·论著·一次研究·

# 基于5种机器学习算法构建抗菌药物致血小板减少症风险预测模型

王 敏，沈晓妍

成都市青白江区人民医院药剂科（成都 610300）

**【摘要】目的** 探讨发生抗菌药物致血小板减少症的影响因素，并基于机器学习（ML）建立预测模型。**方法** 选择2020年1月至2022年12月于成都市青白江区人民医院接受抗菌药物治疗的患者作为研究对象，以发生抗菌药物致血小板减少症作为结局变量，应用最小绝对收缩和选择算子回归及多因素Logistic回归（LR）分析筛选抗菌药物致血小板减少症的独立影响因素；应用合成少数过采样技术-标称连续技术进行过采样，并依据特征变量建立5种不同ML模型，分析和识别最佳模型，使用受试者工作特征曲线下面积（AUC）、校准曲线、Hosmer-Lemeshow检验和决策曲线分析（DCA）评估抗菌药物致血小板减少症预测模型性能，并采用Delong检验比较模型间AUC差异，以传统LR为参考模型计算综合判别改善指数（IDI）与净重分类改善指数（NRI），综合选取最佳预测模型；使用沙普利加和解释（SHAP）方法解释关键特征对模型预测结果的贡献。**结果** 共纳入701例患者，其中41例（5.85%）发生了抗菌药物致血小板减少症。多因素LR分析显示，总胆红素（TBIL）水平较高［OR=1.285，95%CI（1.118，1.477）］是抗菌药物致血小板减少症的独立危险因素，而白蛋白（ALB）水平较高［OR=0.954，95%CI（0.912，0.998）］、肌酐清除率（CCr）水平较高［OR=0.856，95%CI（0.749，0.978）］为保护因素（$P < 0.05$）；极端梯度提升（XGBoost）模型预测效能均最佳；Hosmer-Lemeshow检验和校准曲线显示，XGBoost模型预测风险与观察风险高度一致；DCA曲线显示XGBoost模型在整个阈值范围（0~1.0）内表现最佳、净获益最高；基于Bootstrap内部验证结果，Delong检验显示XGBoost模型与LR、随机森林、决策树、支持向量机模型比较差异具有统计学意义（均$P < 0.05$）；以LR回归为参照，XGBoost模型IDI、NRI明显优于其他模型（均$P < 0.05$）；SHAP结果显示，TBIL水平、ALB水平及CCr水平对预测抗菌药物致血小板减少症具有重要影响。**结论** TBIL水平、ALB水平、CCr水平为抗菌药物致血小板减少症的影响因素。XGBoost模型对预测抗菌药物致血小板减少症具有较好的预测效能，为筛选抗菌药物致血小板减少症高风险患者提供参考。

**【关键词】**机器学习；抗菌药物；血小板减少症；预测模型；危险因素；极端梯度提升；沙普利加和解释

**【中图分类号】**R558+.2 **【文献标志码】**A

## Development of a predictive model for the risk of antimicrobial drug-induced thrombocytopenia using 5 machine learning algorithms

WANG Min, SHEN Xiaoyan

*Department of Pharmacy, People's Hospital of Qingbaijiang District, Chengdu City, Chengdu 610300, China*
*Corresponding author: WANG Min, Email: 546855074@qq.com*

DOI: 10.12173/j.issn.1005-0698.202601056
基金项目：2023年成都市医学科研课题立项项目（2023091）
通信作者：王敏，主管药师，Email：546855074@qq.com

【Abstract】Objective To explore the influencing factors of antimicrobial drug-induced thrombocytopenia and construct a predictive model based on machine learning (ML). Methods Patients who received antimicrobial therapy at Qingbaijiang District People's Hospital in Chengdu from January 2020 to December 2022 were selected as study subjects. Antimicrobial drug-induced thrombocytopenia was set as the outcome variable. Independent risk factors for antimicrobial-induced thrombocytopenia were identified using least absolute shrinkage and selection operator regression and multivariate Logistic regression (LR) analysis. Synthetic Minority Over-sampling Technique-Nominal Continuous was applied for oversampling, and 5 different ML models were constructed based on feature variables to identify the optimal model. The performance of the predictive model for antimicrobial-induced thrombocytopenia was evaluated using area under the receiver operating characteristic curve (AUC), calibration curves, the Hosmer-Lemeshow test, and decision curve analysis (DCA). The Delong test was used to compare differences in AUC among models, and the integrated discrimination improvement index (IDI) and net reclassification improvement index (NRI) were calculated with conventional LR as the reference model. The best predictive model was comprehensively selected. The contribution and impact of key features on model predictions were interpreted using SHapley Additive exPlanations (SHAP). Results A total of 701 patients were included, among whom 41 (5.85%) developed antibiotic-induced thrombocytopenia. Multivariate LR analysis revealed that elevated total bilirubin (TBIL) [OR=1.285, 95%CI (1.118, 1.477)] was an independent risk factor for antibiotic-induced thrombocytopenia, while elevated albumin (ALB) [OR=0.954, 95%CI (0.912, 0.998)] and higher creatinine clearance (CCr) [OR=0.856, 95%CI (0.749, 0.978)] were protective factors ($P$ < 0.05). The eXtreme gradient boosting (XGBoost) model demonstrated the best predictive performance. Hosmer-Lemeshow test and the calibration curve of the XGBoost model showed a high consistency between predicted and observed risks. The DCA curve revealed that the XGBoost model achieved the highest net benefit across the entire threshold range (0-1.0). Based on bootstrap internal validation, the DeLong test showed statistically significant differences between the XGBoost model and the LR, random forest, decision tree, and support vector machine models (all $P$ < 0.05). Compared with LR as the reference, the XGBoost model significantly outperformed other models in terms of IDI and NRI (all $P$ < 0.05). SHAP analysis indicated that TBIL, ALB, and CCr levels played important roles in predicting antibiotic-induced thrombocytopenia. Conclusion TBIL levels, ALB levels and CCr levels are the influencing factors of antimicrobial drug-induced thrombocytopenia. The XGBoost model has a good predictive performance for predicting platelet reduction caused by antibacterial drugs, providing a reference for screening high-risk patients with antimicrobial drug-induced thrombocytopenia.

【Keywords】Machine learning; Antimicrobial drugs; Thrombocytopenia; Prediction model; Risk factors; Extreme gradient boosting; SHapley Additive exPlanation

抗菌药物是临床抗感染治疗的基石，但其引发的药物不良反应已成为威胁患者安全的重要因素[1]。其中，药源性血小板减少症（drug-induced thrombocytopenia，DITP）作为常见且严重的血液系统不良反应，可导致皮肤黏膜出血、脏器出血甚至死亡，不仅明显增加患者住院时间与医疗负担，还可能迫使抗感染治疗中断，影响原发病预后[2]。研究显示[3]，欧美国家DITP的年发病率为1/10万，而在危重患者群体中，该疾病的患病率约达25%；有超过300种药物可诱发DITP，涵盖抗菌药物、肝素及抗肿瘤药等多个类别。在临床实践中，接受利奈唑胺治疗的患者血小板减少症发生率可达37%，该不良反应不仅可快速进展至重度血小板减少症，还可能进而诱发颅内出血等致命性并发症，而颅内出血作为血小板减少症患者的常见死亡原因，严重威胁其生命安全，因此，临床中需密切监测DITP的发生情况[4]。目前，临床对DITP的识别多依赖用药后血小板计数（platelet count，PLT）的动态监测，缺乏早期预警能力[5]。传统风险评估方法多基于单因素分析

或 Logistic 回归模型，仅能纳入有限的临床变量（如年龄、基础疾病、用药剂量等），难以全面捕捉患者个体特征、药物暴露模式及实验室指标间的复杂关联。这种局限性导致DITP的预测准确率偏低，临床干预往往滞后于病情进展。近年来，已有研究将机器学习（machine learning，ML）应用于利奈唑胺等特定抗菌药物致血小板减少症的预测，证实其较传统Logistic回归模型具有更优性能[6-7]。但目前研究多局限于单一抗菌药物，缺乏对多种常用抗菌药物致血小板减少症的综合预测模型，难以满足临床多样化用药场景的需求。此外，现有ML预测研究常直接选用单一算法，然而不同算法在数据适应性、非线性拟合能力及可解释性等方面存在显著差异，单一模型难以保证在所有临床场景下均表现最优。因此，本研究系统纳入β-内酰胺类、喹诺酮类、林可酰胺类等7类常用抗菌药物，覆盖临床90%以上的抗感染治疗场景，并同时构建传统逻辑回归（Logistic regression，LR）、随机森林（random forest，RF）、决策树（decision tree，DT）、极端梯度提升（extreme gradient boosting，XGBoost）及支持向量机（support vector machine，SVM）这5种覆盖不同算法原理的ML模型。通过比较其在同一数据集上的预测性能，筛选出最优模型，以提高模型选择的科学性与结果的外推稳健性。最终，本研究旨在联合多维临床指标，构建适用于多样化用药的综合预测模型，筛选最优预测模型并解析关键影响因素，为临床早期识别高风险患者提供参考。

## 1 资料与方法

### 1.1 研究对象

回顾性选择2020年1月至2022年12月于成都市青白江区人民医院接受抗菌药物治疗的患者作为研究对象。纳入标准：①年龄≥18岁；②住院期间接受抗菌药物治疗，且用药疗程≥3 d；③用药前基线PLT≥$100\times10^9\cdot L^{-1}$者。排除标准：①入院前确诊血液系统疾病、风湿系统疾病、脾功能亢进者；②入院前3个月内接受放化疗、靶向治疗、免疫治疗者；③用药期间合并使用肝素、抗血小板药物、非甾体抗炎药等其他可致血小板减少症的药物者；④用药期间出现脓毒症休克、弥散性血管内凝血、严重感染进展等可导致血小板减少症者；⑤临床资料不完整者。本研究已通过成都市青白江区人民医院伦理委员会审查批准［批号：2023年审（9）号］，并豁免知情同意。

### 1.2 样本量估算

依据样本估算方法 $N=\frac{Z^2_{1-\alpha/2}\times P\times(1-P)}{E^2}$，参考本院预实验中抗菌药物致血小板减少症发病率为5.96%，设定本研究中$P$=0.0596，$\alpha$=0.05，$E$=0.02，脱落率为20%，最终确定本研究最小总样本量为673例。

### 1.3 分组依据

本研究以患者住院期间是否发生抗菌药物致血小板减少症为结局变量，其诊断标准参考文献[8]：①用药前基线PLT≥$100\times10^9\cdot L^{-1}$；②抗菌药物用药≥2 d后，至停药后7 d内出现PLT<$100\times10^9\cdot L^{-1}$，且至少连续2次复查确认；③停药后PLT呈回升趋势，或恢复至正常水平；④排除上述标准中其他可导致血小板减少症的原因。依据是否发生抗菌药物致血小板减少症，将患者分为患病组和未患病组。

### 1.4 数据收集

通过医院电子病历系统收集患者临床资料以及抗菌药物使用情况，包括年龄、性别、红细胞计数（red blood cell count，RBC）、红细胞压积（hematocrit，HCT）、血红蛋白（hemoglobin，Hb）、PLT、碱性磷酸酶（alkaline phosphatase，ALP）、总胆红素（total bilirubin，TBIL）、白蛋白（albumin，ALB）、肌酐（creatinine，Cr）、肌酐清除率（creatinine clearance rate，CCr）、尿素氮（blood urea nitrogen，BUN）、胱抑素C（cystatin C，CysC）、降钙素原（procalcitonin，PCT）、凝血酶原时间（prothrombin time，PT）、活化部分凝血活酶时间（activated partial thromboplastin time，APTT）、抗菌药物类别及限定日剂量（defined daily dose，DDD）。

### 1.5 模型构建、评估及解释

#### 1.5.1 模型构建

先通过单因素分析剔除无统计学关联的噪声变量，降低后续建模维度；将单因素筛选后的变量纳入LASSO回归，经10折交叉验证选取最优λ，进一步压缩冗余特征、避免多重共线性，获取稳健关键特征；再以多因素LR识别独立影响

因素，最终将经统计学确认的独立危险因素纳入ML模型。本研究中患病组与未患病组病例数比值约为1∶16（41∶660），存在明显类别不平衡。研究采用1 000次Bootstrap抽样进行内部验证，每一轮Bootstrap均先将抽样数据集划分为训练子集与测试子集，仅在训练子集内采用合成少数过采样技术-标称连续（synthetic minority oversampling technique-nominal continuous，SMOTE-NC）处理阳性样本，测试子集始终保持原始类别分布、不做任何采样调整，以避免数据泄漏并保证模型评估真实可靠。在训练子集中将阳性样本过采样至与阴性样本数量接近（患病∶未患病≈1∶1.01），再基于采样后的训练子集构建RF、DT、LR、XGBoost及SVM 5种ML模型，所有模型的验证与性能评估均基于未采样的原始分布测试子集完成，通过严格的子集拆分与分步采样流程避免信息泄漏，确保评估结果无偏倚。经LASSO回归10折交叉验证选取最优正则化参数λ，对应Log（λ）=-5.72，在此参数下筛选关键特征并控制多重共线性。ML模型构建过程中，采用10折交叉验证结合网格搜索进行超参数调优，所有调参均在训练子集内完成，以避免过拟合与信息泄露。各模型主要参数设置如下：XGBoost模型决策树数量为100，最大深度为3，学习率为0.1；RF模型决策树数量为300，最大深度为5；DT模型以Gini系数为划分标准，最大深度为5；LR模型采用L2正则化，正则化参数C=1.0；SVM模型采用RBF核函数，惩罚参数C=1.0，gamma=0.1。

#### 1.5.2 模型评估

采用1 000次Bootstrap抽样进行模型内部验证，计算受试者工作特征曲线（receiver operating characteristic curve，ROC）的曲线下面积（area underthe curve，AUC）、Hosmer-Lemeshow检验、校准曲线及决策曲线分析（decision curve analysis，DCA）以综合评估模型效能；同时采用Delong检验比较模型间AUC差异，以LR为参考模型计算综合判别改善指数（integrated discrimination improvement，IDI）与净重分类改善指数（net reclassification improvement，NRI），综合选取最佳预测模型，保证验证结果稳定可靠。

#### 1.5.3 模型解释

本研究选取预测效能最优的XGBoost模型进行特征重要性分析，以权重值评估各变量对预测结果的贡献程度并排序。为进一步提高模型可解释性，采用沙普利加和解释（SHapley Additive exPlanation，SHAP）方法对关键预测因素进行可视化解释，直观展示各特征对模型输出的正向与负向影响，增强模型的临床可解释性与实用性。

### 1.6 统计学分析

利用SPSS 23.0统计软件和R 4.2.1软件进行数据分析。计数资料以例数和百分比［$n$（%）］表示，组间比较采用$\chi^2$检验或*Fisher*确切概率法。计量资料采用Shapiro-Wilk检验判断正态分布性，符合正态分布的计量资料以均数和标准差（$\bar{x}\pm s$）表示，组间比较采用独立样本$t$检验；不符合正态分布的计量资料以中位数和四分位数［$M$（$P_{25}$，$P_{75}$）］表示，组间比较采用Mann-Whitney $U$检验。$P$<0.05为差异具有统计学意义。

## 2 结果

### 2.1 两组患者临床资料对比

共纳入701例接受抗菌药物治疗的患者。患病组41例，未患病组660例。与未患病组患者相比，患病组患者的男性占比、TBIL、Cr、BUN、CysC、PCT、PT、APTT水平更高，RBC、ALB、CCr水平更低（$P$<0.05），见表1。

### 2.2 LASSO回归重要特征变量筛选

以表1中$P$<0.05的指标为自变量，以发生抗菌药物致血小板减少症作为结局变量，通过10折交叉验证分析，当Log（λ）=-5.72时，获得最优模型，并以最优值对应的5个变量作为关键变量，分别为TBIL、ALB、Cr、CCr及CysC（图1）。

### 2.3 多因素Logistic回归分析

将LASSO回归分析筛选出的5个关键变量采用向前逐步回归法纳入多因素Logistic回归分析。结果显示，TBIL水平较高［OR=1.285，95%CI（1.118，1.477）］是抗菌药物致血小板减少症的独立危险因素，而ALB水平较高［OR=0.954，95%CI（0.912，0.998）］、CCr水平较高［OR=0.856，95%CI（0.749，0.978）］为保护因素（$P$<0.05），见表2。

### 2.4 模型比较

本研究中患病患者及未患病患者病例数分别为41例和660例。患病组为少数类样本，经SMOTE-NC处理后该组的病例数为41×16=656例，此时患病组与未患病组的比值约为1∶1.01。经过SMOTE-NC过采样后，将筛选出的3个变量

分别纳入ML模型，计算模型的AUC、准确度、灵敏度、特异度及F1评分。以患者实际是否患病为状态变量，ROC曲线预测验证前后XGBoost模型的预测效能（验证前：AUC=0.906；验证后：0.914）高于其他模型（表3、图2）。校准曲线验证前后模型的预测概率与参考概率拟合优度均较好，其中XGBoost拟合优度更优（$\chi^2$=9.256、7.054，$P$=0.557、0.693），见图3。DCA曲线验证前后XGBoost模型曲线均高于其他模型，有一定的临床决策价值（图4）。以上结果均显示XGBoost模型精准度较高，性能较好。

Delong检验结果显示，XGBoost模型的AUC明显高于LR、RF、DT、SVM模型（$Z$=2.115、2.304、2.672、2.963，均$P$<0.05）；以LR回归为参照，XGBoost模型IDI=0.089［95%*CI*（0.042，0.126），$P$<0.001］，提示相较于LR模型，对抗菌药物致血小板减少症的综合判别能力明显提升；NRI=0.493［95%CI（0.327，0.579），$P$ < 0.001］，提示对抗菌药物致血小板减少症高风险患者的重分类能力明显优化，进一步证实XGBoost模型为最优预测模型（表4）。

## 2.5 最优模型可解释性分析

### 2.5.1 变量重要性分析

在模型性能评估中，XGBoost模型在各项评估指标上的表现均最佳。按照特征的平均SHAP绝对值降序排序，特征越靠上表示特征对整个模型的贡献程度越高，对模型输出结果的影响越大。变量重要性分析显示，TBIL水平是发生抗菌药物致血小板减少症最重要的因素，其次分别为ALB、CCr，见图5。

### 2.5.2 SHAP分析

蜂群图可以实现特征效应方向、作用强度及样本异质性的三维解析，按照SHAP绝对值降序排序，特征越靠上表示特征对疾病发生风险预警的权重越高。结果显示，高水平TBIL，低水平ALB、低水平CCr会增加抗菌药物致血小板减少症发生的风险，见图6。

**表1 两组患者临床资料比较［$M$（$P_{25}$，$P_{75}$）］**

Table 1. Comparison of clinical data between the two groups of patients［$M$（$P_{25}$，$P_{75}$）］

| 临床特征 | 患病组（$n$=41） | 未患病组（$n$=660） | $Z/\chi^2$ | $P$ |
|---|---|---|---|---|
| 年龄（岁） | 47.00（45.00，49.00） | 47.00（44.00，49.00） | 0.875 | 0.381 |
| 性别* | | | 4.261 | 0.039 |
| 男 | 27（65.85） | 325（49.24） | | |
| 女 | 14（34.15） | 335（50.76） | | |
| RBC（×$10^{12}$/L） | 4.15（3.64，4.64） | 4.41（4.08，4.81） | 2.836 | 0.005 |
| HCT（%） | 38.70（35.90，43.15） | 40.45（37.02，44.18） | 1.716 | 0.086 |
| Hb（g/L） | 129.00（119.00，143.50） | 134.00（121.00，147.00） | 1.469 | 0.142 |
| ALP（U/L） | 82.00（66.50，101.50） | 79.00（67.00，96.00） | −1.145 | 0.252 |
| TBIL（μmol/L） | 18.50（12.80，30.05） | 12.00（8.72，15.80） | −4.902 | <0.001 |
| ALB（g/L） | 39.60（35.45，41.05） | 41.55（39.12，43.80） | −4.416 | <0.001 |
| Cr（μmol/L） | 77.00（64.00，97.50） | 67.00（56.00，77.00） | −3.902 | <0.001 |
| CCr（mL/min） | 90.92（77.10，110.44） | 115.51（99.44，133.24） | −5.071 | <0.001 |
| BUN（mmol/L） | 5.60（4.35，6.75） | 4.60（3.60，5.60） | −3.271 | 0.001 |
| CysC（mg/L） | 0.98（0.82，1.14） | 0.78（0.68，0.90） | −5.071 | <0.001 |
| PCT（μg/L） | 0.06（0.00，0.69） | 0.00（0.00，0.05） | −4.286 | <0.001 |
| PT（s） | 13.40（12.10，15.65） | 12.10（11.30，13.20） | −3.976 | <0.001 |
| APTT（s） | 32.70（30.50，41.70） | 30.50（28.00，33.60） | −4.241 | <0.001 |
| DDD（/d） | 5.75（2.46，8.88） | 7.00（4.00，10.00） | −1.709 | 0.087 |
| 抗菌药物类别* | | | | |
| β-内酰胺类 | 37（90.24） | 516（78.18） | 3.372 | 0.066 |
| 大环内酯类 | 1（2.44） | 3（0.45） | – | 0.215# |
| 喹诺酮类 | 13（31.71） | 230（34.85） | 0.168 | 0.682 |
| 磺胺类 | 0（0.00） | 3（0.45） | – | 0.839# |
| 林可酰胺类 | 3（7.32） | 33（5.00） | 0.083 | 0.774 |
| 硝基咪唑类 | 3（7.32） | 23（3.48） | 0.696 | 0.404 |
| 利福霉素类 | 0（0.00） | 1（0.15） | – | 1.000# |

注：RBC.红细胞计数（red blood cell count）；HCT.红细胞压积（hematocrit）；Hb.血红蛋白（hemoglobin）；ALP.碱性磷酸酶（alkaline phosphatase）；TBIL.总胆红素（total bilirubin）；ALB.白蛋白（albumin）；Cr.肌酐（creatinine）；CCr.肌酐清除率（creatinine clearance rate）；BUN.尿素氮（blood urea nitrogen）；CysC.胱抑素C（cystatin C）；PCT.降钙素原（procalcitonin）；PT.凝血酶原时间（prothrombin time）；APTT.活化部分凝血活酶时间（activated partial thromboplastin time）；DDD.限定日剂量（defined daily dose）；*计数资料以例数和百分比［$n$（%）］表示；#采用*Fisher*确切概率法。

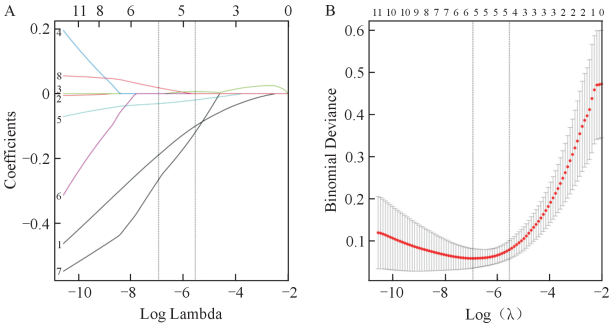



图1 LASSO回归变量筛选

Figure 1. LASSO regression variable screening

注：A：临床特征系数曲线；B：10折交叉验证对最优变量的筛选。

表2 多因素Logistic回归分析

Table 2. Multivariate Logistic regression analysis

| 影响因素 | $\beta$ | $SE$ | $Wald\ \chi^2$ | OR（95%CI） | $P$ |
|---|---|---|---|---|---|
| TBIL | 0.251 | 0.071 | 12.474 | 1.285（1.118，1.477） | <0.001 |
| ALB | −0.047 | 0.023 | 4.192 | 0.954（0.912，0.998） | 0.041 |
| Cr | 0.002 | 0.004 | 0.250 | 1.002（0.994，1.010） | 0.618 |
| CCr | −0.155 | 0.068 | 5.228 | 0.856（0.749，0.978） | 0.023 |
| CysC | 0.019 | 0.022 | 0.732 | 1.019（0.976，1.064） | 0.393 |

注：TBIL.总胆红素（total bilirubin）；ALB.白蛋白（albumin）；Cr.肌酐（creatinine）；CCr.肌酐清除率（creatinine clearance rate）；CysC.胱抑素C（cystatin C）。

表3 5种ML模型的性能评估比较

Table 3. Performance evaluation comparison of 5 ML models

| 模型 | 截断值 | AUC（95%CI） | 灵敏度 | 特异度 | 准确度 | F1评分 |
|---|---|---|---|---|---|---|
| Bootstrap验证前 | | | | | | |
| LR | 0.456 | 0.889（0.713，0.936） | 0.944 | 0.812 | 0.902 | 0.929 |
| RF | 0.453 | 0.802（0.722，0.893） | 0.900 | 0.730 | 0.832 | 0.866 |
| DT | 0.472 | 0.730（0.658，0.827） | 0.874 | 0.851 | 0.867 | 0.899 |
| XGBoost | 0.463 | 0.906（0.805，0.952） | 0.935 | 0.941 | 0.937 | 0.952 |
| SVM | 0.457 | 0.716（0.629，0.846） | 0.799 | 0.792 | 0.797 | 0.842 |
| Bootstrap验证后 | | | | | | |
| LR | 0.452 | 0.897（0.730，0.915） | 0.935 | 0.851 | 0.908 | 0.932 |
| RF | 0.422 | 0.783（0.672，0.853） | 0.846 | 0.842 | 0.844 | 0.881 |
| DT | 0.465 | 0.745（0.653，0.867） | 0.897 | 0.842 | 0.879 | 0.910 |
| XGBoost | 0.473 | 0.914（0.835，0.960） | 0.893 | 0.990 | 0.924 | 0.941 |
| SVM | 0.438 | 0.700（0.647，0.794） | 0.841 | 0.762 | 0.816 | 0.861 |

注：RF.随机森林（random forest）；DT.决策树（decision tree）；LR.传统逻辑回归（Logistic regression）；XGBoost.极端梯度提升（extreme gradient boosting）；SVM.支持向量机（support vector machine）。

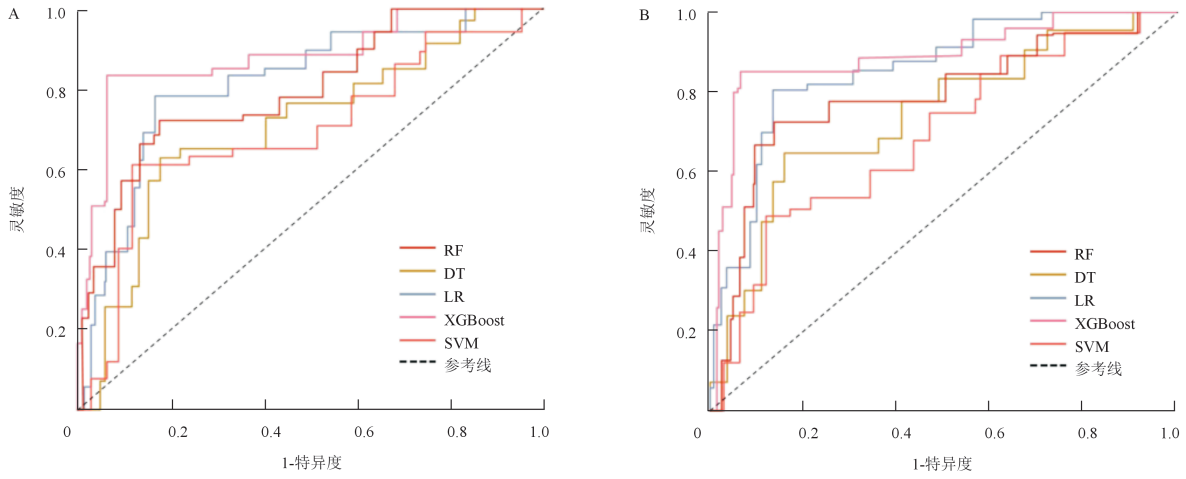



图2 ROC曲线

Figure 2. ROC Curve

注：A.Bootstrap验证前ROC曲线；B.Bootstrap验证后ROC曲线。

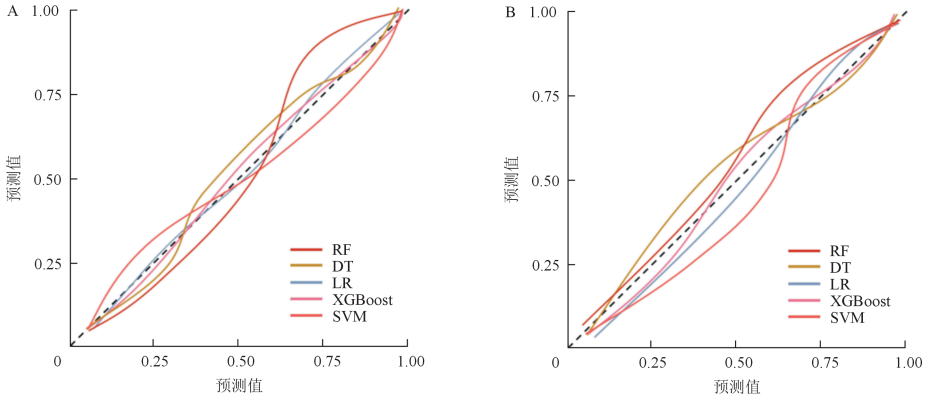



图3 校准曲线

Figure 3. Calibration Curve

注：A.Bootstrap验证前校准曲线；B.Bootstrap验证后校准曲线。

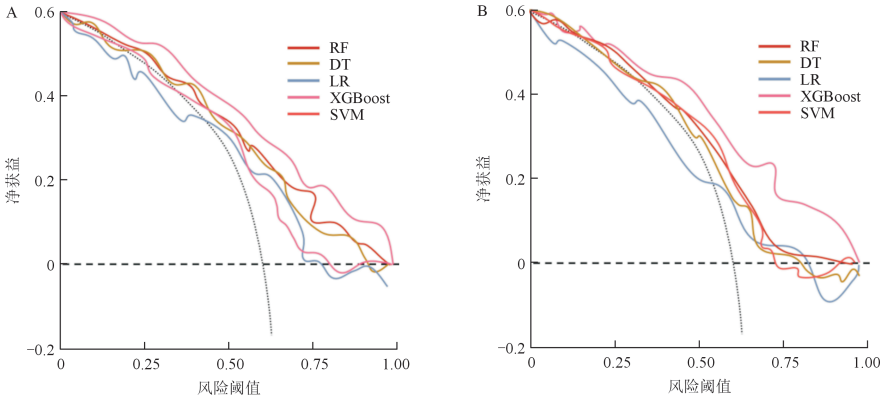



图4 DCA曲线

Figure 4. DCA Curve

注：A.Bootstrap验证前DCA曲线；B.Bootstrap验证后DCA曲线。

表4 5种ML模型的效能比较

Table 4. Performance comparison of 5 ML models

| 模型 | AUC（95%CI） | *P* | IDI（95%CI） | *P* | NRI（95%CI） | *P* |
|---|---|---|---|---|---|---|
| LR | 0.897（0.730，0.915） | | | | | |
| RF | 0.783（0.672，0.853） | 0.087 | −0.027（−0.045，0.011） | 0.062 | −0.204（−0.365，0.056） | 0.073 |
| DT | 0.745（0.653，0.867） | 0.048 | −0.059（−0.094，−0.033） | 0.037 | −0.311（−0.482，−0.200） | 0.034 |
| XGBoost | 0.914（0.835，0.960） | <0.001 | 0.089（0.042，0.126） | <0.001 | 0.493（0.327，0.579） | <0.001 |
| SVM | 0.700（0.647，0.794） | <0.001 | −0.124（−0.131，−0.095） | <0.001 | −0.516（−0.684，−0.410） | <0.001 |

注：RF.随机森林（random forest）；DT.决策树（decision tree）；LR.传统逻辑回归（Logistic regression）；XGBoost.极端梯度提升（extreme gradient boosting）；SVM.支持向量机（support vector machine）。

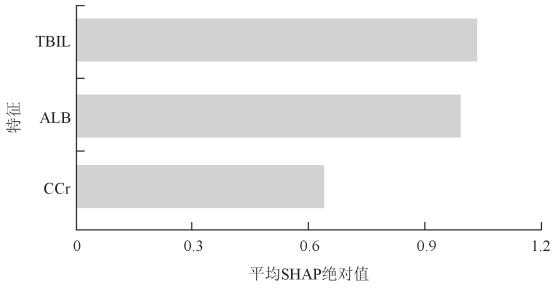



图5 XGBoost模型变量重要性排序条形图

Figure 5. Bar chart of the variable importance ranking for the XGBoost model

注：平均SHAP绝对值：对模型输出的平均影响。

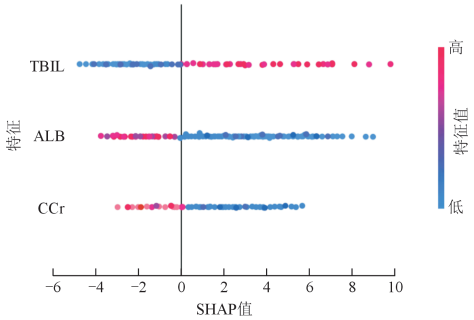



图6 XGBoost模型变量重要性排序蜂群图

Figure 6. Beehive chart of variable importance ranking for XGBoost model

## 3 讨论

血小板减少症是一类常见的血液系统疾病，其特征为血液中PLT低于正常水平，可由药物、感染、骨髓抑制或免疫介导性损伤及遗传等多种因素引发；该病症不仅会提高出血风险，诱发不同程度的出血相关并发症，甚至可能导致患者死亡[9]。当前，针对抗菌药物所致血小板减少症的预测模型研究，多以传统统计方法为基础；然而，抗菌药物引发血小板减少症的危险因素繁杂且可能相互关联，导致模型预测性能下滑，难以有效缩小预测值与实际值之间的偏差。随着大数据在临床研究领域的应用愈发广泛，ML技术也逐渐成为构建医院感染相关风险预测模型的关键手段[10]。因此，本研究通过对比5种算法的表现，构建并验证抗菌药物导致血小板减少症的风险预测模型，为高风险患者的早期识别及临床防控决策提供参考。

本研究基于小样本、阳性事件少、临床变量维度较高的现实条件，采用单因素分析、LASSO回归、多因素LR的多阶段变量筛选策略。当样本量有限且阳性结局较少时，若直接将全部候选变量纳入ML模型，易引发过拟合、特征冗余、模型泛化能力下降等问题；加之临床变量间常存在多重共线性，未经筛选直接建模会进一步降低模型稳定性与可解释性。其中，单因素分析可快速剔除无关噪声变量，LASSO回归能自动压缩系数并筛选关键特征，多因素LR则用于确认独立影响因素，三者联用可最大限度降低假阳性关联，保留具有真实生物学意义的预测因子，让ML模型更贴合临床实际、预测结果更稳健可靠。因此，该筛选策略符合临床预测模型规范，尤其适用于阳性事件率低、样本量有限的药物不良反应回顾性研究。与此同时，本研究中患病组与未患病组例数比约为1∶16，数据存在严重类别不平衡，易导致模型偏向多数类、降低对少数阳性样本的识别效能，出现整体准确率高、阳性预测差的偏差。已有研究显示[11-13]，在类似1∶1的极端不平衡数据中，经SMOTE-NC处理后模型预测效能可明显提升。据此，本研究采用SMOTE-NC技术对训练集阳性样本进行过采样，严格遵循先划分数据集、再采样的规范流程，避免数据泄漏，从而有效缓解类别不平衡对模型性能的影响，提升模型对抗菌药物致血小板减少症的识别能力。

XGBoost模型是一种基于梯度提升决策树的集成算法，其引入了正则化项和二阶导数信息，能够更精准地捕捉数据中的非线性关系及交互作用，在疾病诊断、药物响应预测、医疗资源优化等领域应用广泛[14]。有研究者[15]基于ML构建ICU老年患者呼吸机相关肺炎风险预测模型时，发现XGBoost模型为最优模型；另有相关研究结果表明[16]，在基于ML算法的肺癌四级胸腔镜手术后肺部感染风险预测模型中，XGBoost模型表现最优。本研究对5种ML模型进行内部验证后，XGBoost模型的AUC、准确度、灵敏度、特异度及F1分数均优于其他模型；Delong检验进一步证实XGBoost模型与其余各模型AUC差异具有统计学意义；以LR模型为参照，IDI与NRI结果显示XGBoost模型在综合判别能力与重新分类能力上均有明显提升，进一步从统计学层面支持其为最优模型。同时，校准曲线与DCA曲线显示该模型预测精准度较高、临床净获益最大，因此XGBoost模型为本研究中预测抗菌药物致血小板减少症发生风险的最优模型。XGBoost模型的变量重要性及SHAP分析结果显示，TBIL、ALB及CCr水平是影响抗菌药物致血小板减少症发生的重要因素。

SHAP分析揭示的CCr高重要性，表明了肾功能在抗菌药物致血小板减少症中的核心地位，这与国内外研究中“药物蓄积是触发血小板损伤的主要机制”的结论高度一致[17-18]。多数抗菌药物主要经肾脏排泄，肾小球滤过功能下降会导致药物清除延迟，游离药物浓度升高可能通过两种途径诱发血小板减少症：一是直接抑制骨髓巨核细胞增殖分化，二是作为半抗原与血小板膜蛋白结合，诱导抗体介导的免疫性破坏[19-20]。相关研究证实[21]，肾功能下降是血小板减少症的独立预测因素。肝脏作为药物代谢的主要器官，其功能状态直接影响抗菌药物的生物转化与清除效率。本研究中ALB和TBIL的SHAP高贡献度，揭示了肝功能障碍与抗菌药物致DITP的深层关联。ALB的重要性源于其作为药物结合载体与机体营养状态标志物的双重角色，这与“低蛋白血症加剧药物毒性”的研究[22]结论一致。相关研究表明[23]，肝病合并血小板减少症患者接受特比澳治疗后，

ALB水平上升与PLT恢复呈明显正相关。国外一项研究表明[24]，ALB是严重发热伴血小板减少症患者早期疾病的敏感且独立的预测因子。据此推测其或可预测抗菌药物导致的血小板减少症，本研究结果支持这一推测。此外，ALB作为营养状态指标，其水平低下常提示机体免疫功能与造血储备不足。有研究表明[25]，低ALB患者的骨髓造血微环境受损，巨核细胞对促血小板生成素的反应性下降，更易受抗菌药物的抑制。高胆红素血症提示肝细胞排泄功能障碍或胆汁淤积，可能影响脂溶性抗菌药物的排泄。TBIL升高反映肝细胞损伤或胆红素代谢障碍，而肝脏是多种抗菌药物的代谢场所。肝细胞损伤时，会导致血小板生成素分泌减少，削弱血小板生成代偿能力，从而增加血小板减少症发生风险[26]。

通过对患者TBIL、ALB及CCr水平进行密切监测与综合评估，临床医生能够制定更具个性化的抗菌药物治疗及风险监测方案。通过监测CCr水平，可提前采取措施，调整抗菌药物剂量并加强肾功能观察；针对ALB异常降低的患者，可强化营养支持并增加PLT监测频率，以便及时发现异常；对于TBIL异常升高者，可停用肝毒性药物或调整治疗方案，维持肝功能指标稳定。对这些关键指标进行动态监测，有助于提前实施干预，降低抗菌药物致血小板减少症的发生风险及相关出血并发症的出现概率。

本研究仍存在一定局限性。本研究为单中心回顾性研究，受单中心人群基线、诊疗习惯、抗菌药物使用规范的影响，模型外推性存在一定偏倚；本研究仅采用Bootstrap内部验证，缺乏独立多中心外部数据集对模型泛化性能的验证，模型现阶段仅适用于本院住院患者抗菌药物致血小板减少症的风险初筛，暂不推荐作为临床诊疗决策的唯一依据，不宜直接大范围推广；本研究抗菌药物类别间样本量分布不均衡，可能导致部分药物的风险效应未被充分识别；且本研究纳入的预测特征有限，未纳入患者合并基础疾病、抗菌药物血药浓度、用药时长等潜在影响因素。后续可开展多中心外部验证，纳入不同地域、不同基线特征的患者，全面评估模型的泛化性能与临床适用性；扩大样本量，纳入更多潜在预测特征，进一步优化模型预测效能；此外开展前瞻性队列研究，验证模型在真实临床场景中的预警价值。

综上所述，TBIL、ALB、CCr水平为抗菌药物致血小板减少症的影响因素。不同ML算法中，XGBoost模型对患者发生抗菌药物致血小板减少症具有较强的预测价值，可优先应用于住院患者，尤其是肝肾功能不全者的抗菌药物治疗风险评估，门诊患者需结合用药时长、随访便利性等因素调整阈值后使用。

**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 Kuwabara G, Tazoe K, Imoto W, et al. Isoniazid-induced immune thrombocytopenia[J]. Intern Med, 2021, 60(22): 3639-3643. DOI: 10.2169/internalmedicine.6520-20.

2 周国威, 张宏岩, 矫艳艳, 等. 基于FAERS数据库的药物相关血小板减少症信号挖掘[J]. 药物流行病学杂志, 2025, 34(12): 1382-1389. [Zhou GW, Zhang HY, Jiao YY, et al. Signal mining and analysis of drug-related thrombocytopenia based on FAERS database[J]. Chinese Journal of Pharmacoepidemiology, 2025, 34(12): 1382-1389.] DOI: 10.12173/j. issn. 1005-0698.202507057.

3 Kunyu L, Shuping S, Chang S, et al. An updated comprehensive pharmacovigilance study of drug-induced thrombocytopenia based on fda adverse event reporting system data[J]. J Clin Pharmacol, 2024, 64(4): 478-489. DOI: 10.1002/JCPH.2389.

4 李雪莲, 朱庆东, 马怡静, 等. 利奈唑胺血液系统不良反应发生率及危险因素分析: 一项多中心队列研究[J]. 中国防痨杂志, 2025, 47(6): 719-726. [Li XL, Zhu QD, Ma YJ, et al. Analysis of incidence and risk factorsfor linezolid-related hematological side effects: a multicenter cohort study[J]. Chinese Journal of Antituberculosi, 2025, 47(6): 719-726.] DOI: 10.19982/j.issn.1000-6621.20240539.

5 蔡乐, 汤智慧, 邱子涵, 等. 药源性血小板减少症的危险因素分析与风险预测模型的构建[J]. 临床药物治疗杂志, 2024, 22(8): 22-28. [Cai L, Tang ZH, Qiu ZH, et al. Analysis of risk factors and construction of a risk prediction model for drug-induced thrombocytopenia[J]. Clinical Medication Journal, 2024, 22(8): 22-28.] DOI: 10.3969/j.issn.1672-3384.2024.08.005.

6 Andrès E, El Hassani Hajjam A, Maloisel F, et al. Artificial intelligence (AI) and drug-induced and idiosyncratic cytopenia: the role of ai in prevention, prediction, and patient participation[J]. Hematol Rep, 2025, 17(3): 24. DOI: 10.3390/hematolrep17030024.

7 廖茹, 崔祎, 程晓亮, 等. 基于机器学习算法的利奈唑胺相关血小板减少预测[J]. 医药导报, 2025, 44(4): 676-681. [Liao R, Cui Y, Cheng XL, et al. Prediction of linezolid-induced thrombocytopenia based on machine learning algorithm[J]. Herald of Medicine, 2025, 44(4): 676-681.] DOI: 10.3870/j.issn.1004-0781.2025.04.028.

8 黄翠丽, 高奥, 王嘉熙, 等. 抗菌药物相关血小板减少症的自

动监测与评价研究[J]. 中国药物警戒, 2023, 20(7): 807-811. [Huang CL, Gao A, Wang JX, et al. Automatic monitoring and assessment of antibiotics-related thrombocytopenia[J]. Chinese Journal of Pharmacovigilance, 2023, 20(7): 807-811.] DOI: 10.19803/j.1672-8629.20220348.

9 Nietsch KS, Roach TM, Wilson ZD, et al. Principles and considerations in the early identification and prehospital treatment of thrombocytopenia[J]. J Spec Oper Med, 2022, 22(2): 75-79. DOI: 10.55460/333T-XIYF.

10 Li J, Yan Z. Machine learning model predicting factors for incisional infection following right hemicolectomy for colon cancer[J]. BMC Surg, 2024, 24(1): 279. DOI: 10.1186/s12893-024-02543-8.

11 王梅英, 杨敏, 刘佳微, 等. 基于SMOTE算法的化疗肿瘤患者下呼吸道感染预警模型构建[J]. 中国感染控制杂志, 2021, 20(12): 1094-1101. [Wang MY, Yang M, Liu JW, et al. Construction of early warning model of lower respiratory tract infection in chemotherapy tumor patients based on SMOTE algorithm[J]. Chinese Journal of Infection Control, 2021, 20(12): 1094-1101.] DOI: 10.12138/j.issn.1671-9638.20211135.

12 Yuan H, Xu H. Deep multi-modal fusion network with gated unit for breast cancer survival prediction[J]. Comput Methods Biomech Biomed Engin, 2024, 27(7): 883-896. DOI: 10.1080/10255842.2023.2211188.

13 Gozukara Bag HG, Yagin FH, Gormez Y, et al. Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits[J]. Diagnostics (Basel), 2023, 13(18): 2949. DOI: 10.3390/diagnostics13182949.

14 Liu HQ, Lin SY, Song YD, et al. Machine learning on MRI radiomic features: identification of molecular subtype alteration in breast cancer after neoadjuvant therapy[J]. Eur Radiol, 2023, 33(4): 2965-2974. DOI: 10.1007/S00330-022-09264-7.

15 甘文思, 黄一睿, 王笑青, 等. 基于机器学习的ICU老年患者呼吸机相关肺炎风险预测模型的构建及评价[J]. 中华医院感染学杂志, 2025, 35(2): 290-296. [Gan WS, Huang YR, Wang XQ, et al. Establishment of risk prediction model for ventilator-associated pneumoniain elderly ICU patients based on machine learning and its performance[J]. Chinese Journal of Nosocomiology, 2025, 35(2): 290-296.] DOI: 10.11816/cn.ni.2025-240990.

16 马佳佳, 刘晓芯, 薛蓓, 等. 基于机器学习算法的肺癌四级胸腔镜手术后肺部感染风险预测模型构建[J]. 实用临床医药杂志, 2025, 29(6): 111-117. [Ma JJ, Liu XX, Xue B, et al. Risk prediction model construction of postoperativepulmonary infection in lung cancer patients undergoingfour-level thoracoscopic surgery based on machinelearning algorithms[J]. Journal of Clinical Medicine in Practice, 2025, 29(6): 111-117.] DOI: 10.7619/jcmp.20245679.

17 Shi C, Xia J, Ye J, et al. Effect of renal function on the risk of thrombocytopaenia in patients receiving linezolid therapy: a systematic review and meta-analysis[J]. Br J Clin Pharmacol, 2022, 88(2): 464-475. DOI: 10.1111/bcp.14965.

18 Gou J, Li Q, Fan N, et al. High accumulation of linezolid and its major metabolite in the serum of patients with hepatic and renal dysfunction is significantly associated with thrombocytopenia and anemia[J]. Microbiol Spectr, 2025, 13(7): 0249324. DOI: 10.1128/spectrum.02493-24.

19 Liu Y, Huang J, Li L, et al. Regulatory effect of PDGF/PDGFR on hematopoiesis[J]. Semin Thromb Hemost, 2025, 51(5): 572-577. DOI: 10.1055/S-0044-1796630.

20 Buka RJ, Montague SJ, Moran LA, et al. PF4 activates the c-Mpl-Jak2 pathway in platelets[J]. Blood, 2024, 143(1): 64-69. DOI: 10.1182/blood.2023020872.

21 马勇, 高伟波, 朱继红. 细菌性肝脓肿患者发生血小板减少影响因素研究[J]. 中国全科医学, 2023, 26(17): 2120-2124. [Ma Y, Gao WB, Zhu JH. Risk factors of thrombocytopenia caused by pyogenic liver abscess[J]. Chinese General Practice, 2023, 26(17): 2120-2124.] DOI: 10.12114/j.issn.1007-9572.2022.0742.

22 赵资德, 吴令杰, 张海生, 等. 纠正低蛋白血症减少抗结核药物性肝损伤发生临床研究[J]. 实用肝脏病杂志, 2022, 25(6): 792-795. [Zhao ZD, Wu LJ, Zhang HS, et al. Could the supplement of human blood albumin product reduce the incidence of drug-induced liver injury in patients with pulmonary tuberculosis and hypoproteinemia[J]. Journal of Practical Hepatology, 2022, 25(6): 792-795.] DOI: 10.3969/j.issn.1672-5069.2022.06.009.

23 崔大广, 肖玲燕, 史东阳, 等. 特比澳对肝病合并血小板减少症患者的疗效分析[J]. 肝脏, 2022, 27(12): 1335-1339. [Cui DG, Xiao LY, Shi DY, et al. Analysis of application of recombinant human thrombopoietin in patients with liver disease complicated with thrombocytopenia[J]. Chinese Hepatology, 2022, 27(12): 1335-1339.] DOI: 10.3969/j.issn.1008-1704.2022.12.022.

24 Hao Y, Sun J, Wang X, et al. Difference in hematocrit and plasma albumin levels as an early biomarker of severity and prognosis in patients with severe fever and thrombocytopenia syndrome[J]. J Med Virol, 2024, 96(10): 29941. DOI: 10.1002/jmv.29941.

25 Chen C, Li Y, Yu J, et al. Linezolid-induced thrombocytopenia in patients with acute myeloid leukemia: a matched case-control study[J]. Clin Transl Oncol, 2022, 24(3): 540-545. DOI: 10.1007/s12094-021-02711-9.

26 Lim HI, Cuker A. Thrombocytopenia and liver disease: pathophysiology and periprocedural management[J]. Hematology Am Soc Hematol Educ Program, 2022, 2022(1): 296-302. DOI: 10.1182/hematology.2022000408.

收稿日期：2026年01月15日 修回日期：2026年06月05日
本文编辑：桂裕亮 李绪辉