·论著·一次研究·

基于 XGBoost 的他汀类药物用药者急性肝损伤预测模型的开发与验证

孟祥龙[1,2]，于玥琳[1,2]，孙烨祥[3]，沈　鹏[3]，江志琴[4]，朱　宇[4]，殷玥琪[3]，詹思延[1,2,5]，王胜锋[1,2,5]

1. 北京大学公共卫生学院流行病与卫生统计学系（北京 100191）
2. 教育部重大疾病流行病学重点实验室（北京 100191）
3. 宁波市鄞州区疾病预防控制中心数据中心（浙江宁波 315100）
4. 宁波市鄞州区卫生健康局（浙江宁波 315100）
5. 北京大学人工智能研究院（北京 100191）

【摘要】目的 开发并验证预测模型以识别他汀类药物新用药者在 180 d 内发生急性肝损伤（ALI）的高风险个体，为临床早期干预提供依据。**方法** 选取 2010 年 1 月 1 日—2021 年 10 月 31 日在宁波市鄞州区域健康信息平台中具有诊疗记录的 18 岁及以上他汀类药物新用药者，按用药时间分为开发队列和时序验证队列。采用 LASSO 方法筛选预测变量，利用极限梯度提升（XGBoost）算法并结合代价敏感学习构建模型。通过 Brier 分数、Harrell's C 指数和校准曲线评估模型性能。**结果** 共纳入 126 440 例他汀类药物新用药者，其中开发队列 90 542 例，验证队列 35 898 例。首次用药后 180 d 内，共 412 例（0.33%）发生 ALI，其中开发队列 305 例（0.34%），验证队列 107 例（0.30%）。最终模型纳入 16 个预测变量，涵盖人口学特征、生活方式、家族史、既往史、他汀类药物用药情况及合并用药情况。模型在内部验证中表现出良好的整体性能 [Brier 分数 =0.004 3，95%CI（0.003 8，0.004 9）]、区分度 [Harrell's C 指数 =0.761，95%CI（0.725，0.794）] 和校准度。时序验证中，模型的性能同样良好 [Brier 分数 =0.004 4，95%CI（0.003 6，0.005 2）；Harrell's C 指数 =0.703，95%CI（0.614，0.781）]。**结论** 本研究开发并验证的他汀类药物新用药者 ALI 预测模型，可为临床医生提供可靠的个体化风险评估工具，有助于实现风险分层并减少 ALI 的发生。

【关键词】他汀类药物；急性肝损伤；预测模型；极限梯度提升算法；代价敏感学习

【中图分类号】R 972+.6　　**【文献标识码】**A

Development and validation of an XGBoost-based prediction model for acute liver injury in statin users

MENG Xianglong[1,2], YU Yuelin[1,2], SUN Yexiang[3], SHEN Peng[3], JIANG Zhiqin[4], ZHU Yu[4], YIN Yueqi[3], ZHAN Siyan[1,2,5], WANG Shengfeng[1,2,5]

1. Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing 100191, China

2. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China

3. Department of Data Center, Yinzhou District Center for Disease Control and Prevention of Ningbo,

DOI: 10.12173/j.issn.1005-0698.202502023

基金项目：国家自然科学基金面上项目（82173616）

通信作者：王胜锋，博士，研究员，博士研究生导师，Email：shengfeng1984@126.com

Ningbo 315100, Zhejiang Province, China
4. Yinzhou District Health Bureau of Ningbo, Ningbo 315100, Zhejiang Province, China
5. Institude for Artificial Intelligence, Peking University, Beijing 100191, China
Corresponding author: WANG Shengfeng, Email: shengfeng1984@126.com

【Abstract】Objective To develop and validate a prediction model to identify high-risk individuals who are at-risk to develop acute liver injury (ALI) within 180 days in new statin users, and to support early clinical intervention. Methods Data were sourced from the Yinzhou Regional Health Information Platform, covering statin initiators aged 18 years and older from January 1, 2010, to October 31, 2021. The dataset was divided into a derivation cohort and a temporal validation cohort based on the time of statin initiation. Predictors were selected using LASSO regression, and the model was constructed using the extreme gradient boosting (XGBoost) algorithm combined with cost-sensitive learning. Model performance was evaluated using Brier scores, Harrell's C-index, and calibration curves. Results A total of 126,440 statin initiators were included, with 90,542 in the derivation cohort and 35,898 in the validation cohort. Within 180 days of initial statin use, 412 (0.33%) patients developed ALI, including 305 (0.34%) in the derivation cohort and 107 (0.30%) in the validation cohort. The final model incorporated 16 predictors, which included demographic characteristics, lifestyle factors, family history, medical history, statin use, and concomitant medication use. The model demonstrated excellent overall performance [Brier score=0.0043, 95%CI (0.0038, 0.0049)], discrimination [Harrell's C-index=0.761, 95%CI (0.725, 0.794)], and calibration in internal validation. In temporal validation, the model also performed well [Brier score=0.0044, 95%CI (0.0036, 0.0052), Harrell's C-index= 0.703, 95%CI (0.614, 0.781)]. Conclusion This study develope and validate a prediction model for ALI in statin users, providing clinicians with a reliable tool for individualized risk assessment. This model can help achieve risk stratification and reduce the occurrence of ALI.

【Keywords】Statins; Acute liver injury; Prediction model; XGBoost algorithm; Cost-sensitive learning

随着生活方式的改变和人口老龄化的加剧，高脂血症已成为全球范围内常见的代谢性疾病之一[1]。他汀类药物通过抑制肝脏中 3– 羟基 –3– 甲基戊二酰辅酶 A 还原酶活性以减少体内胆固醇的合成，是临床上高脂血症和心血管疾病治疗和预防的基础药物[2]。尽管他汀类药物在大多数患者中表现出良好的安全性和耐受性[3–4]，但仍有可能引起一系列不良反应，其中急性肝损伤（acute liver injury，ALI）是较为严重的一种[5–7]。

他汀类药物引起的 ALI 虽然在临床上相对罕见，但其严重性不容忽视，我国也规定他汀类药物用药者应定期监测肝生化指标，以早期发现和预防可能的肝损伤[8]。ALI 的发生通常与用药剂量、用药时间、个体差异等因素有关[9]。已有研究[10–11]表明，特定人群如老年人、有肝病史者、合并使用其他肝毒性药物者，更易发生他汀类药物诱发的 ALI。此外，某些遗传因素也可能增加患者的肝损伤风险，如特定的基因多态性已被证实与他汀类药物引起的肌病和肝损伤有关[12–13]。鉴于 ALI 的发生具有显著的异质性，开发一种能更精准和全面预测 ALI 的模型显得尤为重要。目前已有预测他汀类药物不良反应的模型[14]，但尚无专门针对 ALI 且被临床推荐使用的预测模型。本研究通过整合用药者的基本信息、用药情况和疾病史等临床常规指标，开发并验证一个预测模型，旨在帮助临床医师识别发生 ALI 风险较高的他汀类药物用药者，从而优化治疗方案，减少 ALI 的发生。

1 资料与方法

1.1 数据来源

数据来源于宁波市鄞州区区域健康信息平台，鄞州区是浙江省宁波市最大的市辖区，全区常住人口约 136 万，人口流动性较低[15]。该数据平台自 2005 年开启搭建，至 2020 年末，已有超 98% 的该区内常驻居民在平台内建档，累计涵盖当地超 122 万人口的多方面、纵向、完整的卫生信息，如基本人口学特征、疾病诊断随访、

化验检查、处方、死因登记等[16]。本研究涉及的药物信息和疾病信息来源于鄞州区内所有医疗机构的门诊和住院记录。本研究已获得北京大学医学部伦理委员会审查批准（编号：IRB00001052-22010），并豁免患者知情同意。

1.2 研究对象

选取 2010 年 1 月 1 日—2021 年 10 月 31 日在宁波市鄞州区域健康信息平台中具有诊疗记录的 18 岁及以上他汀类药物新用药者为研究对象，纳入标准：① 2010 年 1 月 1 日—2021 年 10 月 31 日在宁波市鄞州区域健康信息平台中有门诊处方记录或住院医嘱记录中存在他汀类药物记录的新用药者；新用药者定义为数据库中记录的首次他汀类药物处方日期之前的 180 d 内无目标药品处方记录[17]，且要求首次他汀类药物处方日期与数据库截止日期的时间差＞ 180 d，以保证足够长的随访时间；②年龄≥ 18 岁。排除标准：①重要变量缺失，如建档编号、建档状态等；②首次用药时年龄＜ 18 岁；③患有恶性肿瘤等严重疾病或他汀类药物禁忌证；④在首次处方日期前 7 d 内出现与目标结局相关的严重 ALI 诊断记录[18]。

研究对象从首次用药时开始随访，直至出现 ALI 结局、死亡或研究结束（2021 年 10 月 31 日），以三者中最早发生的时间为准。

1.3 结局

参考药源性 ALI 不良反应因果关系评价方法[19]、药物性肝损伤诊治指南[20]和既往研究[21-22]，ALI 定义为满足以下任一条：①丙氨酸转氨酶（alanine aminotransferase，ALT）水平超过 5 倍正常值上限（upper limit of normal，ULN）；②碱性磷酸酶（alkaline phosphatase，ALP）水平超过 2 倍 ULN，且伴随 γ- 谷氨酰转肽酶（gamma glutamyl transpeptidase，GGT）升高；③ ALT 水平超过 3 倍 ULN，且总胆红素（total bilirubin，TBIL）水平超过 2 倍 ULN。为排除其他非药源性因素引起的肝损伤，要求以上生化指标异常的检查记录时间均需与末次他汀类药物用药记录间隔在 180 d 内。满足以上条件的首次生化指标检查日期记为 ALI 的发生时点。

1.4 预测变量

参考既往文献[23-24]，本研究共纳入了 6 个维度的临床易于收集的变量，包括：人口学特征、生活方式、家族史、既往史和合并症、他汀类药物用药情况以及合并用药情况，共计 34 个变量。

1.5 统计学分析

采用 R 4.3.2 软件进行进行统计学分析。计量资料使用 Kolmogorov-Smirnov 检验进行正态性检验，符合正态分布的连续变量以 $\bar{x} \pm s$ 表示，2组间比较采用 t 检验；若不符合正态分布，则以 M（P_{25}，P_{75}）表示，2 组间比较采用 Wilcoxon 秩和检验；分类变量以 n（%）表示，2 组间比较采用 χ^2 检验或 Fisher 精确性检验，等级资料比较采用秩和检验。检验水准为 α=0.05，所有检验均采用双侧检验。

根据首次他汀类药物用药时点，以 7∶3 的比例将研究对象划分为开发队列（2010 年 1 月 1 日—2017 年 12 月 31 日）和时序验证队列（2018 年 1 月 1 日—2021 年 10 月 31 日）。缺失比例≥ 50% 的预测变量不再纳入之后的分析。对于缺失比例＜ 50% 的变量，采用基于链式方程的多重插补法处理缺失数据[25]，共生成 10 组缺失数据，采用 Rubin 规则合并研究结果[26]。预测变量的缺失率在开发队列和验证队列中基本一致。

在开发队列中，采用最小绝对收缩与选择算子（least absolute shrinkage and selection operator，LASSO）方法筛选变量[27]，通过十折交叉验证，选择使得 LASSO-Cox 模型误差在最小误差的一个标准误之内的最大值作为最优惩罚参数[28]。在开发队列中，以 ALI 结局的发生情况为分层变量，按照 8∶2 的比例随机划分为训练集和测试集。训练集用于极限梯度提升（extreme gradient boosting，XGBoost）模型开发，测试集用于评估模型的内部性能。XGBoost 是一个快速学习框架，使用梯度提升决策树以优化损失函数，能高效地处理大规模数据集并具有较高的预测精度[29]。由于 ALI 的发生率较低，存在严重的类别不平衡问题，因此在模型训练中结合了代价敏感学习方法，将阳性结局误分类的代价设置为超参数，后续进行调整优化[30]。通过网格搜索结合十折交叉验证进行模型优化。

由于 XGBoost 模型只能提供相对风险的估计，而无法直接得出个体的绝对风险，本研究进一步估计 ALI 的基线风险，并计算个体的绝对风险，具体步骤如下[31]：①基于 XGBoost 模型计算个体的相对风险评分；②以风险评分为预测变量，拟合单因素 Cox 比例风险模型，计算绝对风险，即 ALI 的发生概率。从整体表现、区分度

和校准度3个维度评价模型表现。使用Brier分数评价模型的整体表现；使用Harrell's C指数和时间依赖性受试者工作特征（receiver operating characteristic，ROC）曲线评价模型的区分度，其中，Harrell's C指数的取值范围为0~1，C指数＞0.7表明模型的区分度表现良好[32]；使用校准曲线评价模型的校准度。

为了全面评估模型的表现，进行亚组分析，评估模型在不同用药种类和不同剂量人群中的性能。研究依据透明报告多变量预测模型个体预后或诊断标准（transparent reporting of a multivariable prediction model for individual prognosis or diagnosis，TRIPOD）进行报告[33]。

2 结果

2.1 研究对象的基本特征

共发现126 440例他汀类药物新用药者，其中，开发队列90 542例，验证队列35 898例，总用药达1 718 452人次，研究人群筛选流程图及数据集拆分见图1。首次用药类型为阿托伐他汀、辛伐他汀和瑞舒伐他汀者最多，分别62 880例（49.7%）、38 623例（30.5%）和21 408例（16.9%）。用药剂量方面，单次用药剂量≤10 mg、＞10~20 mg和＞20 mg的用药者分别为41 672例（33.0%）、54 354例（43.0%）和30 414例（24.1%）。药品剂型以片剂为主（99.7%），给药途径则主要为口服（97.3%）。此外，用药记录大部分来自门诊处方（97.9%），见表1。

研究人群平均随访时间为2.58人年。在首次用药180 d内，共有412例（0.33%）他汀类药物用药者发生ALI，其中，开发队列305例（0.34%），验证队列107例（0.30%）。

2.2 模型表现

经过LASSO-Cox正则化方法筛选变量后

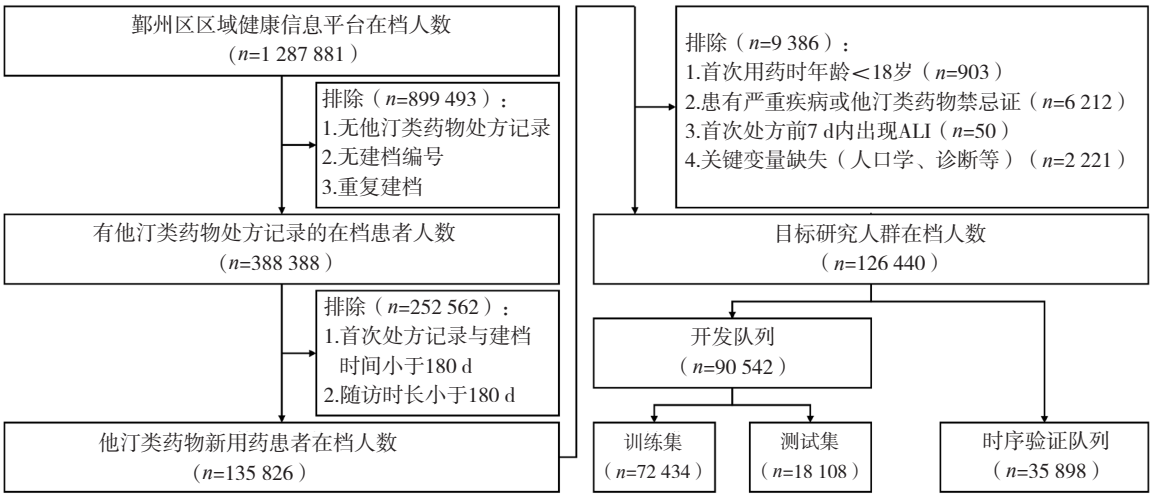

图1 研究人群筛选流程图及数据集拆分

Figure 1. Flow chart of study population and dataset splitting

表1 开发队列和验证队列中发生ALI与未发生ALI的他汀类药物用药者特征比较[M（P_{25}，P_{75}），n（%）]

Table 1. Comparison of characteristics between statin users with and without ALI in the development and validation cohorts [M (P_{25}, P_{75}), n (%)]

预测变量	开发队列（n=90 542）				时序验证队列（n=35 898）			
	未发生ALI（n=90 237）	发生ALI（n=305）	t/χ^2	P	未发生ALI（n=35 791）	发生ALI（n=107）	t/χ^2	P
首次他汀类药物用药年龄（岁）	61（52，69）	63（52，74）	2.7	0.003	62（53，70）	62（53，71）	0.74	0.771
性别			1.3	0.251			0.34	0.558
男性	41 062（45.7）	149（49.0）			17 802（50.0）	50（47.2）		
女性	48 740（54.3）	155（51.0）			17 788（50.0）	56（52.8）		
民族			–	0.428			–	0.999
汉族	88 287（99.8）	295（99.7）			35 174（99.8）	105（100）		

续表1

预测变量	开发队列（*n*=90 542）				时序验证队列（*n*=35 898）			
	未发生ALI（*n*=90 237）	发生ALI（*n*=305）	t/χ^2	*P*	未发生ALI（*n*=35 791）	发生ALI（*n*=107）	t/χ^2	*P*
其他	166（0.2）	1（0.3）			72（0.2）	0（0.0）		
受教育程度			3.9	0.140			0.23	0.891
初中及以下	74 157（85.2）	263（89.2）			24 874（73.9）	75（75.8）		
高中	8 515（9.8）	23（7.8）			5 335（15.8）	14（14.1）		
大学及以上	4 338（5.0）	9（3.1）			3 457（10.3）	10（10.1）		
婚姻状况			0.1	0.771			2.4	0.123
单身	16 741（28.0）	53（27.0）			6 467（24.9）	21（33.3）		
非单身	43 103（72.0）	143（73.0）			19 484（75.1）	42（56.7）		
职业			9.7	0.008			4.3	0.115
非体力劳动	54 051（60.2）	171（57.0）			21 150（59.4）	74（69.2）		
体力劳动	25 013（27.9）	105（35.0）			7 982（22.4）	17（15.9）		
综合性劳动	10 658（11.9）	24（8.0）			6 450（18.1）	16（15.0）		
吸烟状况			0.89	0.628			–	0.930
从不吸烟	66 821（77.2）	228（79.4）			26 752（77.7）	79（79.0）		
戒烟	4 987（5.8）	16（5.6）			1 607（4.7）	4（4.0）		
吸烟	14 722（17.1）	43（15.0）			6 093（17.7）	17（17.0）		
饮酒状况			–	0.616			–	0.699
从不饮酒	48 221（55.9）	168（58.5）			16 054（46.6）	50（50.0）		
戒酒	767（0.9）	3（1.0）			305（0.9）	0（0.0）		
饮酒	37 315（43.2）	116（40.4）			18 094（52.5）	50（50.0）		
锻炼强度			4.0	0.263			7.0	0.071
基本无	16 480（19.1）	43（15.0）			6 999（20.3）	14（14.0）		
轻度	26 367（30.6）	99（34.5）			8 469（24.6）	35（35.0）		
低强度规律性	8 029（9.3）	27（9.4）			3 098（9.0）	10（10.0）		
经常	35 241（40.9）	118（41.1）			15 865（46.1）	41（41.0）		
饮食情况			2.1	0.358			–	0.695
荤素均衡	72 422（92.5）	232（91.3）			30 106（95.6）	83（95.4）		
高脂高糖高盐	1 977（2.5）	10（3.9）			615（2.0）	1（1.1）		
素食	3 858（4.9）	12（4.7）			764（2.4）	3（3.4）		
环境危险因素暴露史	117（0.2）	0（0.0）	–	0.999	102（0.3）	0（0.0）	–	0.999
药品过敏史	952（1.1）	3（1.0）	–	0.999	132（0.4）	0（0.0）	–	0.999
高血压家族史	3 945（6.6）	5（3.2）	5.4	0.085	1 613（5.2）	0（0.0）	–	0.033
糖尿病家族史	1 118（1.9）	1（0.6）	–	0.378	462（1.5）	0（0.0）	–	0.631
冠心病家族史	160（0.3）	1（0.6）	–	0.345	77（0.2）	0（0.0）	–	0.999
病毒性肝炎家族史	5（0.0）	0（0.0）	–	0.999	5（0.0）	0（0.0）	–	0.999
CCI 得分			102.2	＜0.001			–	＜0.001
0	14 518（16.1）	12（3.9）			2 498（7.0）	2（1.9）		
1	20 770（23.0）	34（11.1）			5 374（15.0）	10（9.3）		
2	18 795（20.8）	55（18.0）			6 633（18.5）	10（9.3）		
≥3	36 154（40.1）	204（66.9）			21 286（59.5）	85（79.4）		
高脂血症	80 581（89.3）	251（82.3）	15.6	＜0.001	31 539（88.1）	89（83.2）	2.5	0.115
高血压	71 153（78.9）	256（83.9）	4.7	0.030	27 142（75.8）	89（83.2）	3.1	0.076
糖尿病（无并发症）	20 180（22.4）	86（28.2）	6.0	0.015	7 776（21.7）	24（22.4）	0.0	0.860
糖尿病（伴并发症）	2 750（3.0）	11（3.6）	0.3	0.571	1 541（4.3）	2（1.9）	–	0.334
心血管疾病	79 852（88.5）	279（91.5）	2.7	0.103	32 942（92.0）	102（95.3）	1.6	0.209
轻度肝脏疾病	16 864（18.7）	86（28.2）	18.1	＜0.001	9 873（27.6）	41（38.3）	6.1	0.013
中重度肝脏疾病	618（0.7）	8（2.6）	16.6	0.001	763（2.1）	10（9.3）	26.3	＜0.001

续表1

预测变量	开发队列（n=90 542）				时序验证队列（n=35 898）			
	未发生ALI（n=90 237）	发生ALI（n=305）	t/χ^2	P	未发生ALI（n=35 791）	发生ALI（n=107）	t/χ^2	P
肾脏疾病	2 717（3.0）	19（6.2）	10.7	0.001	1 612（4.5）	11（10.3）	8.2	0.009
他汀类药物种类			4.0	0.135			0.9	0.637
脂水双溶型	41 162（45.6）	148（48.5）			23 184（64.8）	74（69.2）		
亲脂型	37 346（41.4）	110（36.1）			2 735（7.6）	7（6.5）		
亲水型	11 729（13.0）	47（15.4）			9 872（27.6）	26（24.3）		
他汀类药物单次剂量（mg）			2.1	0.353			16.4	0.002
≤10	25 072（27.8）	91（29.8）			16 467（46.0）	38（35.5）		
>10~20	35 349（39.2）	125（41.0）			18 816（52.6）	63（58.9）		
>20	29 811（33.0）	89（29.2）			508（1.4）	6（5.6）		
他汀类药物剂型			–	0.999			33.9	<0.001
片剂/胶囊	90 177（99.9）	305（100.0）			35 527（99.3）	101（94.4）		
其他	60（0.1）	0（0.0）			264（0.7）	6（5.6）		
他汀类药物给药途径			10.8	0.004			6.7	0.010
口服	88 811（98.4）	293（96.1）			33822（94.5）	95（88.8）		
其他	1 426（1.6）	12（3.9）			1969（5.5）	12（11.2）		
他汀类药物用药记录来源			85.3	<0.001			34.7	<0.001
门诊处方	87 865（97.4）	271（88.9）			35 532（99.3）	101（94.4）		
住院记录	2 372（2.6）	34（11.1）			259（0.7）	6（5.6）		
既往用药情况	77 208（85.6）	281（92.1）	10.6	0.001	33 001（92.2）	93（86.9）	4.1	0.042
既往用药品种数	3（1, 5）	4（2, 6）	5.9	<0.001	3（2, 5）	4（2, 5）	4.6	0.148
合并用药品种数	5（3, 7）	6（4, 7）	7.9	<0.001	3（2, 4）	4（3, 5）	7.0	<0.001

注：CCI. 查尔森合并症指数（Charlson comorbidity index）。

（图2），共纳入16个预测变量：性别、受教育程度、饮酒状况、锻炼强度、轻度肝损伤既往史、中重度肝损伤既往史、肾损伤既往史、心血管疾病既往史、查尔森合并症指数（Charlson comorbidity index，CCI）得分、既往用药史、既往用药类数、合并用药类数、他汀类药物用药类型、单次用药剂量、给药途径和药品记录来源。

网格搜索结合十折交叉验证确定XGBoost模型的最佳超参数，具体如下：学习率为0.19，最大深度为10，最小叶子节点样本数为1，子样本比例为0.9，列采样比例为1，最小分裂增益为0.1，阳性样本误分类代价为30。使用特征增益作为评价指标，对各个变量对模型预测的贡献进行了评估，前5位重要的变量如下：合并用药类数（0.193）、既往用药类数（0.171）、锻炼强度（0.098）、单次用药剂量（0.089）、CCI得分（0.088）。180 d时研究对象发生ALI的基线风险为0.001 9。

在内部验证中，模型的Brier分数为0.004 3 [95%CI（0.003 8，0.004 9）]。区分度方面，模型在180 d时的Harrell's C指数为0.761 [95%CI（0.725，0.794）]，时间依赖性ROC曲线下面积为0.743 [95%CI（0.737，0.748）]，表明模型具有良好的区分度。在时序验证中，模型同样表现出良好的性能，Brier分数为0.004 4 [95%CI（0.003 6，0.005 2）]，Harrell's C指数和时间依赖性ROC曲线下面积分别为0.703 [95%CI（0.614，0.781）和0.694 [95%CI（0.636，0.751）]。模型在180 d时的校准度良好，见图2，在开发队列和验证队列中，预测概率与实际概率均具有较高的一致性。

进一步分析模型在不同亚组人群中的表现，结果显示模型在各种他汀类药物用药类型和不同单次用药剂量的亚组中均具有良好的整体表现和区分度，见表2。特别是在脂水双溶型和单次剂量≤10 mg的亚组中，模型的表现尤为突出。此外，该模型在首次用药后1个月到1年的随访时间内表现出良好的区分度（Harrell's C指数>0.7），见表3。校准曲线图也表明，模型在首次用药后30 d、90 d、180 d，预测概率和实际概率具有良好的一致性，见图3。

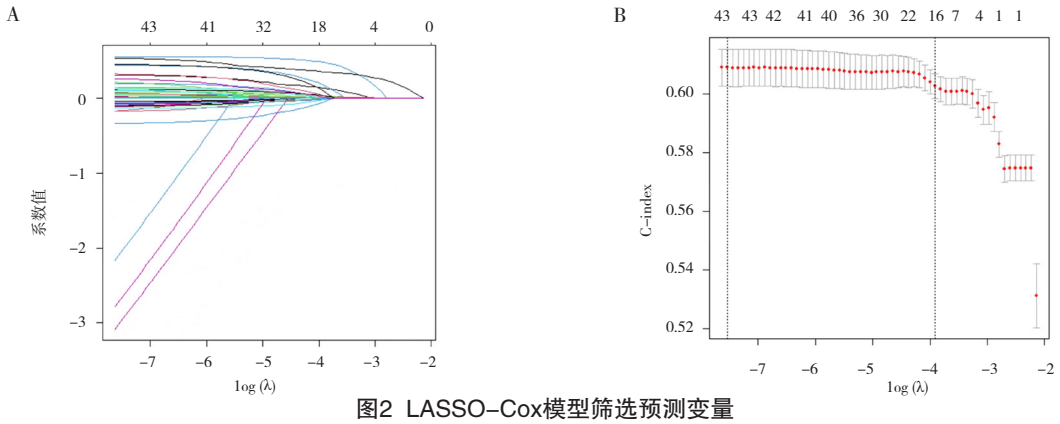

图2 LASSO–Cox模型筛选预测变量

Figure 2. LASSO–Cox model for predictors selection

注：A. 系数路径图；B. 交叉验证误差图。

表2 首次用药后180 d时模型在不同用药特征亚组中的表现

Table 2. Performance of the model across different medication characteristic subgroups at 180 days after initial statin use

亚组	具体分类	评价维度		
		Brier 分数（95%CI）	Harrell's C指数（95%CI）	AUC（95%CI）
他汀类药物用药种类	脂水双溶型	0.004 5（0.003 7，0.005 3）	0.760（0.704，0.808）	0.742（0.699，0.785）
	亲脂型	0.003 9（0.003 1，0.004 7）	0.748（0.693，0797）	0.721（0.669，0.773）
	亲水型	0.005 0（0.003 4，0.006 6）	0.803（0.692，0.881）	0.802（0.734，0.870）
他汀类药物单次剂量（mg）	≤10	0.004 6（0.003 5，0.005 6）	0.790（0.713，0.850）	0.773（0.719，0.827）
	＞10~20	0.004 4（0.003 5，0.005 3）	0.752（0.694，0.803）	0.722（0.675，0.769）
	＞20	0.004 1（0.003 2，0.004 9）	0.747（0.685，0.801）	0.744（0.688，0.799）

表3 首次用药后180 d内不同时间点的模型表现

Table 3. Model performance at different time points within 180 days after initial statin use

队列	具体分类	评价维度		
		Brier 分数（95%CI）	Harrell's C指数（95%CI）	AUC（95%CI）
开发队列	1个月	0.002 2（0.001 9，0.002 6）	0.749（0.701，0.796）	0.732（0.724，0.740）
	3个月	0.003 4（0.002 9，0.003 8）	0.762（0.720，0.804）	0.744（0.738，0.751）
	6个月	0.004 3（0.003 8，0.004 9）	0.761（0.725，0.794）	0.742（0.737，0.748）
时序验证队列	1个月	0.002 0（0.001 5，0.002 5）	0.684（0.582，0.787）	0.672（0.597，0.748）
	3个月	0.003 0（0.002 4，0.003 7）	0.682（0.587，0.775）	0.668（0.602，0.735）
	6个月	0.004 4（0.003 6，0.005 2）	0.703（0.614，0.781）	0.693（0.636，0.751）

注：AUC. 曲线下面积（area under curve）。

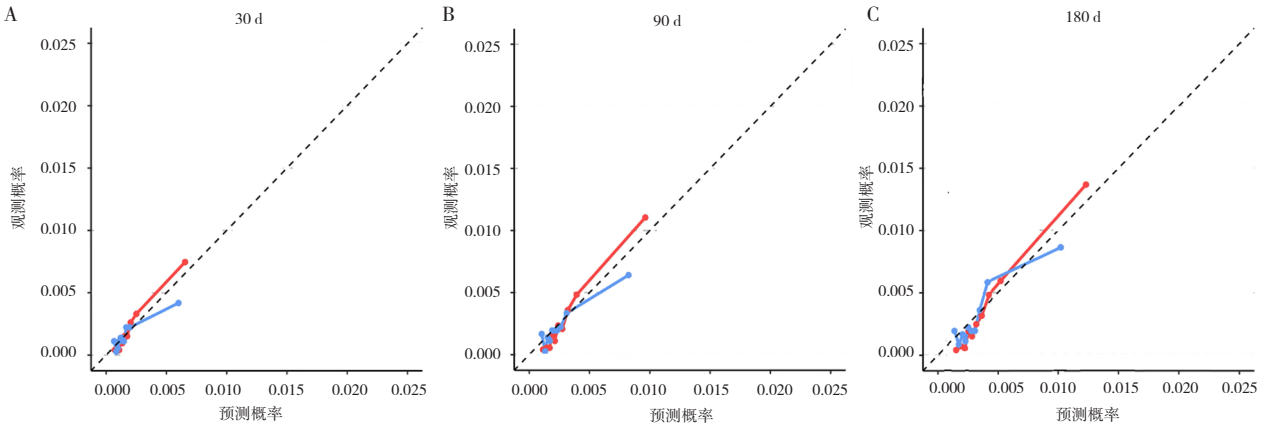

图3 不同时间点的校准曲线图

Figure 3. Calibration plots at different time points

注：A. 30 d时点校准曲线图；B. 90 d时点校准曲线图；C. 180 d时点校准曲线图。

3 讨论

本研究利用鄞州区区域健康信息平台，基于 XGBoost 算法针对首次他汀类药物用药人群开发了用药后 180 d 内发生 ALI 不良反应的预测模型，并通过时序验证评估了其性能。该模型整合了 6 个维度的 16 个预测变量，这些变量涵盖人口学特征、生活方式、家族史、既往史、他汀类药物用药情况以及合并用药情况，均为临床中常规收集或容易获取的指标，确保了模型在实际应用中的可行性和便捷性。模型在区分度和校准度方面表现良好，显示出较强的预测能力和稳定性，能够为临床医生提供可靠的个体化风险评估工具。

预测因子方面，合并用药类数和既往用药类数是模型中最主要的预测变量，与现有文献 [34] 中的发现一致，表明复杂用药情况可能增加他汀类药物引起的 ALI 风险。此外，患者的锻炼强度、饮酒习惯和 CCI 得分也对模型的预测效能贡献显著，说明生活方式和患者的整体健康状况都是评估 ALI 风险的重要因素。医生应鼓励他汀类药物用药者戒酒并加强运动，以降低 ALI 的风险 [35]。通过改善生活方式，患者的整体健康状况也能得到提升，从而减少其他潜在并发症的发生。大剂量他汀类药物可能增加 ALI 的风险，医生在开具处方时应谨慎评估患者的初始用药剂量，必要时可逐步调整剂量，以确保疗效的同时最大限度地减少 ALI 发生 [36]。脂水双溶型他汀类药物，如阿托伐他汀，更易导致 ALI，这一发现与既往文献 [37] 相吻合。临床医生应特别关注使用阿托伐他汀的患者，定期监测其肝功能，发现异常时及时调整治疗方案 [37–38]。基因、表型等维度的预测变量在预测药物性肝损伤方面具有更高的敏感性和特异性，但其检测成本较高，考虑到本研究的目标是开发一个简便、经济且易于临床应用的模型，因此未纳入上述指标 [39–41]。

考虑到 ALI 在他汀类药物用药者中的罕见性，本研究采用 XGBoost 算法结合代价敏感学习的方法，通过增加出现 ALI 的用药者样本的权重，提高了模型对 ALI 的分类性能。进一步地，以 XGBoost 输出的风险评分为预测变量构建 Cox 模型，为他汀类药物用药者在 180 d 内发生 ALI 的个体化绝对风险提供了预测，有助于医生进行患者精准分层，从而提高预防和管理的针对性。亚组分析和敏感性分析结果表明，该模型在不同用药情况和用药后不同时间段内均表现出良好的预测效果，尤其是在临床较为常见用药场景，即脂水双溶型他汀类药物和单次剂量小于 10 mg 的亚组中 [42]，模型的性能尤为突出，显示出较大的应用价值。因此，推荐临床医生在开具他汀类药物处方时首先使用该模型对患者的 ALI 风险进行预测，并对高风险患者进行重点关注，增加肝功能监测频率，以便早期发现问题并及时调整治疗方案，最大限度地减少 ALI 的发生。

本研究具有以下优势：①开发了首个专门针对 ALI 的预测模型，基于首次他汀类药物用药者的数据，为医生提供了一个有效的工具来评估患者发生 ALI 的风险；②通过使用 XGBoost 算法结合代价敏感学习方法，模型有效处理了 ALI 发生率低导致的类不平衡问题，提高了阳性结局的预测效果；③该模型整合了多个维度的临床变量，包括人口学特征、生活方式、既往史和用药情况等，均为临床上易于获得的数据，能够帮助医生实现个体化风险评估，精准识别高风险患者并进行有效管理。

本研究也存在一些局限性：① ALI 结局的定义主要依赖于生化指标，未进行病例验证，但这部分影响有限，因为 ALI 定义综合了多项基于真实世界数据开展的药物肝毒性研究，具有较高的准确性，而同类研究中进行病例验证的情况也较少 [43]；②模型的预测变量均为基线时的固定变量，未考虑患者在用药状态的动态变化，这可能限制模型的预测能力；③尽管本研究进行了时序验证，但验证队列与开发队列仍来自同一人群，因此结果的外推性受到一定限制，未来研究中应使用多中心数据进行地域外部验证 [44]。

综上，本研究开发并验证了一个基于 XGBoost 的他汀类药物用药者发生 ALI 的预测模型。该模型在不同用药情形的患者中均表现出良好的区分度和校准度，为临床医生提供了可靠的个体化评估工具，有助于实现风险分层，优化治疗方案，从而减少 ALI 的发生。

利益冲突声明：作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

1 Kim DS, Scherer PE. Obesity, diabetes, and increased cancer progression[J]. Diabetes Metab J, 2021, 45(6): 799–812. DOI: 10.4093/dmj.2021.0077.

2 中国血脂管理指南修订联合专家委员会 . 中国血脂管理指南（基层版 2024 年）[J]. 中华心血管病杂志 , 2024, 52(4): 330–337. [Joint Committee on the Chinese Guidelines for Lipid Management. Chinese guideline for lipid management (primary care version 2024)[J]. Chinese Journal of Cardiology, 2024, 52(4): 330–337.] DOI: 10.3760/cma.j.cn112148–20240102–00002.

3 Newman CB, Preiss D, Tobert JA, et al. Statin safety and associated adverse events: a scientific statement from the American Heart Association[J]. Arterioscler Thromb Vasc Biol, 2019, 39(2): e38–e81. DOI: 10.1161/atv.0000000000000073.

4 韦诗友 , 郑莉 . 他汀类药物在老年患者临床应用安全性的评价进展 [J]. 华西医学 , 2017, 32(2): 290–297. DOI: 10.7507/1002–0179.201507088.

5 Björnsson E, Jacobsen EI, Kalaitzakis E. Hepatotoxicity associated with statins: reports of idiosyncratic liver injury post–marketing[J]. J Hepatol, 2012, 56(2): 374–380. DOI: 10.1016/j.jhep.2011.07.023.

6 Law M, Rudnicka AR. Statin safety: a systematic review[J]. Am J Cardiol, 2006, 97(8a): 52c–60c. DOI: 10.1016/j.amjcard.2005.12.010.

7 孔飞飞 , 谭兴起 , 郭良君 . 阿托伐他汀钙致肝功能异常 [J]. 药物流行病学杂志 , 2011, 20(5): 267–267. DOI: 10.19960/j.cnki.issn1005–0698.2011.05.022.

8 他汀类药物安全性评价工作组 . 他汀类药物安全性评价专家共识 [J]. 中华心血管病杂志 , 2014, 42(11): 890–894. [Working Group of the Evaluation on the Safety of Statins. Experts consensus of the evaluation on the statin drugs safty[J]. Chinese Journal of Cardiology, 2014, 42(11): 890–894.] DOI: 10.3760/cma.j.issn.0253–3758.2014.11.002.

9 Wang B, Huang S, Li S, et al. Hepatotoxicity of statins: a real–world study based on the US Food and Drug Administration Adverse Event Reporting System database[J]. Front Pharmacol, 2024, 15: 1502791. DOI: 10.3389/fphar.2024.1502791.

10 Horodinschi RN, Stanescu AMA, Bratu OG, et al. Treatment with statins in elderly patients[J]. Medicina (Kaunas), 2019, 55(11): 721. DOI: 10.3390/medicina55110721.

11 Björnsson ES. Hepatotoxicity of statins and other lipid–lowering agents[J]. Liver Int, 2017, 37(2): 173–178. DOI: 10.1111/liv.13308.

12 Allen RM, Marquart TJ, Albert CJ, et al. miR–33 controls the expression of biliary transporters, and mediates statin– and diet–induced hepatotoxicity[J]. EMBO Mol Med, 2012, 4(9): 882–895. DOI: 10.1002/emmm.201201228.

13 Cooper–DeHoff RM, Niemi M, Ramsey LB, et al. The clinical pharmacogenetics implementation consortium guideline for *SLCO1B1*, *ABCG2*, and *CYP2C9* genotypes and statin–associated musculoskeletal symptoms[J]. Clin Pharmacol Ther, 2022, 111(5): 1007–1021. DOI: 10.1002/cpt.2557.

14 Xiong Y, Liu X, Wang Q, et al. Machine learning–based prediction model for the efficacy and safety of statins[J]. Front Pharmacol, 2024, 15: 1334929. DOI: 10.3389/fphar.2024.1334929.

15 Lin H, Tang X, Shen P, et al. Using big data to improve cardiovascular care and outcomes in China: a protocol for the Chinese Electronic Health Records Research in Yinzhou (CHERRY) Study[J]. BMJ Open, 2018, 8(2): e019698. DOI: 10.1136/bmjopen–2017–019698.

16 赵厚宇 , 柴三葆 , 孙烨祥 , 等 . 二甲双胍与 2 型糖尿病患者痴呆症发病风险相关性的回顾性队列研究 [J]. 中国糖尿病杂志 , 2024, 32(8): 567–575. [Zhao HY, Chai SB, Sun YX, et al. Retrospective cohort study on the relationship between metformin and the risk of dementia in patients with type 2 diabetes mellitus[J]. Chinese Journal of Diabetes, 2024, 32(8): 567–575.] DOI: 10.3969/j.issn.1006–6187.2024.08.002.

17 Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application[J]. Curr Epidemiol Rep, 2015, 2(4): 221–228. DOI: 10.1007/s40471–015–0053–5.

18 De Vriese AS. Should statins be banned from dialysis?[J]. J Am Soc Nephrol, 2017, 28(6): 1675–1676. DOI: 10.1681/asn.2017020201.

19 Danan G, Teschke R. Roussel uclaf causality assessment method for drug–induced liver injury: present and future[J]. Front Pharmacol, 2019, 10: 853. DOI: 10.3389/fphar.2019.00853.

20 中国医药生物技术协会药物性肝损伤防治技术专业委员会 , 中华医学会肝病学分会药物性肝病学组 . 中国药物性肝损伤诊治指南（2023 年版）[J]. 中华肝脏病杂志 , 2023, 31(4): 355–384. [Technology Committee on DILI Prevention and Management, Chinese Medical Biotechnology Association; Study Group of Drug–Induced Liver Disease, Chinese Medical Association for the Study of Liver Diseases. Chinese guideline for diagnosis and management of drug–induced liver injury (2023 version)[J]. Chinese Journal of Hepatology, 2023, 31(4): 355–384.] DOI: 10.3760/cma.j.cn501113–20230419–00176–1.

21 García–Rodríguez LA, González–Pérez A, Stang MR, et al. The safety of rosuvastatin in comparison with other statins in over 25,000 statin users in the Saskatchewan Health Databases[J]. Pharmacoepidemiol Drug Saf, 2008, 17(10): 953–961. DOI: 10.1002/pds.1602.

22 Douglas IJ, Langham J, Bhaskaran K, et al. Orlistat and the risk of acute liver injury: self controlled case series study in UK Clinical Practice Research Datalink[J]. BMJ, 2013, 346: f1936. DOI: 10.1136/bmj.f1936.

23 Falconer N, Nand S, Liow D, et al. Development of an electronic patient prioritization tool for clinical pharmacist interventions[J]. Am J Health Syst Pharm, 2014, 71(4): 311–320. DOI: 10.2146/ajhp130247.

24 Alfirevic A, Neely D, Armitage J, et al. Phenotype standardization for statin–induced myotoxicity[J]. Clin Pharmacol Ther, 2014, 96(4): 470–476. DOI: 10.1038/clpt.2014.121.

25 White IR, Royston P, Wood AM. Multiple imputation using

chained equations: Issues and guidance for practice[J]. Stat Med, 2011, 30(4): 377–399. DOI: 10.1002/sim.4067.

26 Heidarian Miri H, Hassanzadeh J, Rajaeefard A, et al. Multiple imputation to correct for nonresponse bias: application in non–communicable disease risk factors survey[J]. Glob J Health Sci, 2015, 8(1): 133–142. DOI: 10.5539/gjhs.v8n1p133.

27 Mao B, Ma J, Duan S, et al. Preoperative classification of primary and metastatic liver cancer via machine learning–based ultrasound radiomics[J]. Eur Radiol, 2021, 31(7): 4576–4586. DOI: 10.1007/s00330–020–07562–6.

28 Bradic J, Fan J, Jiang J. Regularization for Cox's proportional hazards model with NP–dimensionality[J]. Ann Stat, 2011, 39(6): 3092–3120. DOI: 10.1214/11–aos911.

29 Friedman JH. Stochastic gradient boosting[J]. Comput Stat Data An, 2002, 38(4): 367–378. DOI: 10.1016/S0167–9473(01)00065–2.

30 Aldraimli M, Soria D, Grishchuck D, et al. A data science approach for early–stage prediction of patient's susceptibility to acute side effects of advanced radiotherapy[J]. Comput Biol Med, 2021, 135: 104624. DOI: 10.1016/j.compbiomed.2021.104624.

31 Pan Z, Zhang R, Shen S, et al. OWL: an optimized and independently validated machine learning prediction model for lung cancer screening based on the UK Biobank, PLCO, and NLST populations[J]. EBioMedicine, 2023, 88: 104443. DOI: 10.1016/j.ebiom.2023.104443.

32 Hosmer DW Jr LS, Sturdivant RX, eds. Applied logistic regression[M]. Hoboken: John Wiley & Sons, Inc., 2013: 117.

33 Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement[J]. BMJ, 2015, 350: g7594. DOI: 10.1136/bmj.g7594.

34 郭恒，李丹丹，温爱萍，等. 9城市老年患者使用他汀类药物潜在药物相互作用分析[J]. 中南药学，2023, 21(5): 1377–1382. [Guo H, Li DD, Wen AP, et al. Potential drug–drug interactions with statins among elderly patients in 9 cities in China[J]. Central South Pharmacy, 2023, 21(5): 1377–1382.] DOI: 10.7539/j.issn.1672–2981.2023.05.042.

35 Bytyçi I, Penson PE, Mikhailidis DP, et al. Prevalence of statin intolerance: a Meta–analysis[J]. Eur Heart J, 2022, 43(34): 3213–3223. DOI: 10.1093/eurheartj/ehac015.

36 Escobar C, Echarri R, Barrios V. Relative safety profiles of high dose statin regimens[J]. Vasc Health Risk Manag, 2008, 4(3): 525–533. DOI: 10.2147/vhrm.s2048.

37 国家药品不良反应监测中心. 国家药品不良反应监测年度报告（2022年）[J]. 中国病毒病杂志，2023, 13(4): 245–251. [Chinese national center for adverse drug reaction monitoring. Annual report of national adverse drug reaction monitoring (2022)[J]. Chinese Journal of Viral Diseases, 2023, 13(4): 245–251.] DOI: 10.16505/j.2095–0136.2023.4003.

38 Cai T, Abel L, Langford O, et al. Associations between statins and adverse events in primary prevention of cardiovascular disease: systematic review with pairwise, network, and dose–response Meta–analyses[J]. BMJ, 2021, 374: n1537. DOI: 10.1136/bmj.n1537.

39 Kaniwa N, Saito Y. Pharmacogenomics of severe cutaneous adverse reactions and drug–induced liver injury[J]. J Hum Genet, 2013, 58(6): 317–326. DOI: 10.1038/jhg.2013.37.

40 Wang CW, Preclaro IAC, Lin WH, et al. An updated review of genetic associations with severe adverse drug reactions: translation and implementation of pharmacogenomic testing in clinical practice[J]. Front Pharmacol, 2022, 13: 886377. DOI: 10.3389/fphar.2022.886377.

41 Floyd JS, Bloch KM, Brody JA, et al. Pharmacogenomics of statin–related myopathy: Meta–analysis of rare variants from whole–exome sequencing[J]. PLoS One, 2019, 14(6): e0218115. DOI: 10.1371/journal.pone.0218115.

42 江林双，陈茂伟. 阿托伐他汀药物性肝损伤的临床特征分析[J]. 中国全科医学，2024, 27(30): 3772–3775, 3783. [Jiang LS, Chen MW. Clinical characteristics of liver injury induced by atorvastatin[J]. Chinese General Practice, 2024, 27(30): 3772–3775, 3783.] DOI: 10.12114/j.issn.1007–9572.2023.0698.

43 郜隽，于森，张琳，等. 基于自然语言处理技术的儿童病例注射用尖吻蝮蛇血凝酶不良反应主动监测方法构建与验证研究[J]. 中国药物警戒，2023, 20(9): 1011–1016. [Tai J, Yu M, Zhang L. Active monitoring of adverse reactions related to haemocoagulase agkistrodon for injection in children[J]. Chinese Journal of Pharmacovigilance, 20(9): 1011–1016.] DOI: 10.19803/j.1672–8629.20220736.

44 Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment[J]. Heart, 2012, 98(9): 691–698. DOI: 10.1136/heartjnl–2011–301247.

收稿日期：2025年02月10日 修回日期：2025年04月02日
本文编辑：杨 燕 冼静怡