Journal of Blood Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Development and Validation of an Interpretable Machine Learning Model for Predicting Thrombocythemia Risk During Third Generation Cephalosporin Therapy

Kailei Du[1], Maofeng Wang[2], Ping Yu[3]

[1]Intensive Care Medicine, Affiliated Dongyang Hospital, Wenzhou Medical University, Dongyang, Zhejiang, 322100, People's Republic of China; [2]Department of Biomedical Sciences Laboratory, Affiliated Dongyang Hospital, Wenzhou Medical University, Dongyang, Zhejiang, 322100, People's Republic of China; [3]Department of Gynecology, Affiliated Dongyang Hospital, Wenzhou Medical University, Dongyang, Zhejiang, 322100, People's Republic of China

Correspondence: Maofeng Wang; Ping Yu, Email wzmcwmf@wmu.edu.cn; 13505897866@163.com

Objective: Third-generation cephalosporins are widely used for severe infections but carry thrombocythemia risks complicating therapeutic decisions. Current predictive tools lack accuracy and clinical interpretability. This study aimed to develop an interpretable machine learning (ML) model for thrombocythemia risk stratification during cephalosporin therapy.

Methods: A retrospective cohort of 45,779 adults treated with third-generation cephalosporins (2019–2023) was analyzed. After exclusions (age <18, missing data, baseline platelet anomalies), 25,707 patients were included. Thrombocythemia was defined as platelet count $>400\times10^9$/L within 30 days post-treatment. Predictors encompassed demographics, comorbidities, medications, and laboratory parameters. Data preprocessing included multiple imputation and stratified partitioning (70% training, 30% testing). Three ML algorithms (XGBoost, Random Forest, LightGBM) were evaluated using ROC-AUC, Brier score, and clinical utility metrics. SHAP analysis provided model interpretability.

Results: XGBoost demonstrated superior performance, achieving the highest test-set discrimination (AUC=0.858, 95% CI:0.814–0.902) and calibration (Brier score=0.0088). SHAP analysis identified Baseline platelet count (PLT), red blood cell count (RBC), creatinine (CRE), daily usage frequency, and sex as key drivers. PLT was the strongest predictor (SHAP range: −1.67 to +1.48), with lower PLT exerting protective effects. RBC and CRE ranked second and third in importance, showing nonlinear risk relationships. Key clinical interactions included amplified risk from malignancies (SHAP=−0.215) and protective effects of female sex (SHAP=−0.194).

Conclusion: This interpretable ML framework enables precise thrombocythemia risk prediction during cephalosporin therapy, balancing algorithmic performance with clinical actionability.

Keywords: thrombocythemia, risk prediction, machine learning, third-generation cephalosporin, XGBoost

Introduction

Third-generation cephalosporins can trigger thrombocytosis via immune responses, infection-driven inflammation, or bone marrow activation during treatment.[1,2] Though usually self-limiting, persistent cases risk thrombosis, complicating clinical management.[3] Therefore, when treating patients with third-generation cephalosporins, it is necessary to closely monitor platelet levels and adjust the treatment regimen as needed to reduce associated risks.[4]

Third generation cephalosporins are associated with multisystem adverse effects requiring clinical vigilance.[5] Severe cutaneous adverse reactions, including life-threatening conditions such as Stevens-Johnson syndrome, have been reported, emphasizing the need for prompt recognition and drug discontinuation.[6,7]

https://doi.org/10.2147/JBM.S567773
Received: 15 September 2025
Accepted: 8 January 2026
Published: 11 February 2026

Thrombocytosis is a adverse reaction cause by third-generation cephalosporins. Researchers found that the widespread use of third generation cephalosporins might lead to thrombocytosis in some patients, especially with long-term or high-dose use.[8] In addition, another study examined the effectiveness of third generation cephalosporins in treating severe infections. The results showed that while these drugs have significant efficacy in combating infections, they can also lead to adverse hematological reactions, including thrombocytosis.[9] Researchers also found that the use of third-generation cephalosporins is associated with an increase in certain bacterial resistances, which may further complicate treatment processes and potentially indirectly affect changes in platelet counts.[10] Studies show that while third-generation cephalosporins are effective against Salmonella infections, their use may be associated with an increased risk of thrombocytosis, especially in pediatric patients.[11] To sum up, although the third generation cephalosporins have good therapeutic effect in treating bacterial infections, they may also cause hematological adverse reactions such as thrombocytosis. Therefore, in clinical application, it is necessary to closely monitor the hematological indicators of patients to ensure the safety and effectiveness of drug use.[12] Therefore, it is crucial to establish an effective predictive model to assess the risk of thrombocytosis caused by third-generation cephalosporin therapy. Predictive models[13,14] can integrate factors such as patient's clinical characteristics, past medication history, and laboratory test results to help doctors better assess the risk of thrombocytosis when using third-generation cephalosporins, thereby optimizing treatment plans and reducing the occurrence of adverse reactions.

To bridge this clinical gap, we engineered an interpretable machine learning framework capable of providing thrombocythemia risk stratification for patients undergoing cephalosporin treatment.

Methods

Study Population

This retrospective cohort study analyzed anonymized electronic medical records from 45,779 adults who received third-generation cephalosporin therapy at Affiliated Dongyang Hospital of Wenzhou Medical University (January 2019-December 2023). After applying exclusion criteria: age <18 years (n=7,858), absent baseline (n=3,684) and follow-up platelet counts (n=3,492), post-treatment platelet count < 100×10^9/L (n=1,622), baseline thrombocytopenia (<100×10^9/L; n=2,598) and thrombocytosis (>400×10^9/L; n=818) – the final cohort comprised 25,707 patients. The requirement for informed consent was formally waived by Affiliated Dongyang Hospital of Wenzhou Medical University Institutional Review Board (No. 2025-YX-012), because this retrospective study used only anonymized data and involved no more than minimal risk to participants. This waiver is consistent with the principles of the Declaration of Helsinki.

Outcome Definition

Thrombocythemia[15] was defined as platelet count >400×10^9/L occurring within 30 days following the initiation of third-generation cephalosporin therapy in hospitalized patients. Cases were identified through structured medication classification codes in the hospital's EMRs system.

Risk Factors

Variable selection adhered to pharmacoepidemiologic principles encompassing biological plausibility, clinical relevance to antimicrobial stewardship protocols, and availability of quantifiable measures meeting CLIA-certified laboratory standards. We extracted de-identified patient data from institutional electronic medical records (EMRs).(1) Demographics: Sex and age; (2) Clinical History: Lifestyle factors (tobacco use, alcohol consumption); (3) Comorbidities: malignancies, diabetes mellitus, hypertension, cerebrovascular/cardiovascular events; (4) Organ dysfunction (hepatic cirrhosis, renal insufficiency); (5) Medication exposures (anticoagulants, systemic antifungals); (6) Laboratory Parameters: Hematologic indices: white blood cell count (WBC), red blood cell count (RBC), platelet count (PLT), serum albumin, serum lactate.Temporal specification: Lowest recorded levels within 30 days pre-cephalosporin initiation. All baseline risk factors were ascertained prior to the initiation of antibiotic therapy, while data on antibiotic use were collected through the point of hospital discharge.This structured approach minimized recall bias while ensuring data comparability across the cohort.

Data Pre-Processing

The analysis employed a two-phase preprocessing strategy. In the initial quality control phase, systematic completeness assessment of 29 candidate covariates revealed 8 predictors exceeding predefined missingness thresholds, which were subsequently excluded via preanalysis exclusion protocol. For variables with acceptable missing data (<20% prevalence), multiple imputation was conducted using chained equations (MICE algorithm; 5 imputation cycles, predictive mean matching).[16,17] Model validation employed stratified partitioning (70% training; 30% testing) with outcome frequency preservation across subsets. This dual-stage approach complied with STROBE guidelines for observational data handling while mitigating selection bias through probabilistic imputation and rigorous validation design.

Model Building

Three ensemble algorithms were selected for their complementary strengths in handling high-dimensional biomedical data:[18,19] XGBoost leveraged histogram-optimized gradient boosting with parallel tree construction and L2 regularization to balance sparsity adaptation and overfitting prevention. Random Forest employed bootstrap-aggregated decision trees with dual randomization to enhance prediction stability through consensus voting. LightGBM prioritized computational efficiency via leaf-wise growth with depth constraints and exclusive feature bundling, achieving 5–10× acceleration over conventional gradient boosting methods in pilot benchmarks.

Model Evaluation

The validation framework employed tripartite metric evaluation aligned with TRIPOD guidelines: (1) Discrimination analysis combined ROC-AUC[20] with PR-AUC metrics to address spectrum bias in imbalanced outcomes, (2) Calibration fidelity was quantified through Brier score decomposition,[21] (3) Clinical utility metrics incorporated PPV/NPV thresholds and F1 scores to reflect operational risk-benefit ratios. Model optimization followed a hierarchical approach: maximizing the ROC-AUC (to prioritize discrimination) was followed by minimizing the Brier score (to achieve probabilistic calibration), as outlined in Figure 1.

Model Interpretation

The model interpretability framework incorporated two orthogonal techniques aligned with TRIPOD-AI explainability standards: 1) Global feature importance was quantified through permutation-based error reduction analysis, 2) Individualized attribution utilized SHAP values derived from cooperative game theory to decompose prediction-level contributions.

Statistical Methods

Statistical analysis and data visualization were conducted using R4.4.2 software on Windows. Continuous variables were presented as means with standard deviations or medians with interquartile ranges and compared using appropriate tests such as Student's *t*-test or Mann–Whitney *U*-test. Categorical variables were expressed as frequencies with percentages and compared using χ^2-test or Fisher's exact test. Multiple imputation techniques were implemented using the "mice" package. Baseline description and differences analysis utilized the "comparegroups" package. Discrimination analysis involved the use of packages such as "pROC", "ggROC", and "fbroc". Calibration assessment was performed using the "rms" and "riskregression" packages. Models such as XGBoost, RF, and GBM were developed and assessed with the R package "mlr3". For model interpretation, the R packages "fastshap" "and" "pdp" " were employed". Additional packages included " xgboost", "random Forest", " gbm", "ggplot2", and their dependencies. All statistical tests were two-sided, and a significance level of $p < 0.05$ was considered statistically significant.

Results

Study Population Characteristics

The study initially screened 45,779 patients treated between January 2019 and December 2023. After applying exclusion criteria, 25,707 subjects were included in the final analysis (Figure 1). The cohort demonstrated balanced demographic

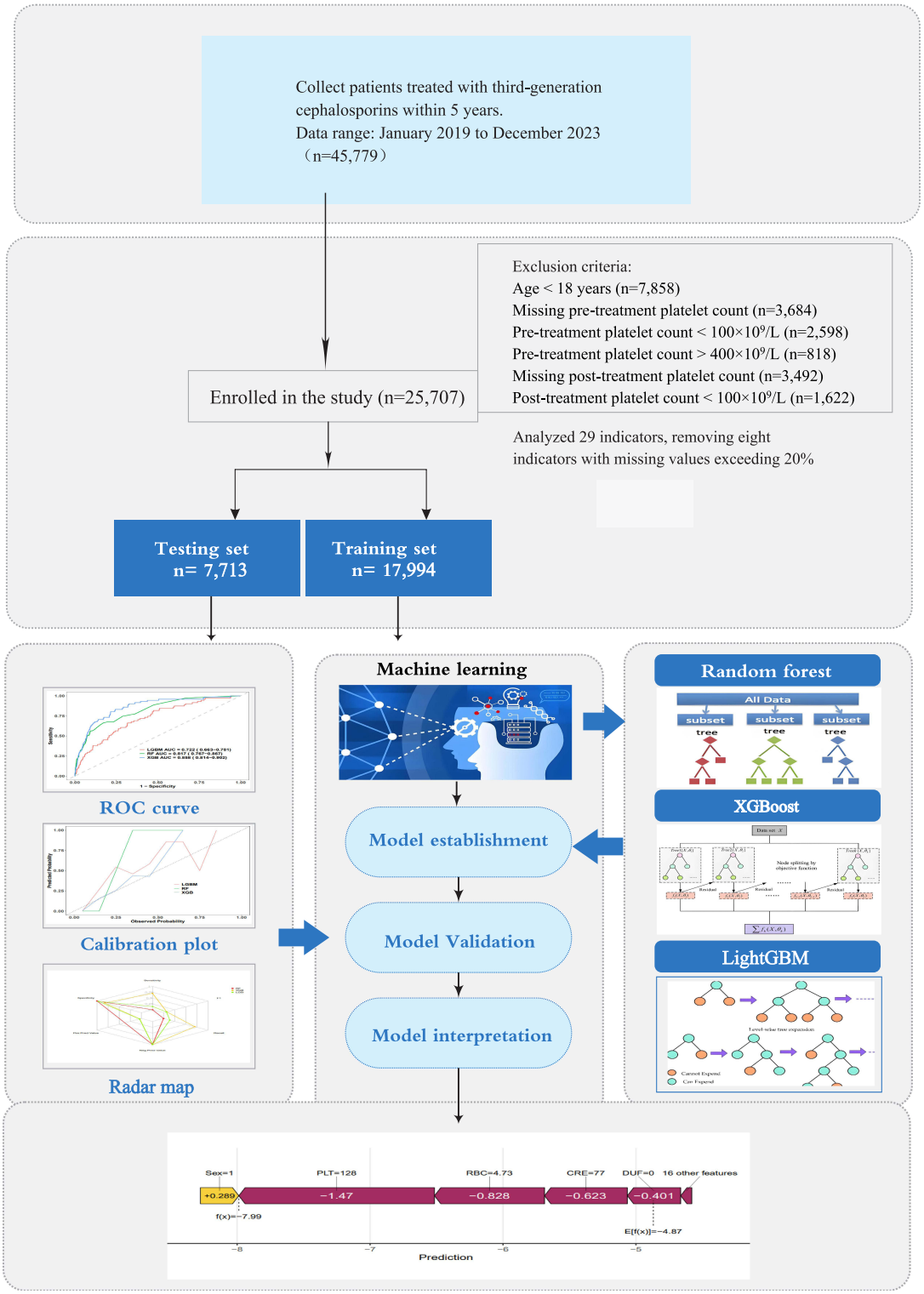

Figure 1 Study process flowchart.

characteristics, with 56.8% males and a median age of 64 years (IQR:49–76), showing no significant intergroup differences in gender distribution (p=0.321) or age (p=0.118) between thrombocythemia (n=245) and non-thrombocythemia groups (n=25,462). Clinically significant differences emerged in laboratory parameters: the thrombocythemia group exhibited elevated platelet counts (median 300 vs 206×10^9/L, p<0.001) and white blood cells (8.81 vs

7.76×10^9/L, p=0.002), but lower creatinine (55.0 vs 59.0 μmol/L, p<0.001) and red blood cell counts (3.85 vs 4.28×10^{12}/L, p<0.001). Medication patterns revealed higher meropenem (8.16% vs 3.91%, p=0.001) and ofloxacin (20.41% vs 12.78%, p=0.001) usage in thrombocythemia patients. Table 1 presents the baseline characteristics of our hospitalized patients.

The dataset was partitioned into training (n=17,994) and testing (n=7,713) sets, with rigorous validation confirming comparable distributions of all baseline characteristics (p>0.05 for sex, age, laboratory values, and comorbidity profiles), ensuring model generalizability. Table 2 displays the baseline characteristics of patients in training and testing sets.

Table 1 Baseline Characteristics of Subjects

Variables	Total N=25707	No Thrombocythemia N=25462	Thrombocythemia N=245	p
Sex				0.321
Female	11105 (43.20%)	10,991 (43.17%)	114 (46.53%)	
Male	14602 (56.80%)	14,471 (56.83%)	131 (53.47%)	
Age (years)	64.0 [49.0;76.0]	64.0 [49.0;76.0]	61.0 [49.0;74.0]	0.118
PLT (10^9/L)	206 [167;252]	206 [166;251]	300 [229;357]	<0.001
CRE (μmol/L)	59.0 [49.0;72.0]	59.0 [49.0;72.0]	55.0[46.0;67.0]	<0.001
WBC (10^9/L)	7.76 [5.65;11.20]	7.75 [5.65;11.19]	8.81[6.33;12.01]	0.002
RBC (10^{12}/L)	4.28 [3.80;4.72]	4.28 [3.81;4.73]	3.85 [3.46;4.31]	<0.001
Meropenem, n (%)				0.001
No	24692 (96.05%)	24,467 (96.09%)	225 (91.84%)	
Yes	1015 (3.95%)	995 (3.91%)	20 (8.16%)	
Ofloxacin, n (%)				0.001
No	22403 (87.15%)	22,208 (87.22%)	195 (79.59%)	
Yes	3304 (12.85%)	3254 (12.78%)	50 (20.41%)	
Smoke, n (%)				0.744
No	700 (2.72%)	692 (2.72%)	8 (3.27%)	
Yes	25007 (97.28%)	24,770 (97.28%)	237 (96.73%)	
Drink, n (%)				0.744
No	700 (2.72%)	692 (2.72%)	8 (3.27%)	
Yes	25007 (97.28%)	24,770 (97.28%)	237 (96.73%)	
DM, n (%)				0.072
No	21763 (84.66%)	21,545 (84.62%)	218 (88.98%)	
Yes	3944 (15.34%)	3917 (15.38%)	27 (11.02%)	
Hypertension, n (%)				0.341
No	15236 (59.27%)	15,083 (59.24%)	153 (62.45%)	
Yes	10471 (40.73%)	10,379 (40.76%)	92 (37.55%)	
Tumor, n (%)				1.000
No	20239 (78.73%)	20,046 (78.73%)	193 (78.78%)	
Yes	5468 (21.27%)	5416 (21.27%)	52 (21.22%)	
MI, n (%)				0.633
No	25118 (97.71%)	24,877 (97.70%)	241 (98.37%)	
Yes	589 (2.29%)	585 (2.30%)	4 (1.63%)	
CI, n (%)				0.319
No	21150 (82.27%)	20,942 (82.25%)	208 (84.90%)	
Yes	4557 (17.73%)	4520 (17.75%)	37 (15.10%)	
Anticoagulants, n (%)				0.347
No	12505 (48.64%)	12,378 (48.61%)	127 (51.84%)	
Yes	13202 (51.36%)	13,084 (51.39%)	118 (48.16%)	

(*Continued*)

https://doi.org/10.2147/JBM.S567773

Table 1 (Continued).

Variables	Total N=25707	No Thrombocythemia N=25462	Thrombocythemia N=245	p
RI, n (%)				0.588
No	23987 (93.31%)	23,761 (93.32%)	226 (92.24%)	
Yes	1720 (6.69%)	1701 (6.68%)	19 (7.76%)	
LC, n (%)				0.448
No	25283 (98.35%)	25,040 (98.34%)	243 (99.18%)	
Yes	424 (1.65%)	422 (1.66%)	2 (0.82%)	
AD, n (%)				0.149
No	25217 (98.09%)	24,980 (98.11%)	237 (96.73%)	
Yes	490 (1.91%)	482 (1.89%)	8 (3.27%)	
DUF (n)	1.00 [0.00;3.00]	1.00 [0.00;3.00]	1.00 [1.00;3.00]	<0.001
Duration (day)	3.93 [0.01;6.50]	3.93 [0.01;6.50]	4.00 [1.06;6.27]	0.325

Abbreviations: PLT, platelet count; CRE, creatinine; WBC, White blood cell count; RBC, Red blood cell count; DM, Diabetes Mellitus; MI, Myocardial infarction; CI, Cerebral infarction; RI, renal insufficiency; LC, liver cirrhosis; AD, Antifungal drugs; DUF, Daily usage frequency.

Table 2 The Baseline Characteristics of the Training and Testing Set

Variables	Total N=25707	Testing N=7713	Training N=17994	p
Sex				0.882
Female	11105 (43.2%)	3326 (43.1%)	7779 (43.2%)	
Male	14602 (56.8%)	4387 (56.9%)	10,215 (56.8%)	
Age (years)	64.0 [49.0;76.0]	65.0 [49.0;76.0]	64.0 [49.0;76.0]	0.388
PLT (10^9/L)	206 [167;252]	206 [166;252]	206 [167;252]	0.677
CRE (μmol/L)	59.0 [49.0;72.0]	60.0 [49.0;73.0]	59.0 [49.0;72.0]	0.318
WBC (10^9/L)	7.8 [5.7;11.2]	7.9 [5.7;11.1]	7.7 [5.6;11.2]	0.247
RBC (10^{12}/L)	4.3 [3.8;4.7]	4.3 [3.8;4.7]	4.3 [3.8;4.7]	0.497
Meropenem, n (%)				0.578
No	24692 (96.1%)	7400 (95.9%)	17,292 (96.1%)	
Yes	1015 (3.9%)	313 (4.1%)	702 (3.9%)	
Ofloxacin, n (%)				0.325
No	22403 (87.1%)	6697 (86.8%)	15,706 (87.3%)	
Yes	3304 (12.9%)	1016 (13.2%)	2288 (12.7%)	
Smoke, n (%)				0.647
No	700 (2.7%)	216 (2.8%)	484 (2.7%)	
Yes	25007 (97.3%)	7497 (97.2%)	17,510 (97.3%)	
Drink, n (%)				0.647
No	700 (2.7%)	216 (2.8%)	484 (2.7%)	
Yes	25007 (97.3%)	7497 (97.2%)	17,510 (97.3%)	
DM, n (%)				0.184
No	21763 (84.7%)	6494 (84.2%)	15,269 (84.9%)	
Yes	3944 (15.3%)	1219 (15.8%)	2725 (15.1%)	
Hypertension, n (%)				0.378
No	15236 (59.3%)	4539 (58.8%)	10,697 (59.4%)	
Yes	10471 (40.7%)	3174 (41.2%)	7297 (40.6%)	
Tumor, n (%)				0.053
No	20239 (78.7%)	6131 (79.5%)	14,108 (78.4%)	
Yes	5468 (21.3%)	1582 (20.5%)	3886 (21.6%)	

(*Continued*)

Table 2 (Continued).

Variables	Total N=25707	Testing N=7713	Training N=17994	p
MI, n (%)				0.731
No	25118 (97.7%)	7532 (97.7%)	17,586 (97.7%)	
Yes	589 (2.3%)	181 (2.3%)	408 (2.3%)	
CI, n (%)				0.965
No	21150 (82.3%)	6344 (82.3%)	14,806 (82.3%)	
Yes	4557 (17.7%)	1369 (17.7%)	3188 (17.7%)	
Anticoagulants, n (%)				0.576
No	12505 (48.6%)	3773 (48.9%)	8732 (48.5%)	
Yes	13202 (51.4%)	3940 (51.1%)	9262 (51.5%)	
RI, n (%)				0.762
No	23987 (93.3%)	7203 (93.4%)	16,784 (93.3%)	
Yes	1720 (6.7%)	510 (6.6%)	1210 (6.7%)	
LC, n (%)				1.000
No	25283 (98.4%)	7586 (98.4%)	17,697 (98.3%)	
Yes	424 (1.6%)	127 (1.6%)	297 (1.7%)	
AD, n (%)				0.295
No	25217 (98.1%)	7577 (98.2%)	17,640 (98.0%)	
Yes	490 (1.9%)	136 (1.8%)	354 (2.0%)	
DUF (n)	1.0 [0.0;3.0]	1.0 [0.0;3.0]	1.0 [0.0;3.0]	0.581
Duration (day)	3.9 [<0.1;6.5]	3.9 [<0.1;6.3]	3.9 [<0.1;6.6]	0.463

Abbreviations: PLT, platelet count; CRE, creatinine; WBC, White blood cell count; RBC, Red blood cell count; DM, Diabetes Mellitus; MI, Myocardial infarction; CI, Cerebral infarction; RI, renal insufficiency; LC, liver cirrhosis; AD, Antifungal drugs; DUF, Daily usage frequency.

Risk Factors Screening for Thrombocythemia

The variable importance analysis across three machine learning models (LightGBM, Random Forest, and XGBoost) revealed both consistent and model-specific risk factors associated with thrombocythemia (Figure 2). Baseline PLT emerged as the most critical predictor in both random forest (Figure 2A) and XGBoost (Figure 2B) models, ranking first in importance, but less emphasized by LightGBM (sixth) (Figure 2C). Age and RBC were consistently identified as high-impact factors in LightGBM (ranked first and second). Notably, Baseline PLT, RBC, CRE, WBC, treatment duration maintained top-five rankings in both random forest and XGBoost models. Surprisingly, Baseline PLT, RBC, CRE, WBC, age, daily usage frequency and treatment duration were consistently ranked as the most importance variables across all models.

Model Development and Optimal Algorithm Selection

In the process of model development, three machine learning algorithms—LightGBM, random forest, and XGBoost—were rigorously evaluated based on their performance across multiple classification metrics. AUC-ROC analysis (as shown in Figure 3A) demonstrated that XGBoost achieved the highest discrimination capability (highest AUC value), followed by random forest and LightGBM, though all models exhibited strong predictive power. Further evaluation using comprehensive classification metrics (Figure 3B) revealed nuanced differences among the algorithms. XGBoost achieved the highest sensitivity, indicating its robustness in identifying true positive cases, which is critical for clinical applications requiring minimal false negatives. In contrast, LightGBM and random forest excelled in specificity, suggesting its utility in scenarios where avoiding false positives is prioritized. Calibration analysis further confirmed XGBoost's superior reliability, as evidenced by its lower Brier score (0.0088) compared to LightGBM (0.0101) and random forest (0.0089).

https://doi.org/10.2147/JBM.S567773

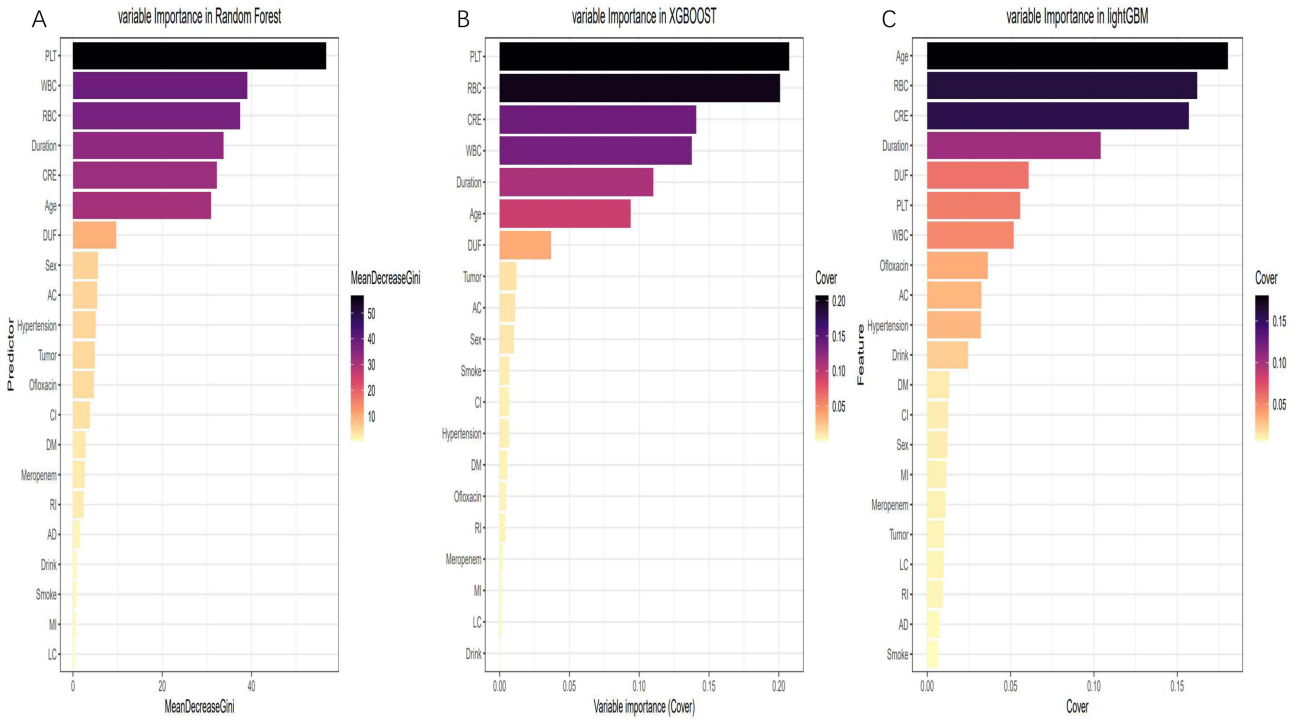

Figure 2 Machine learning-driven variable importance analysis for clinical outcome prediction (**A**) Random Forest importance measured by Mean Decrease Gini, highlighting, baseline PLT, WBC, RBC, duration and CRE as dominant factors. (**B**) XGBoost analysis with SHAP value-weighted importance, identifying baseline PLT, RBC, CRE, WBC, and duration as dominant factors. (**C**) LightGBM-based importance ranking using split frequency metrics. Top predictors include age, RBC, CRE, duration and DUF.

Abbreviations: PLT, platelet count; CRE, creatinine; WBC, White blood cell count; RBC, Red blood cell count; DM, Diabetes Mellitus; MI, Myocardial infarction; CI, Cerebral infarction; RI, renal insufficiency; LC, liver cirrhosis; AD, Antifungal drugs; DUF, Daily usage frequency; AC, Anticoagulants.

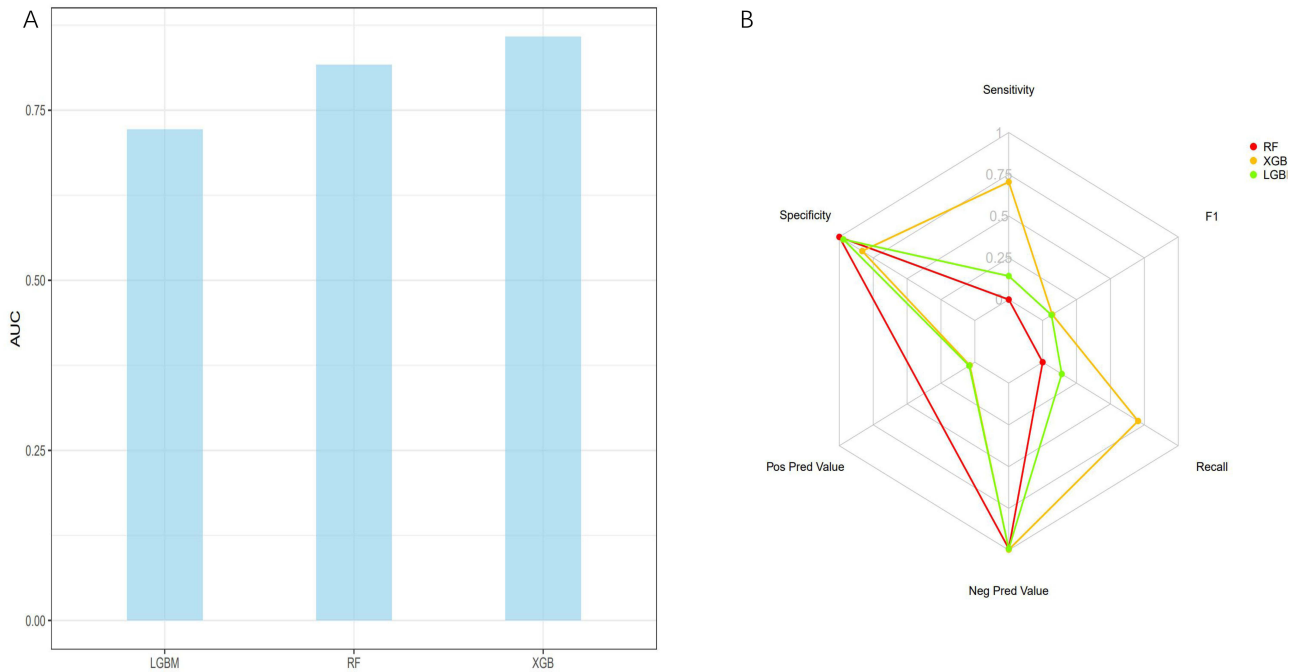

Figure 3 Comparative performance evaluation of machine learning models across classification metrics. (**A**) Comparative analysis of AUC scores among three models. (**B**) Performance metrics comparison among three models. Bar plots illustrate key classification metrics including Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Recall, and F1 Score.

Abbreviations: XGB, XGBoost; RF, Random Forest; LGBM, LightGBM.

https://doi.org/10.2147/JBM.S567773

This comprehensive profile combining high sensitivity, robust calibration, and exceptional discriminatory power—establishes XGBoost as the most clinically actionable model for risk stratification during cephalosporin therapy.

Model Evaluation

The model exhibited strong discrimination in both the training and testing cohorts (Figure 4A and B, respectively). Furthermore, its calibration performance was also excellent across these cohorts (Figure 4C and D, respectively). In the test set, XGBoost achieved the highest discriminative performance with an AUC of 0.858 (0.814−0.902), marginally outperforming Random Forest (AUC= 0.817, 0.767−0.867) and LightGBM (AUC= 0.722, 0.663−0.781). Calibration analysis revealed superior reliability for XGBoost, with its predicted probabilities closely aligned with observed outcomes (Brier score=0.0088). Notably, while Random Forest showed perfect training performance (AUC=1.000), its test-

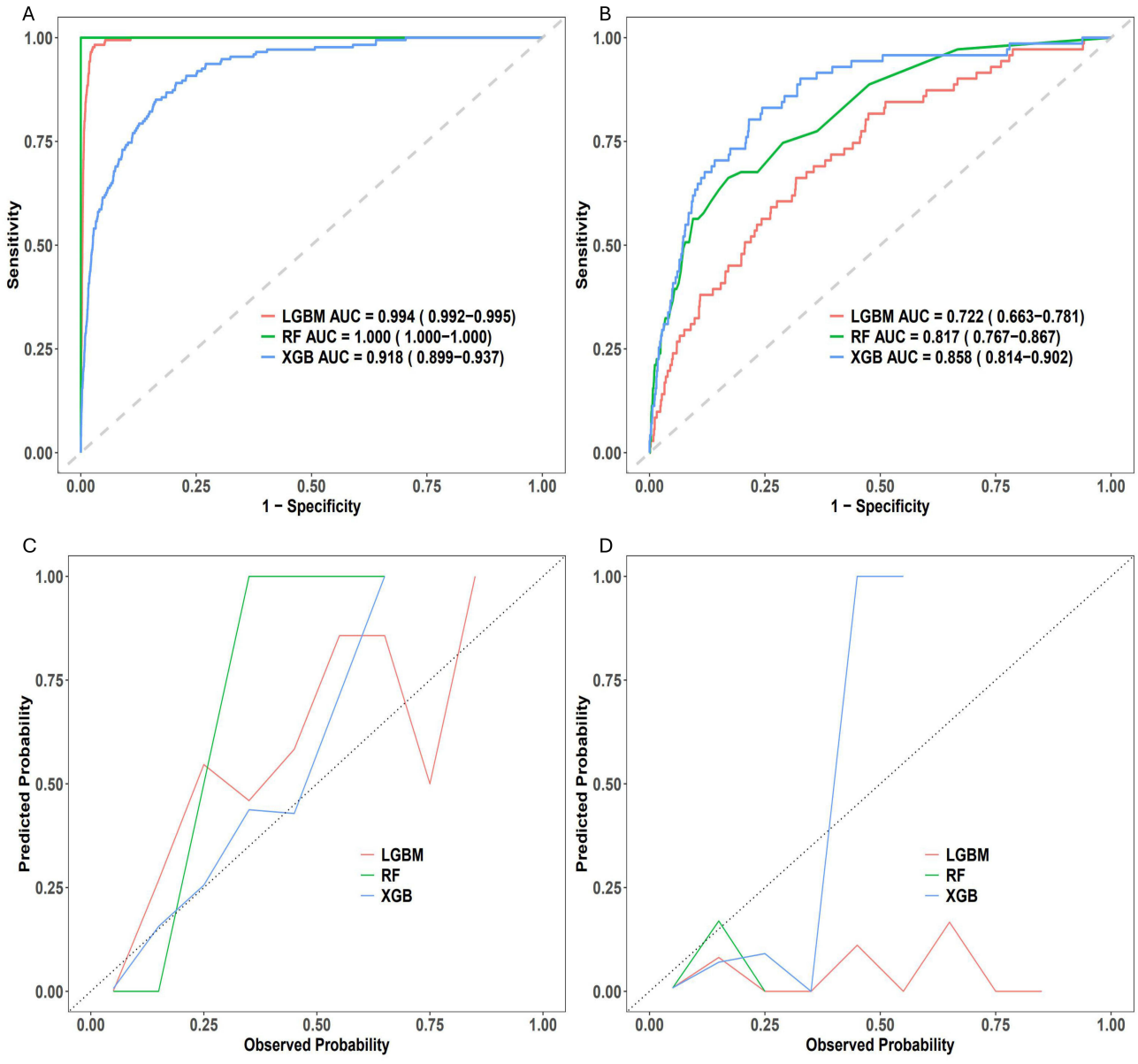

Figure 4 ROC and calibration curves of three models. (**A**) ROC curves in the training set; (**B**) in the validation set. (**C**) Calibration curves in the training set; (**D**) in the validation set. Calibration curves illustrate the correspondence between predicted bleeding risk (x-axis) and actual diagnosed cases (y-axis). The grey diagonal dotted line represents perfect predictions by an ideal model. A closer alignment between the model line and diagonal dotted lines suggests better prediction performance.

https://doi.org/10.2147/JBM.S567773

set AUC declined to 0.817, indicating overfitting—a limitation not observed in XGBoost's stable training-to-test performance (Training AUC=0.918 vs Test AUC=0.858).

Model Interpretation

The SHAP analysis elucidated the complex risk dynamics underlying thrombocythemia prediction during third-generation cephalosporin therapy. As depicted in Figure 5, baseline PLT emerged as the most influential predictor, with SHAP values spanning −1.67 to +1.48, reflecting its dual role as both a protective (higher counts) and risk-enhancing (lower counts) factor. RBC and CRE ranked second and third in global importance, demonstrating nonlinear relationships with thrombocythemia risk, as evidenced by SHAP value distributions across their physiological ranges. Individualized force plots (Figure 6) revealed context-specific risk modulation. Figure 6A (Case 1); Figure 6B (Case 2); Figure 6C (Case 3); Figure 6D (Case 4). For instance, in a representative case (Age=47, PLT=201×10^9/L), reduced platelet count contributed a risk elevation (SHAP=−0.765), while elevated RBC (3.72×10^{12}/L) exerted a protective effect (SHAP=+0.37). Comorbidities such as active malignancies amplified risk (SHAP=−0.215), whereas female sex showed modest protection (SHAP=−0.194).

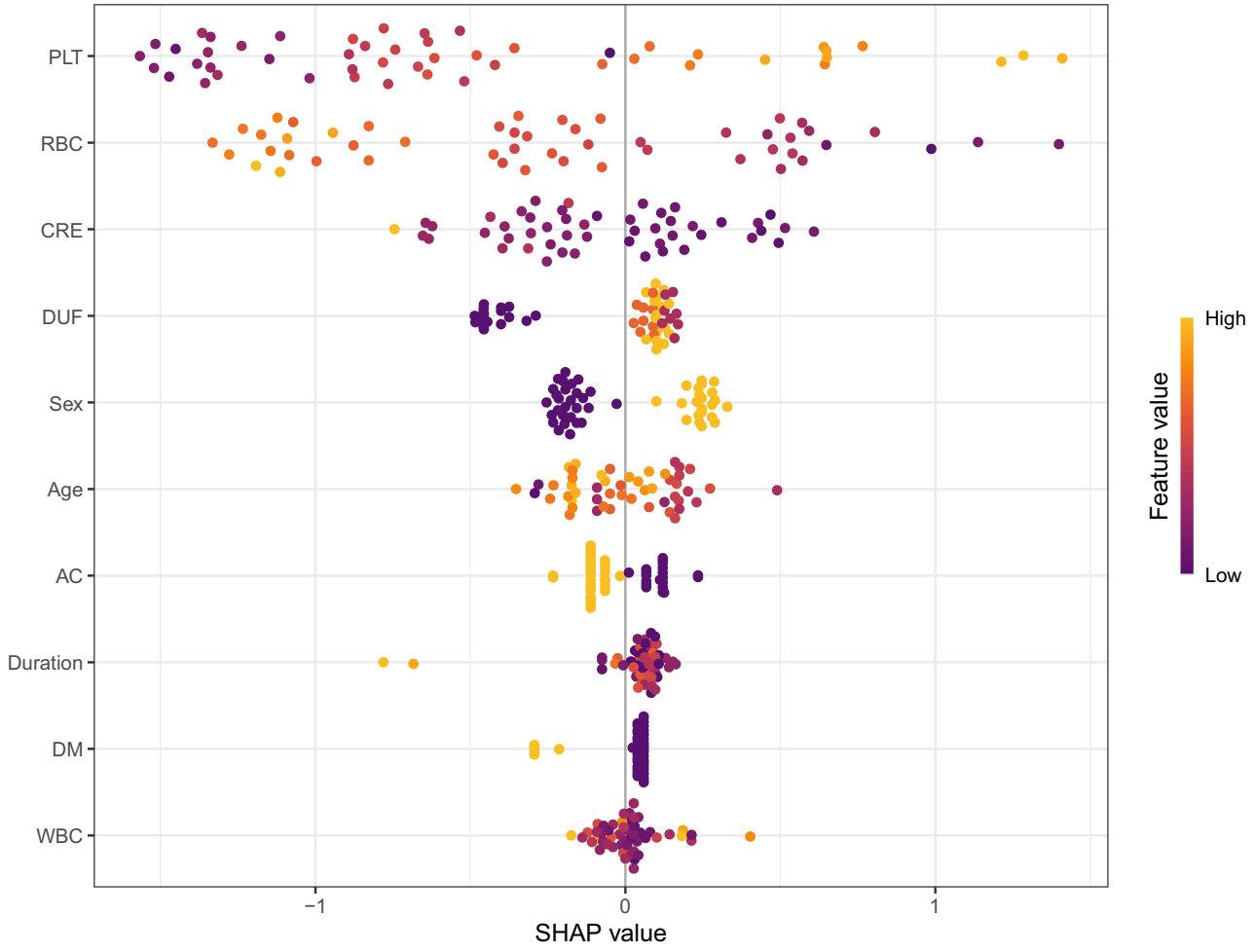

Figure 5 SHAP value interpretation of the XGBoost model predictions. The beeswarm plot visualizes feature contributions to individual predictions. Each point corresponds to a participant's variable, with its horizontal position reflecting the directional influence on the model output (right: positive impact; left: negative impact). SHAP value magnitude (SHAP) quantifies the variable's effect strength, where higher absolute values correlate with stronger associations with risk. Variable values are encoded by color: purple (high values, enhancing prediction likelihood) and yellow (low values, reducing prediction likelihood). The plot highlights variables with the largest absolute SHAP values as key determinants of model decisions.

https://doi.org/10.2147/JBM.S567773

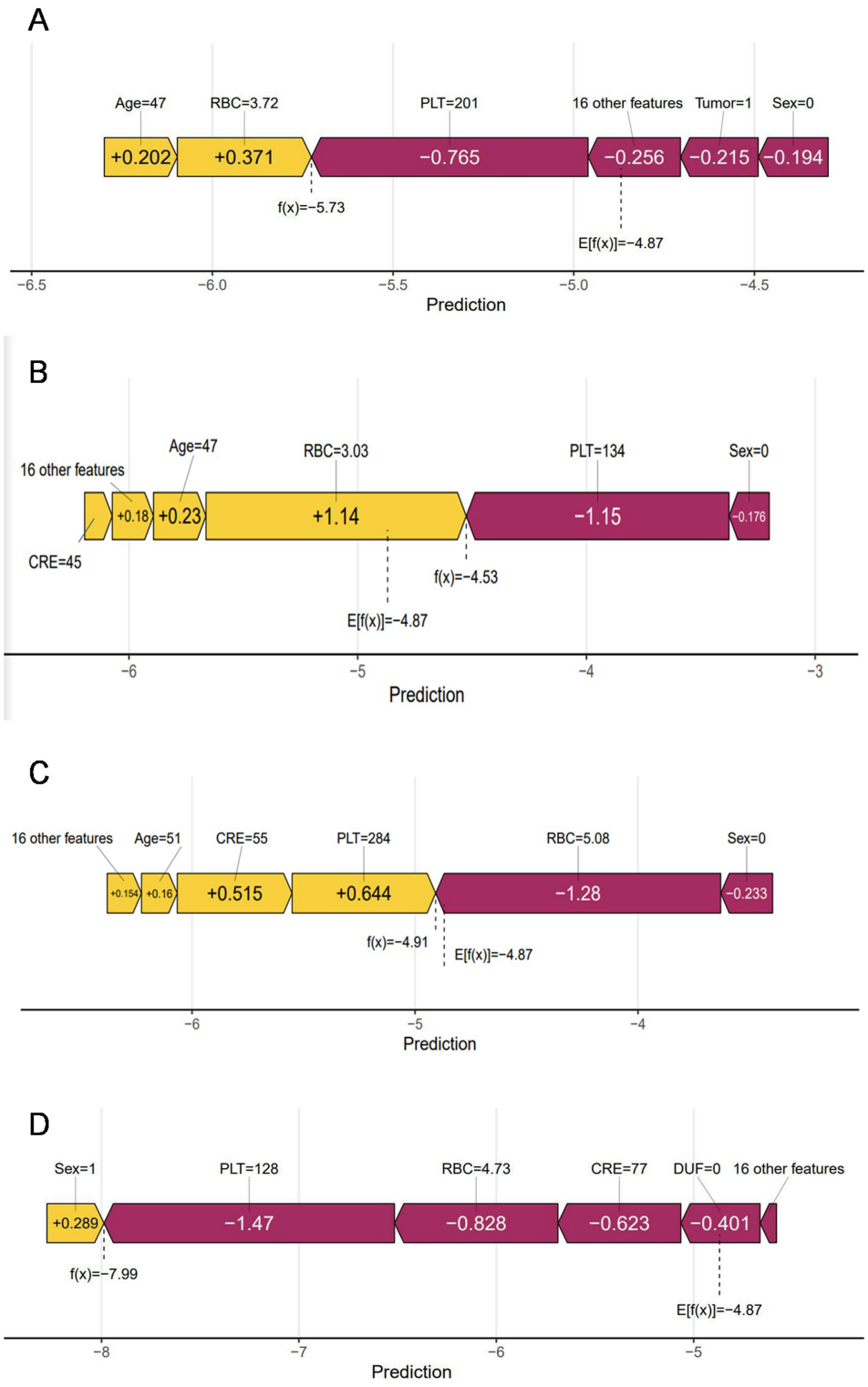

Figure 6 Feature contribution analysis (SHAP values) for XGBoost model prediction (**A**–**D**) SHAP plots visualize key feature contributions to individual predictions across four independent study cases (Case 1–4). Model and interpretation: Predictions were generated by an Extreme Gradient Boosting (XGBoost) model, with SHAP values (SHapley Additive exPlanations) quantifying directional contributions of features to individual outcomes. X-axis: SHAP value magnitude and direction. Positive values (right) indicate features increasing prediction probability; negative values (left) indicate reduced probability. 6A (Case 1); 6B (Case 2); 6C (Case 3); 6D (Case 4).

https://doi.org/10.2147/JBM.S567773

Discussion

This study presents the first interpretable machine learning framework for predicting thrombocythemia risk during third-generation cephalosporin therapy, addressing critical gaps in antibiotic safety monitoring. The XGBoost model demonstrated superior predictive performance, achieving the highest discrimination (test-set AUC=0.858) and calibration accuracy (Brier score=0.0088) among three machine learning algorithms. Baseline platelet count emerged as the dominant predictor, with SHAP values reflecting its dual role as both a protective (higher counts) and risk-enhancing (lower counts) factor.

Third-generation cephalosporins are widely used in clinical settings to treat various bacterial infections. However, these drugs may cause thrombocytosis, an abnormality of the blood system that can lead to serious consequences such as clot formation. Studies have shown that baseline platelet levels and red blood cell counts may be crucial factors influencing thrombocytosis during the treatment with third-generation cephalosporins. Researchers found that patients with higher baseline platelet levels were more likely to experience thrombocytosis when treated with third-generation cephalosporin antibiotics. This may be because high baseline platelet levels themselves indicate a certain degree of hematological activity in the body, which can further increase platelet production when stimulated by medication.[22] It has also been found in our research to be the most important influencing factor. Red blood cell count is also associated with the occurrence of thrombocytosis. An increase in the red blood cell count can lead to changes in blood viscosity, thereby affecting platelet production and function. Some studies have indicated that an abnormally elevated red blood cell count may be related to the development of thrombocytosis, especially in certain hematological disorders, where this association is more pronounced.[23] Our research further confirms that the duration of use is an important factor affecting thrombocythemia. Changes in creatinine levels may be an important risk factor. It was found that elevated creatinine levels were significantly associated with the occurrence of thrombocytosis. This could be due to impaired drug metabolism and excretion in the body caused by renal insufficiency, thereby increasing the toxic effects of drugs.[24] In this study, we also found that creatinine is the risk of thrombocythemia. Studies have shown that the duration and frequency of cephalosporin use are significant factors influencing thrombocytosis. Specifically, patients who use cephalosporins for more than 14 days are more likely to experience thrombocytosis.[25] The daily dose of cephalosporins is also a significant risk factor, especially when the daily dose reaches or exceeds 6 grams, significantly increasing the risk of thrombocytosis.[25] These two factors have also been found to have similar effects in our research. In addition, the use of anticoagulants is also an important risk factor for thrombocytosis.[26,27] Our research also confirms this fact.

Elderly patients, due to physiological decline, may be more susceptible to drug-related adverse reactions, including thrombocytosis.[28] In addition, the metabolism and excretion of drugs are also associated with the occurrence of thrombocytosis. In patients with renal insufficiency, the metabolism and excretion of third-generation cephalosporins may be affected, resulting in increased drug concentration in the body and an increased risk of thrombocytosis.[28] These two factors were also found to have a significant impact on thrombocytosis in our study.

Notably, female sex was consistently associated with a lower risk of thrombocythemia, a finding that aligns with known sex-based differences in drug metabolism and hematopoietic responses, which may be related to gender-specific physiological differences.[29] Patients with diabetes may face a higher risk of thrombocytosis when using third-generation cephalosporins. Diabetes itself is a chronic metabolic disease often accompanied by vascular lesions and hematological abnormalities. The platelet function in diabetic patients may already be affected, making them more prone to abnormal changes in platelet counts when using third-generation cephalosporins.[30] These two common factors were also found to have a significant impact in our study.

Our innovative risk model integrates machine learning with multidimensional variables (demographics, laboratory parameters, comedications, comorbidities) for personalized risk stratification. By integrating real-time risk stratification into EHR systems, we bridge the gap between algorithmic prediction and clinical stewardship. The model's SHAP interface enables personalized risk visualization. A key consideration for the utility of our model is its potential integration into routine clinical workflows. We envision it functioning as a decision-support tool within electronic health record systems, where it could automatically flag patients at high risk for thrombocythemia at the time of antibiotic prescription. This would prompt clinicians to institute pre-emptive monitoring strategies, such as scheduling earlier follow-up platelet counts, thereby shifting the paradigm from reaction to prevention.

This study should be interpreted in the context of its limitations. A key limitation stems from its retrospective design, which restricts causal inference and may introduce selection bias, a factor that future prospective studies are needed to address. Another significant constraint is the lack of external validation, which is a crucial next step before any clinical application can be considered; we explicitly acknowledge this and are planning a multi-center collaboration to this end. Lastly, the strategy of removing variables with excessive missing data was adopted to ensure model stability, but it carries a risk of information bias. Advanced data collection protocols in subsequent research will be crucial to minimize this issue and capture a more complete set of predictive features.

Conclusion

This study demonstrates that interpretable machine learning effectively predicts thrombocytosis risk during third-generation cephalosporin therapy. While retrospective design limits causal inference, this framework advances precision antimicrobial stewardship and provides a replicable template for drug safety analytics. Future work should validate risk archetypes prospectively and incorporate pharmacogenomic biomarkers to optimize personalized dosing strategies.

Data Sharing Statement

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author (Maofeng Wang) upon reasonable request. Data are located in controlled access data storage at the Affiliated Dongyang Hospital of Wenzhou Medical University.

Ethics Approval

This study received approval from the Medical Ethics Committee of the Affiliated Dongyang Hospital of Wenzhou Medical University (approval #2025-YX-012), and informed consent was waived by the Ethics Committee of Affiliated Dongyang Hospital of Wenzhou Medical University. Patient records/information were anonymized and deidentified before analysis.

Funding

This research was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LTGY23H200002.

Disclosure

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Liaskas A, Vassilakopoulos TP. Thrombocytosis. *Blood*. 2024;143(17):1782. doi:10.1182/blood.2023023702
2. Vo QT, Thompson DF. A review and assessment of drug-induced thrombocytosis. *Ann Pharmacother*. 2019;53(5):523–536. doi:10.1177/1060028018819450
3. Tefferi A, Vannucchi AM, Barbui T. Essential thrombocythemia: 2024 update on diagnosis, risk stratification, and management. *Am J Hematol*. 2024;99(4):697–718. doi:10.1002/ajh.27216
4. Donnelly PC, Sutich RM, Easton R, Adejumo OA, Lee TA, Logan LK. Ceftriaxone-associated biliary and cardiopulmonary adverse events in neonates: a systematic review of the literature. *Paediatr Drugs*. 2017;19(1):21–34. doi:10.1007/s40272-016-0197-x
5. Li X, Lei W, Wang M, Xu L. Risk stratification for cephalosporin-induced thrombocytopenia: development and validation of a multidimensional predictive model in older adults. *Risk Manage Healthcare Policy*. 2025;18:2107–2120. doi:10.2147/rmhp.S529488
6. Duong TA, Valeyrie-Allanore L, Wolkenstein P, Chosidow O. Severe cutaneous adverse reactions to drugs. *Lancet*. 2017;390(10106):1996–2011. doi:10.1016/s0140-6736(16)30378-6
7. Lin Y-F, Yang C-H, Sindy H, et al. Severe cutaneous adverse reactions related to systemic antibiotics. *Clin Infect Dis*. 2014;58(10):1377–1385. doi:10.1093/cid/ciu126
8. Hsieh CC, Lee CH, Li MC, et al. Empirical third-generation cephalosporin therapy for adults with community-onset Enterobacteriaceae bacteraemia: impact of revised CLSI breakpoints. *Int J Antimicrob Agents*. 2016;47(4):297–303. doi:10.1016/j.ijantimicag.2016.01.010
9. Maillard A, Delory T, Bernier J, et al. Effectiveness of third-generation cephalosporins or piperacillin compared with cefepime or carbapenems for severe infections caused by wild-type AmpC β-lactamase-producing Enterobacterales: a multi-centre retrospective propensity-weighted study. *Int J Antimicrob Agents*. 2023;62(1):106809. doi:10.1016/j.ijantimicag.2023.106809

10. Lin W-P, Huang Y-S, Wang J-T, Chen Y-C, Chang S-C. Prevalence of and risk factor for community-onset third-generation cephalosporin-resistant Escherichia coli bacteremia at a medical center in Taiwan. *BMC Infect Dis*. 2019;19(1):245. doi:10.1186/s12879-019-3880-z
11. Mendes IF, Completo S, Vieira de Carvalho R, et al. Salmonellosis in children at a Portuguese hospital: a retrospective study. *Acta Médica Portuguesa*. 2023;36(2):96–104. doi:10.20344/amp.18906
12. Bouiller K, Gbaguidi-Haore H, Hocquet D, et al. The effects of switching from ceftriaxone to cefotaxime on the occurrence of third-generation cephalosporin-resistant Enterobacterales: a stepped-wedge cluster randomized trial. *Infect Dis Now*. 2024;54(1):104806. doi:10.1016/j.idnow.2023.104806
13. Robert W, Denis A, Thomas A, et al. A comprehensive review on cryptographic techniques for securing internet of medical things: a state-of-the-art, applications, security attacks, mitigation measures, and future research direction. *Mesopot J Art Intellige Healthcare*. 2024;(2024):135–169. doi:10.58496/MJAIH/2024/016
14. Alkattan H, Al-Nuaimi BT, Subhi AA. Machine learning techniques to predictive in healthcare: hepatitis C diagnosis. *Mesopota J Art Intellige Healthcare*. 2024;2024:128–134. doi:10.58496/MJAIH/2024/015
15. Chen C, Guo D-H, Cao X, et al. Risk factors for thrombocytopenia in adult Chinese patients receiving linezolid therapy. *Curr Ther Res Clin Exp*. 2012;73(6):195–206. doi:10.1016/j.curtheres.2012.07.002
16. Ling F, Jianling Q, Maofeng W. Development and validation of a novel model to predict pulmonary embolism in cardiology suspected patients: a 10-year retrospective analysis. *Open Med*. 2024;19(1):20240924. doi:10.1515/med-2024-0924
17. Mzili T, Mzili M, Bouderba SI, Abatal A, Aribowo W, Oughannou Z. The role of artificial intelligence in early tumor detection: an XGBoost risk assessment model for Egyptian patients. *Mesopota J Art Intellige Healthcare*. 2025;2025:85–92. doi:10.58496/MJAIH/2025/008
18. Qian Y, Wanlin L, Maofeng W. Machine learning derived model for the prediction of bleeding in dual antiplatelet therapy patients. *Front Cardiovasc Med*. 2024;11:1402672. doi:10.3389/fcvm.2024.1402672
19. Chen T, Lei W, Wang M. Predictive model of internal bleeding in elderly aspirin users using XGBoost machine learning. *Risk Manage Healthcare Policy*. 2024;17:2255–2269. doi:10.2147/rmhp.S478826
20. Liang C, Wanling L, Maofeng W. LASSO-derived model for the prediction of bleeding in aspirin users. *Sci Rep*. 2024;14(1):12507. doi:10.1038/s41598-024-63437-6
21. Jing J, Wanling L, Maofeng W. A practical nomogram for predicting the bleeding risk in patients with a history of myocardial infarction treating with aspirin. *Clin Appl Thromb Hemost*. 2024;30:10760296241262789. doi:10.1177/10760296241262789
22. Mikič T B, Vratanar B, Pajič T, et al. Is it possible to predict clonal thrombocytosis in triple-negative patients with isolated thrombocytosis based only on clinical or blood findings? *J Clin Med*. 2021;10(24). doi:10.3390/jcm10245803
23. Grenier JMP, El Nemer W, De Grandis M. Red blood cell contribution to thrombosis in polycythemia vera and essential thrombocythemia. *Int J Mol Sci*. 2024;25(3):1417. doi:10.3390/ijms25031417
24. Chen X, Jin H, Wang D, et al. Serum creatinine levels, traditional cardiovascular risk factors and 10-year cardiovascular risk in Chinese patients with hypertension. *Front Endocrinol*. 2023;14:1140093. doi:10.3389/fendo.2023.1140093
25. Zhu B, Jin P, Li J, Zhu Y. Retrospective analysis of risk factors for cefoperazone/sulbactam-induced thrombocytopenia in adult chinese patients: a six-year real-world study. *Infect Drug Resist*. 2024;17:3901–3911. doi:10.2147/idr.S475590
26. Grandhi R, Harrison G, Voronovich Z, et al. Preinjury warfarin, but not antiplatelet medications, increases mortality in elderly traumatic brain injury patients. *J Trauma Acute Care Surg*. 2015;78(3):614–621. doi:10.1097/ta.0000000000000542
27. Piel-Julian M-L, Mahévas M, Germain J, et al. Risk factors for bleeding, including platelet count threshold, in newly diagnosed immune thrombocytopenia adults. *J Thromb Haemost*. 2018;16(9):1830–1842. doi:10.1111/jth.14227
28. Bathini L, Jandoc R, Kuwornu P, et al. Clinical outcomes of failing to dose-reduce cephalosporin antibiotics in older adults with CKD. *Clin J Am Soc Nephrol*. 2019;14(2):197–205. doi:10.2215/cjn.10710918
29. Venkat RK, Redd RA, Harris AC, et al. Risk of bleeding in patients with essential thrombocythemia and extreme thrombocytosis. *Blood Adv*. 2024;8(23):6043–6054. doi:10.1182/bloodadvances.2024013777
30. Herczeg G, Somogyi A, Herold M, et al. Does diabetes affect paraneoplastic thrombocytosis in colorectal cancer? *Open Med*. 2022;17(1):160–173. doi:10.1515/med-2021-0407