23.5)	604(21.4)		
2 ~ 4	233(33.0)	1 049(37.1)		

（续表）

变量	肝硬化患者（n = 1 064）	未患肝硬化者（n = 236 801）	t/χ^2 值	P 值
5 ~ 7	247(34.9)	1 027(36.3)		
不知道	51(7.2)	138(4.9)		
不愿回答	10(1.4)	10(0.4)		
血液生化指标				
天冬氨酸氨基转移酶(AST,U/L)	42.7 ± 28.9	26.0 ± 8.2	−66.768	<0.001
丙氨酸氨基转移酶(ALT,U/L))	43.5 ± 36.4	23.5 ± 13.7	−48.167	<0.001
白蛋白(ALB,g/L)	44.4 ± 2.9	45.4 ± 2.6	13.190	<0.001
总胆红素(TBIL,μmol/L)	10.3 ± 5.1	10.0 ± 4.5	−7.042	0.003
肌酐(Cr,umol/L)	75.5 ± 20.1	73.8 ± 14.9	−2.599	0.009
尿素(Urea,mmol/L)	5.6 ± 1.8	5.4 ± 1.4	−1.943	<0.005
碱性磷酸酶(ALP,U/L)	105.8 ± 65.5	81.5 ± 23.8	−33.081	<0.001
γ − 谷氨酰转移酶(GGT,U/L)	113.3 ± 142.7	36.0 ± 32.9	−84.245	<0.001
脂蛋白 A(Lp a,g/L)	42.5 ± 48.4	45.0 ± 49.7	1.488	0.306
直接胆红素(DBil,umol/L)	2.2 ± 1.2	1.9 ± 0.9	−20.554	<0.001
睾酮(Testosterone,nmol/L)	8.2 ± 6.5	7.3 ± 6.2	−9.595	<0.001
性激素结合球蛋白(SHBG,nmol/L)	53.7 ± 29.6	50.2 ± 27.1	−7.446	0.002
载脂蛋白 A(Apo A,g/L)	1.4 ± 0.3	1.5 ± 0.3	7.194	<0.001
磷酸盐(Phosphate,mmol/L)	1.1 ± 0.2	1.1 ± 0.2	4.598	<0.001
葡萄糖(GLU,mmol/L)	5.7 ± 2.0	5.1 ± 1.2	−18.722	<0.001
总蛋白(TP,g/L)	74.0 ± 4.8	72.6 ± 4.1	−11.829	<0.001
钙(Ca,mmol/L)	2.4 ± 0.1	2.4 ± 0.1	2.976	0.050
维生素 D(VD,nmol/L)	44.3 ± 21.6	48.5 ± 21.0	9.200	<0.001
糖化血红蛋白(HbA1c,mmol/mol)	39.9 ± 10.4	35.8 ± 6.7	−20.843	<0.001
胰岛素生长因子 1(IGF − 1,nmol/L)	17.3 ± 6.2	21.6 ± 5.6	29.463	<0.001
载脂蛋白 B(Apo B,g/L)	0.9 ± 0.2	1.0 ± 0.2	9.574	<0.001
C 反应蛋白(CRP,g/L)	4.5 ± 6.4	2.5 ± 4.1	−16.422	<0.001
直接低密度脂蛋白(LDL D,g/L)	3.1 ± 0.8	3.5 ± 0.8	13.963	0.030
胱抑素 C(CysC,mg/L)	1.0 ± 0.2	0.9 ± 0.2	−25.363	<0.001
胆固醇(CHOL,mmol/L)	5.1 ± 1.1	5.5 ± 1.0	15.067	<0.001
甘油三酯(TG,mmol/L)	1.9 ± 1.1	1.7 ± 0.9	−9.182	<0.001
尿酸盐(Urate,umol/L)	353.9 ± 92.1	314.8 ± 81.3	−15.602	<0.001
高密度脂蛋白胆固醇(HDL,mmol/L)	1.3 ± 0.4	1.4 ± 0.4	10.798	<0.001

2.2　基于 XGBoost 的肝硬化特征变量筛选结果　本研究基于训练集，利用 XGBoost 算法筛选肝硬化关键预测变量，并将变量按照重要性进行排序，通过 Gain 确定变量重要性，研究结果见图 1。在评估不同变量组合模型的性能时，本研究根据计算得到的准确度（Accuracy）、AUC 、F1 值（F1 − score）进行评估（见表 2），决定使用前 10 个变量来平衡简单性和准确性。本研究基于 XGBoost 算法筛选的重要性排序前 10 个的特征变量变量分别为 GGT、IGF − 1、AST、SHBG、CysC、GLU 、ALB、BMI、Testosterone、TP。

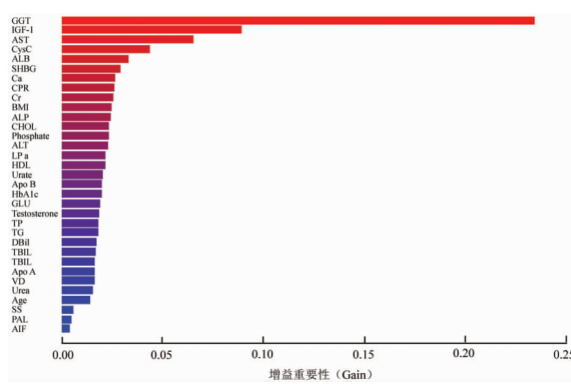


图 1　基于 XGBoost 的变量重要性排序图

Fig. 1　Variable importance ranking based on XGBoost

表 2　评估不同变量组合模型性能的参数

Table 2　Parameters to evaluate the performance of models with different combinations of variables

前 n 个变量	Accuracy	AUC	F1 值
5	0.995 557 6	0.987 283 3	0.997 773 7
10	0.995 731 7	0.997 730 6	0.997 860 7
15	0.995 779 7	0.999 156 0	0.997 884 8
20	0.995 875 7	0.999 848 8	0.997 932 8

2.3　基于 Lasso − Cox 回归的肝硬化特征变量筛选结果　本研究基于训练集，采用 Lasso − Cox 回归筛选肝硬化潜在预测变量，将所有的变量通过 10 折交叉验证选择最优参数 λ（λ = 0.004 327 759）进行模型构建（见图 2）。筛选出的前 28 个变量根据系数从大到小排列，具体情况见表 3。

表 3 Lasso – Cox 回归筛选的变量

Table 3 Lasso – Cox regression screening variables

编号	变量	系数
1	CysC	1.593 524 578
2	Apo B	0.821 664 110
3	Phosphate	0.278 924 995
4	SS	0.243 387 559
5	DBil	0.118 589 562
6	AIF	0.100 298 476
7	TP	0.065 999 561
8	BMI	0.057 121 474
9	Testosterone	0.050 807 354
10	TG	0.043 750 489
11	Urea	0.029 720 335
12	Age	0.011 868 706
13	AST	0.009 832 830
14	CRP	0.009 085 691
15	ALT	0.004 188 530
16	GGT	0.003 815 536
17	Urate	0.001 606 998
18	HbA1c	0.001 246 085
19	ALP	–0.001 017 578
20	VD	–0.003 912 683
21	HDL	–0.008 389 843
22	Cr	–0.013 432 115
23	ALB	–0.045 594 826
24	PAI)	–0.066 784 097
25	IGF – 1	–0.100 089 142
26	CHOL	–0.269 037 675
27	LDL D	–0.323 063 609
28	Ca	–0.624 568 177

2.4 基于随机生存森林的肝硬化特征变量筛选结果 本研究基于训练集，采用随机生存森林方法，分别根据变量重要性法和最小深度法将变量进行重要性排序，研究结果见图 3。由图 3A 可知，基于 VIMP 值筛选的重要性排名前 5 的变量分别为 GGT、AST、IGF – 1、CysC、ALT；由图 3B 可知，基于最小深度法筛选的重要性排名前 5 的变量分别为 AST、GGT、ALP、IGF – 1、CysC。

图 4 为综合两种方法的散点图。其中，高于红色对角虚线的点代表其 VIMP 值排名更高，低于对角虚线的点则代表其最小深度法排名更高，位在红色对角虚线上的点代表两种方法对该变量的排名相同。

此外，本研究评估了不同变量组合模型的 Harrell C 指数。C 指数从前 5 个变量的 0.914 增加到前 10 个变量的 0.923，然后略微增加到前 15 个变量的 0.929。当包括所有 34 个变量时，C 指数达到 0.933。因此，决定使用前 15 个变量来平衡简单性和准确性。重要性排序前 15 个的特征变量分别为 AST、GGT、ALP、IGF – 1、CysC、ALT、CHOL、DBil、ALB、CRP、BMI、HbA1c、LDL D、GLU、SHBG。

2.5 三种机器学习方法共同筛选变量的结果 本研究基于训练集，联合三种的机器学习算法筛选肝硬化发病风险潜在预测变量。其中随机生存森林确定了 15 个潜在的预测变量；Lasso – Cox 回归确定了 28 个

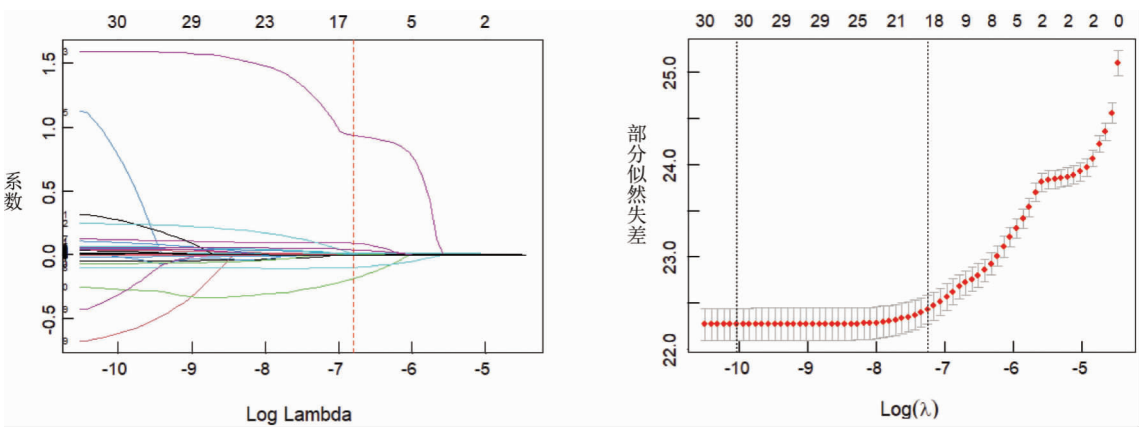


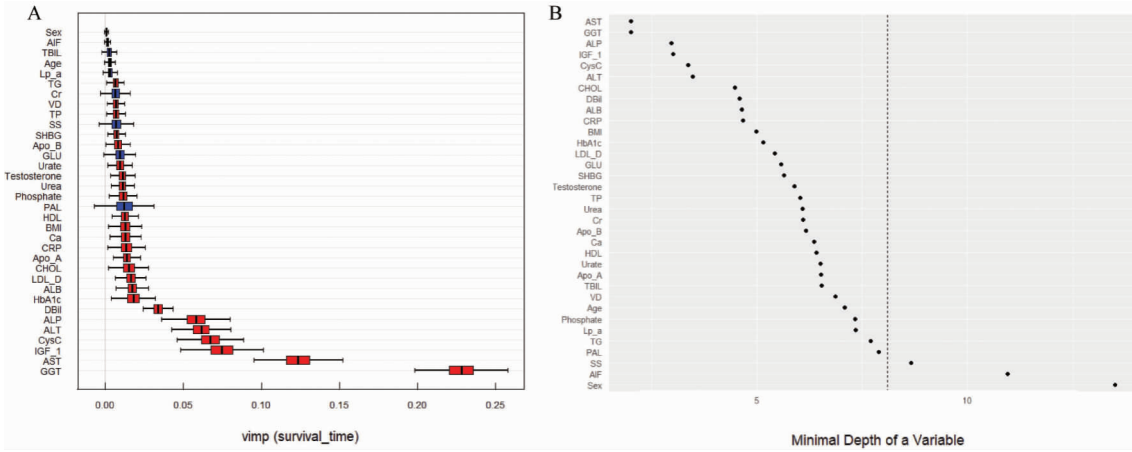
图 2 基于 Lasso – Cox 回归的变量筛选及交叉验证构建模型

Fig. 2 The model was constructed based on Lasso – Cox regression variable screening and cross – validation

潜在的预测变量；XGBoost 算法确定了 10 个潜在预测变量，具体结果见图 5。由图 5 可知，三种方法共同确定了 6 个特征变量，分别是 GGT、AST、IGF – 1、CysC、BMI、ALB。

2.6 肝硬化发病风险预测模型构建 考虑到 IGF – 1 在日常临床实践中可能不会常规测量，故本研究选择其他 5 个特征变量构建 Cox 比例风险模型。Cox 模型结果显示，ALB 是肝硬化发生的保护因素[HR = 0.91，95% CI：0.88 ~ 0.93，P < 0.005]；而 GGT、AST、CysC 和体质量指数 BMI 是肝硬化的独立危险因素（P < 0.05），其中 CysC 危害最大[HR = 2.51，95% CI：2.22 ~ 2.83，P < 0.005]，见图 6。同时，本研究建立

了用于肝硬化发病风险预测列线图,见图 7。由图 7 可知,列线图中总分是相关预测变量对应的分数之和,根据个体风险总分可预测该个体发生肝硬化的概率大小。



注:图 A 以 VIMP 值为依据,图 B 最小深度为依据。

图 3　基于随机生存森林的变量重要性排序

Fig. 3　Variable importance ranking based on random survival forest

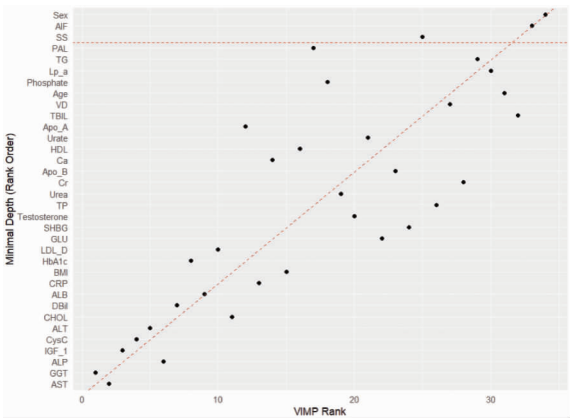


图 4　随机生存森林模型 VIMP 法和最小深度法的散点图

Fig. 4　Scatter plot of random survival forest model VIMP method and minimum depth method

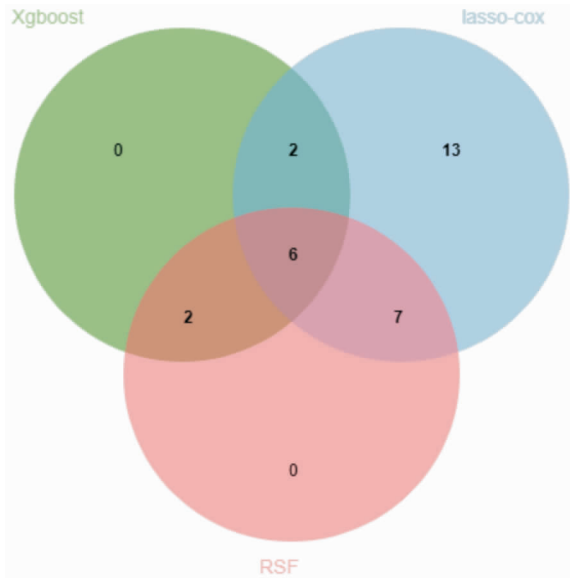


图 5　基于机器学习结果得到的特征变量

Fig. 5　Potential predictors were obtained based on machine learning results

2.7　肝硬化发病风险预测模型评价

2.7.1　各模型的预测性能比较　为了评价各肝硬化发病风险预测模型的预测性能,本研究采用 ROC 曲线下面积大小,分别比较了本研究构建的预测模型、FIB-4 评分模型和 APRI 评分模型的预测性能大小。研究结果显示,在训练集中,三个模型的 AUC 分别为 0.827(0.806~0.848)、0.665(0.637~0.693)、0.740(0.713~0.766);在验证集中,三个模型的 AUC 分别 0.811(0.776~0.845)、0.664(0.621~0.708)和 0.711(0.668~0.755),结果见图 8。本研究构建的预测模型,在训练集和验证集中的预测性能均优于传统评分的预测模型,模型之间的效能差异均有统计学意义(Delong 检验 $P<0.05$)。

2.7.2　各模型的临床决策诊断性能比较　为了评价模型的临床决策诊断性能,本研究分别采用净重新分类指数(NIR)和综合判别改善指数(IDI)对本研究构建的肝硬化发病风险预测模型与传统 FIB-4 评分模型、APRI 评分模型进行了对比分析。研究结果显示,本研究构建的肝硬化发病风险预测模型,其净 NRI 和 IDI 均表现出显著优势($P<0.001$),且在训练集和测试集中的结果一致性较高,具体结果见表 4。

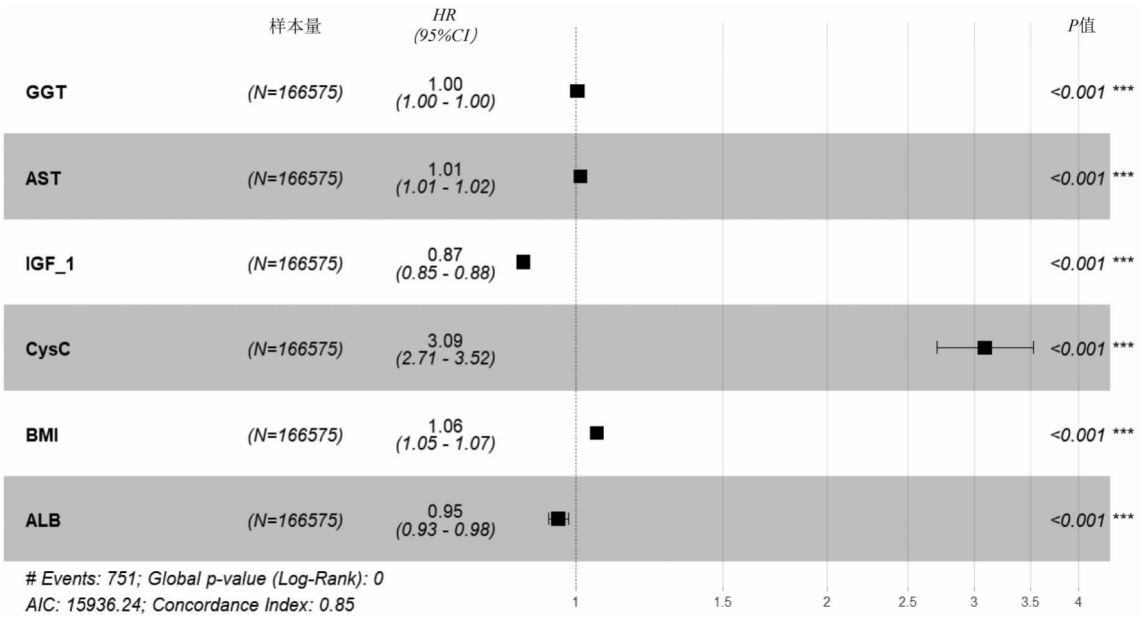


图 6 多因素 Cox 比例风险回归

Fig. 6 Multivariate Cox proportional risk regression

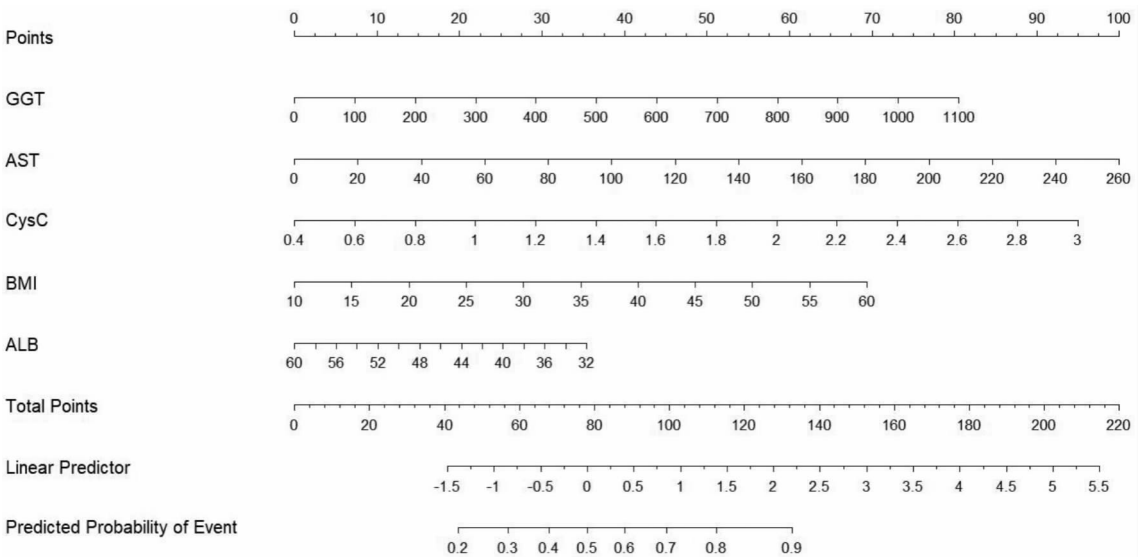


图 7 预测肝硬化发病风险的列线图

Fig. 7 Nomogram for predicting the risk of liver cirrhosis development

表 4 两模型净重新分类指数(NRI)与综合判别改善指数(IDI)比较

Table 4 Comparison of the net reclassification index(NRI) of the model with the integrated discrimination improvement index(IDI)

数据集	模型比较	NIR(95% *CI*)	P_{NIR}	IDI(95% *CI*)	P_{IDI}
训练集	本研究模型与 FIB－4 评分模型	0.137(0.072～0.229)	<0.001	0.111(0.095～0.127)	<0.001
	本研究模型与 APRI 评分模型	0.138(0.050～0.232)	<0.001	0.111(0.096～0.127)	<0.001
测试集	本研究模型与 FIB－4 评分模型	0.150(0.049～0.248)	<0.001	0.097(0.063～0.131)	<0.001
	本研究模型与 APRI 评分模型	0.146(0.023～0.240)	<0.001	0.085(0.063～0.130)	<0.001

3 讨 论

肝硬化是全球常见的死亡病因之一,预计未来 10 年肝硬化相关死亡人数将增加[13]。因此,积极开发肝硬化早期发病风险预测模型具有重要临床意义[3]。本研究分别采用 XGBoost 算法、Lasso－Cox 回

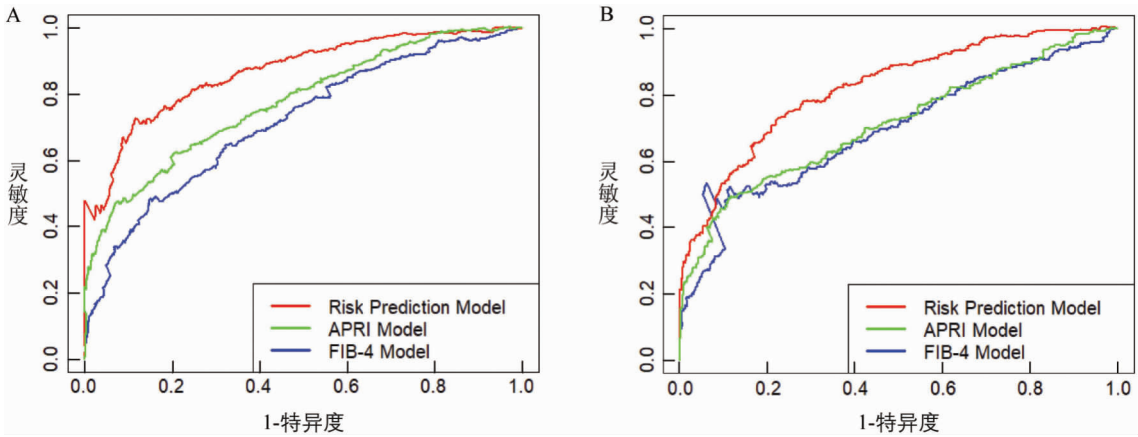


图 8　训练集(A)和测试集(B)APRI、FIB－4 和本研究模型的受试者工作曲线(ROC)下面积比较

Fig. 8　Comparison of the area under the receiver operating characteristic curve(ROC) for APRI, FIB－4, and the proposed model in the training(A) and test(B) sets

归和随机生存森林三种机器学习算法,并结合临床应用实际,共筛选的 5 个关键特征变量,其中包含 BMI 和 4 个生物标志物,分别是 GGT、AST、CysC、ALB。并基于筛选的 5 个关键特征变量构建了肝硬化发病风险预测模型。研究结果显示,GGT、AST、CysC 和 BMI 是肝硬化发生的独立危险因素,ALB 是其独立保护因素。该研究结果与其他研究结果相似。

谷丙转氨酶(GGT)已被广泛研究并证实是肝脏疾病的重要生物标志物。研究表明,GGT 水平的升高常常提示着肝脏功能的损害和肝纤维化的进展[14－15]。AST 水平的升高与肝硬化的进展密切相关。相关研究表明,AST 不仅反映了肝细胞的损伤,还与肝脏纤维化的程度成正比[16]。导致 AST 升高的机制是肝纤维化的进展可能会损害 AST 清除,从而导致其在血液中保留[17]。因此,AST 水平的提高为预测肝硬化的严重程度和进展提供了重要线索。

最新研究表明,胱抑素 C(CysC)作为一个新兴的生物标志物[18],在纤维化疾病的预测中具有较高的特异性和价值[19],CysC 水平的升高与肝纤维化进展密切相关[20－21]。这与本研究的结果高度一致。另外,本研究结果表明,BMI 是肝硬化的潜在危险因素之一,其原因可能与肥胖引起的代谢紊乱及其对肝脏的负面影响有关。有研究表明,肥胖是西方国家人口研究中肝硬化发病的独立危险因素[22－23]。

综上所述,本研究采用多种生物标志物及关键指标构建的肝硬化发病风险预测模型可为一般人群肝硬化早期筛查提供实践依据。然而,本研究也存在一定局限性。首先,本研究主要基于横断面数据开展的研究,因此该研究无法确定变量之间的因果时序关系,可能导致关联估计具有一定偏差。其次,本研究所用数据库为 UK Biobank 数据,该数据库主要为英国人群,因此本研究所得结论在其他人群中是否成立,需要进一步通过外部人群验证。最后,尽管本研究人群样本量较大,但是随访期间肝硬化病例数量有限,可能会影响研究结果的统计学效能和外推性。

利益冲突声明　本研究不存在任何利益冲突

参考文献

[1] Xu XY, Ding HG, Li WG, et al. Chinese guidelines on the management of liver cirrhosis (abbreviated version)[J]. World Journal of Gastroenterology, 2020, 26(45): 7088－7103.

[2] Huang DQ, Mathurin P, Cortez－Pinto H, et al. Global epidemiology of alcohol－associated cirrhosis and HCC: trends, projections and risk factors[J]. Nature Reviews Gastroenterology & Hepatology, 2023, 20(1): 37－49.

[3] Huang DQ, Terrault NA, Tacke F, et al. Global epidemiology of cirrhosis — aetiology, trends and predictions[J]. Nature Reviews Gastroenterology & Hepatology, 2023, 20(6): 388－398.

[4] Gastroenterology O, Association MC. Chinese consensus on the management of liver cirrhosis[J]. Journal of Digestive Diseases, 2024, 25(6): 332－352.

[5] Patel K, Sebastiani G. Limitations of non－invasive tests for assessment of liver fibrosis[J]. JHEP REPORTS, 2020, 2(2): 100067.

[6] Song J, Gao Z, Lai L, et al. Machine learning－based plasma metabolomics for improved cirrhosis risk stratification[J]. BMC Gastroenterology, 2025, 25(1): 61.

[7] Graindorge PH, Paoli J, Yildirim B, et al. Early stage of metabolic dysfunction associated steatotic liver disease disrupts circadian rhythm and induces neuroinflammation in rats[J]. Scientific Reports, 2025, 15(1): 10616.

[8] Åberg F, Asteljoki J, Männistö V, et al. Combined use of the CLivD score and FIB－4 for prediction of liver－related outcomes in the population[J]. Hepatology, 2024, 80(1): 163－172.

[9] Kunutsor SK, Apekey TA, Seddoh D, et al. Liver enzymes and risk of all－cause mortality in general populations: a systematic review and meta－analysis[J]. International Journal of Epidemiology, 2014, 43(1): 187－201.

(下转第 4131 页)

参考文献

[1] 健康中国研究网络专家组,梁万年. 创新医防融合 共筑健康中国——2023 年健康中国研究网络专家共识[J]. 中国全科医学,2024,27(22):2685 - 2688.
Jian Kang Zhong Guo Yan Jiu Wang Lao Zhuan Jia Zu, Liang WN. Innovative integration of treatment and prevention to build a healthy China together: expert consensus from the healthy China research network in 2023[J]. Chinese General Practice, 2024, 27(22): 2685 - 2688. (In Chinese)

[2] 汪辉,孙磊. 构筑职业防线守护劳动者健康[N]. 郑州日报, 2024 - 10 - 16(005).
Wang H, Sun L. Building occupational defenses to protect workers' health[N]. Zhengzhou Daily, 2024 - 10 - 16(005). (In Chinese)

[3] 潘杰,张田丰,张雨萌,等. 数智驱动医疗卫生服务体系高质量发展:发展机制与实施路径[J]. 四川大学学报(医学版), 2024,55(5):1055 - 1062.
PAN J, Zhang TF, Zhang YM, et al. Digital intelligence drives the high - quality development of the healthcare service system: Development mechanisms and implementation pathway[J]. Journal of Sichuan University(Medical Sciences), 2024, 55(5): 1055 - 1062. (In Chinese)

[4] Kilpeläinen K, Ståhl T, Ylöstalo T, et al. Citizens' digital footprints to support health promotion at the local level - PUHTI study, Finland[J]. European Journal of Public Health, 2024, 34(4): 676 - 681.

[5] Damilos S, Saliakas S, Karasavvas D, et al. An overview of tools and challenges for safety evaluation and exposure assessment in industry 4.0[J]. Applied Sciences, 2024, 14(10): 4207.

[6] Jacobs JV, Hettinger LJ, Huang YH, et al. Employee acceptance of wearable technology in the workplace[J]. Applied Ergonomics, 2019, 78: 148 - 156.

[7] Rama RAJ, Madrigal PL, Mustelier CLJ, et al. Monitoring and ensuring worker health in controlled environments using economical particle sensors[J]. Sensors, 2024, 24(16): 5267.

[8] Bernier T, Shah A, Ross LE, et al. The needs and preferences of Eastern Canadian sex workers in mitigating occupational health and safety risks through the use of Information and Communication Technologies: A qualitative study[J]. PLOS One, 2022, 17(6): e0269730.

[9] Beckman S, Andamel N, Harrison R. California warehouse industry worker injury rates in the occupational health and safety administration's injury tracking application, 2018 - 2023[J]. Am J Ind Med, 2025, 68(3): 295 - 303.

[10] Felsen M. Addressing worker safety and health through the lens of strategic Enforcement - Part 2[J]. New Solutions: A Journal of Environmental and Occupational Health Policy, 2024, 34(3): 227 - 240.

[11] Koppisch D, Zschiesche W, Goebel A, et al. Schwei brauchex positionen in deutschland und bewertung der gesundheitsschädigenden wirkunge - insbesondere im hinblick auf lungenkrebserkrankungen[J]. Arbeitsmed Sozialmed Umweltmed, 2023, 58(3): 185 - 195.

[12] Siefer A. Berichterstattung zu sicherheit und gesundheit bei der arbeit[Reporting on safety and health at work][J]. Bundes gesundheit sblatt, Gesundheits for schung, Gesundheits schutz, 2020, 63(9): 1076 - 1083.

收稿日期:2025-07-01

(上接第 4045 页)

[10] 陈庄,周余. 基于 MIC - XGBoost 算法的居民用水量数据预测[J]. 计算机应用与软件,2021,38(10):125 - 130.
Chen Z, Zhou D. Prediction of residential water consumption data based on MIC - XGBoost algorithm[J]. Computer Applications and Software, 2021, 38(10): 125 - 130. (In Chinese)

[11] Yosefian I, Farkhani EM, Baneshi MR. Application of random forest survival models to increase generalizability of decision trees: a case study in acute myocardial infarction[J]. Computational and Mathematical Methods in Medicine, 2015, 2015: 576413.

[12] Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests[J]. Annals of Applied Statistics, 2008, 2(3): 841 - 860.

[13] Chen QJI, Li FW, Gao YZ, et al. Identification of energy metabolism genes for the prediction of survival in hepatocellular carcinoma[J]. Frontiers in Oncology, 2020, 10: 1210.

[14] Kwak J, Seo IH, Lee YJ. Serum γ - glutamyltransferase level and incidence risk of metabolic syndrome in community dwelling adults: longitudinal findings over 12 years[J]. Diabetology & Metabolic Syndrome, 2023, 15(1): 29.

[15] Xuan YY, He FF, Liu Q, et al. Elevated GGT to HDL ratio as a marker for the risk of NAFLD and liver fibrosis[J]. Scientific Reports, 2025, 15(1): 10.

[16] Dao TD, Tran HV, Nguyen KH, et al. Prevalence and serological predictors of liver fibrosis in vietnamese chronic hepatitis B patients with persistently normal or mildly elevated aminotransferase level[J]. Med Arch, 2025, 79(2): 147 - 154.

[17] Okuda M, Li K, Beard MR, et al. Mitochondrial injury, oxidative stress, and antioxidant gene expression are induced by hepatitis C virus core protein[J]. Gastroenterology, 2002, 122(2): 366 - 375.

[18] Wang XY, Xu W, Yao L, et al. Association of cystatin C level with all - cause mortality in patients with liver cirrhosis: a meta - analysis[J]. Current Medicinal Chemistry, 2024, 31(25): 3977 - 3986.

[19] Lalmanach G, Rigoux B, David A, et al. Human cystatin C in fibrotic diseases[J]. Clinica Chimica Acta, 2025, 565: 120016.

[20] Nagy E, Abdelfattah AH, Sayed - Ahmed N, et al. Cystatin C as a predictor of renal recovery and in hospital mortality in patients with acute kidney injury and liver cirrhosis[J]. BMC Nephrology, 2025, 26(1): 413.

[21] Lytvyak E, Hirschfield G, Shreekumar D, et al. Pathogenesis, Non - Invasive assessments and treatment of hepatic fibrosis in autoimmune liver diseases[J]. Liver International, 2025, 45(8): e70190.

[22] Abeysekera KWM, Fernandes GS, Hammerton G, et al. Prevalence of steatosis and fibrosis in young adults in the UK: a population - based study[J]. Lancet Gastroenterol Hepatol, 2020, 5(3): 295 - 305.

[23] Glyn - Owen K, Böhning D, Parkes J, et al. The combined effect of alcohol and body mass index on risk of chronic liver disease: A systematic review and meta - analysis of cohort studies[J]. Liver International, 2021, 41(6): 1216 - 1226.

收稿日期:2025-06-28