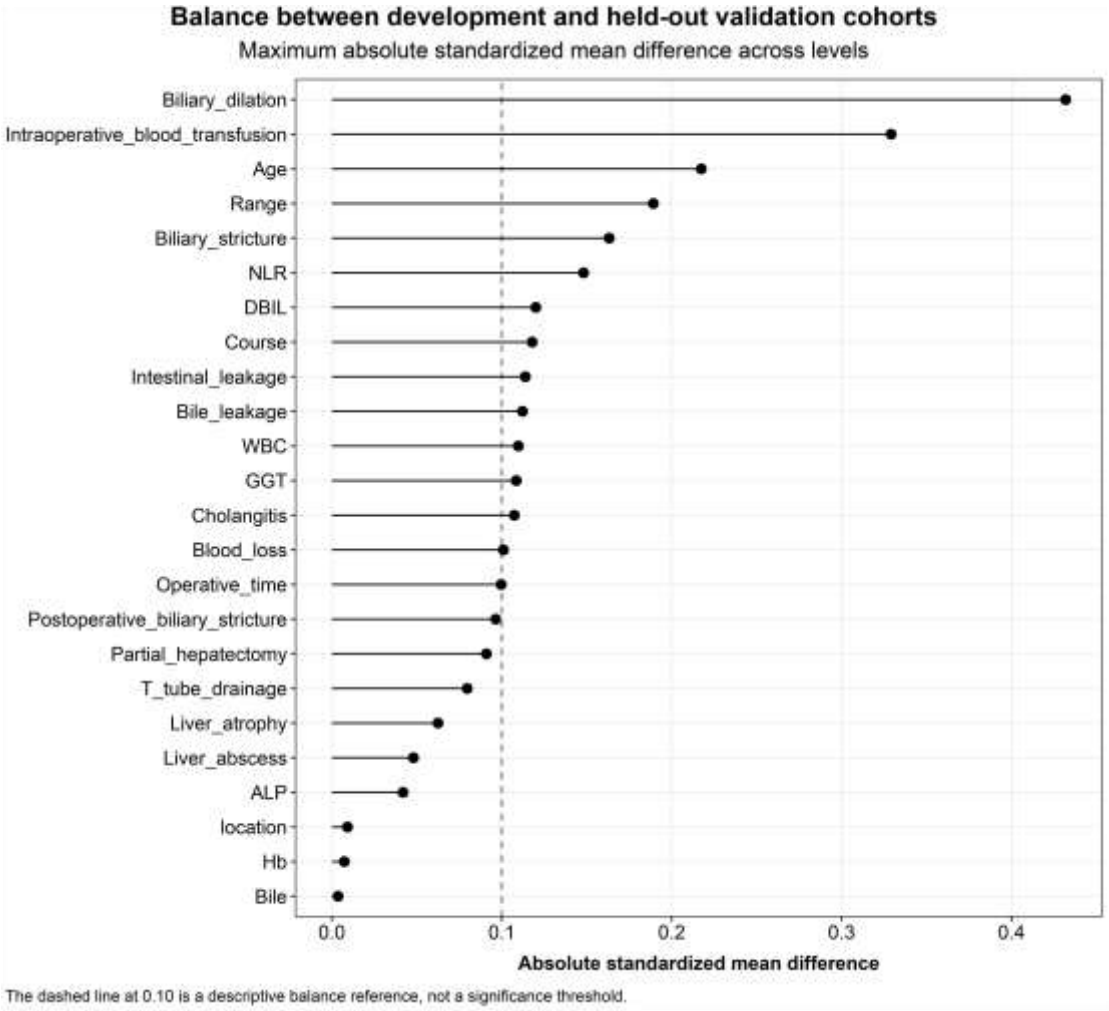


Supplementary Figure 1 Preoperative laboratory characteristics of the study cohort. Absolute standardized mean differences values across baseline demographic, clinical, and laboratory characteristics evaluated between the development and validation cohorts. The vertical dashed line indicates the conventional descriptive threshold of 0.10 for covariate balance. DBIL, direct bilirubin; GGT, gamma-glutamyl transferase; Hb, hemoglobin; NLR, neutrophil-to-lymphocyte ratio; WBC, white blood cell.

Supplementary Table 1 Preoperative laboratory characteristics of the study cohort

Item	Overall (*n* = 300)	Non–residual stone group (*n* = 218)	Residual stone group (*n* = 82)	*P* value
Laboratory values				
TP (g/L)	69.40±7.60	68.77±7.43	71.06±7.83	0.020
ALB (g/L)	39.75 (35.80–42.10)	39.55 (36.02–41.98)	40.30 (35.25–42.85)	0.425
GLB (g/L)	30.20 (26.60–33.30)	29.60 (26.40–32.95)	32.30 (28.22–35.20)	0.002
A/G	1.29 (1.17–1.46)	1.32 (1.19–1.47)	1.23 (1.11–1.42)	0.013
PAB (mg/L)	206.55±57.00	209.09±55.25	199.80±61.24	0.209
ALT (U/L)	36.00 (19.00–66.25)	36.00 (19.00–65.75)	34.50 (20.25–67.00)	0.691
AST (U/L)	31.00 (22.00–47.25)	31.00 (22.00–49.00)	32.00 (23.00–46.00)	0.742
ALP (U/L)	135.50 (93.00–236.75)	126.50 (91.00–209.75)	169.00 (101.25–304.50)	0.018
GGT (U/L)	145.00 (52.75–329.50)	133.00 (47.50–292.25)	203.00 (70.50–381.25)	0.036
TBA (μmol/L)	4.70 (2.70–8.90)	4.50 (2.62–8.17)	4.95 (3.32–9.55)	0.288
TBIL (μmol/L)	14.50 (10.20–22.85)	14.05 (9.90–20.98)	16.75 (11.03–25.10)	0.051
DBIL (μmol/L)	5.00 (3.30–10.70)	4.85 (3.10–9.45)	6.50 (3.60–12.52)	0.029

IBIL (μmol/L)	8.40 (6.20–12.43)	8.05 (5.82–12.40)	9.30 (6.90–12.47)	0.112
LDH (U/L)	176.00 (152.00–202.00)	179.00 (154.00–202.00)	172.50 (148.00–199.25)	0.242
CREA (μmol/L)	63.00 (55.00–75.00)	63.00 (55.00–77.00)	62.50 (55.25–70.50)	0.743
UA (μmol/L)	284.50 (230.00–343.00)	290.50 (237.25–352.50)	265.00 (218.50–332.75)	0.169
TC (mmol/L)	4.58 (3.88–5.66)	4.59 (3.96–5.52)	4.51 (3.80–6.24)	0.998
TG (mmol/L)	1.23 (0.91–1.69)	1.23 (0.94–1.70)	1.25 (0.83–1.62)	0.895
PLT (10^9/L)	227.00 (179.75–294.50)	232.00 (182.25–305.75)	210.50 (176.00–270.75)	0.121
WBC (10^9/L)	5.64 (4.58–6.99)	5.54 (4.55–6.90)	5.87 (4.71–7.21)	0.414
NEUT (10^9/L)	3.27 (2.40–4.32)	3.17 (2.38–4.27)	3.38 (2.54–4.75)	0.255
LYMPH (10^9/L)	1.62 (1.27–2.10)	1.60 (1.25–2.04)	1.65 (1.30–2.26)	0.518
MONO (10^9/L)	0.37 (0.27–0.48)	0.37 (0.27–0.47)	0.36 (0.25–0.51)	0.891
RBC (10^{12}/L)	4.48±0.57	4.48±0.56	4.46±0.59	0.772
Hb (g/L)	135.00 (124.00–146.00)	135.00 (124.00–145.00)	135.50 (124.50–147.00)	0.523
Hct (L/L)	0.41 (0.37–0.44)	0.41 (0.37–0.43)	0.41 (0.38–0.44)	0.513
NLR	1.85 (1.35–2.90)	1.84 (1.35–2.86)	2.05 (1.40–2.99)	0.517
LMR	4.53 (3.20–6.52)	4.49 (3.25–6.47)	4.75 (3.16–6.76)	0.714

PLR	141.50 (102.92–186.65)		144.03 (103.29–192.09)		135.82 (100.00–178.01)		0.241
SII	442.20 (291.55–719.21)		436.89 (292.06–815.43)		450.48 (292.44–669.74)		0.912
PT (S)	11.40 (10.90–12.20)		11.50 (10.90–12.30)		11.40 (11.00–11.90)		0.482
APTT (S)	26.20 (24.80–28.00)		26.20 (24.72–28.10)		26.40 (25.15–27.60)		0.682
TT (S)	16.70 (15.80–17.50)		16.70 (16.00–17.50)		16.70 (15.40–17.50)		0.366
FIB (g/L)	3.48 (2.80–4.52)		3.41 (2.79–4.51)		3.71 (2.88–4.64)		0.175
INR	0.97 (0.94–1.03)		0.97 (0.94–1.03)		0.97 (0.94–1.01)		0.767
D–dimer (μg/mL)	0.50 (0.30–0.99)		0.47 (0.29–0.96)		0.57 (0.35–1.10)		0.059
FDPs (μg/mL)	2.50 (2.50–3.31)		2.50 (2.50–3.29)		2.50 (2.50–3.33)		0.787

Data are presented as mean ± SD, median (IQR), or n (%). TP, total protein; ALB, albumin; GLB, globulin; A/G, albumin-to-globulin ratio; PAB, prealbumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALP, alkaline phosphatase; GGT, gamma-glutamyl transpeptidase; TBA, total bile acid; TBIL, total bilirubin; DBIL, direct bilirubin; IBIL, indirect bilirubin; LDH, lactate dehydrogenase; CREA, creatinine; UA, uric acid; TC, total cholesterol; TG, triglycerides; PLT, platelet count; WBC, white blood cell count; NEUT, neutrophil count; LYMPH, lymphocyte count; MONO, monocyte count; RBC, red blood cell count; Hb, hemoglobin; Hct, hematocrit; NLR, neutrophil-to-lymphocyte ratio; LMR, lymphocyte-to-monocyte ratio; PLR, platelet-to-lymphocyte ratio; SII, systemic immune-inflammation index; PT, prothrombin time; APTT, activated partial thromboplastin time; TT, thrombin time; FIB,

fibrinogen; INR, international normalized ratio; FDPs, fibrin degradation products.

Supplementary Table 2 All 12 candidate models: Development OOF performance

Model	AUC	Brier	Log loss	Calibration intercept	Calibration slope
Random forest	0.803	0.154	0.472	0.016	1.085
XGBoost	0.817	0.153	0.469	0.227	0.705
RBF-SVM	0.785	0.158	0.481	-0.055	0.939
LASSO-LR	0.849	0.138	0.432	0.022	1.748
k-NN	0.743	0.180	0.596	0.723	0.709
PLSR	0.784	0.175	0.535	-0.581	2.237
GBM	0.813	0.150	0.457	0.076	0.938
ANN	0.757	0.172	0.568	0.047	0.475
Naive Bayes	0.788	0.187	0.689	0.937	0.334
LDA	0.796	0.171	0.565	0.122	0.434
CatBoost	0.801	0.167	0.518	0.510	0.589
LightGBM	0.780	0.183	0.679	0.855	0.338

All performance metrics were derived from aggregated out-of-fold (OOF) predictions across repeated 5-fold cross-validation (5-fold×3 repeats, n=211) in the development cohort. Perfect calibration is indicated by a calibration intercept of 0 and a calibration slope of 1.0. The LASSO-logistic model achieved optimal performance across discrimination (AUC=0.849) and overall fit (Brier=0.138), establishing it as the frozen primary candidate for validation. ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; CatBoost, categorical boosting; GBM, gradient boosting machine; k-NN, k-nearest neighbors; LASSO-LR, least absolute shrinkage and

selection operator-penalized logistic regression; LDA, linear discriminant analysis; LightGBM, light gradient boosting machine; OOF, out-of-fold; PLSR, partial least squares regression; RBF-SVM, radial basis function support vector machine; XGBoost, extreme gradient boosting.

Supplementary Table 3 All 12 candidate models: Held-out validation performance

Model	AUC	Brier	Log loss	Calibration intercept	Calibration slope
Random forest	0.855	0.133	0.404	-0.024	1.301
XGBoost	0.854	0.136	0.405	0.180	0.762
RBF-SVM	0.877	0.127	0.396	-0.033	1.487
LASSO-LR	0.877	0.134	0.417	0.006	1.939
k-NN	0.779	0.169	0.518	0.796	1.404
PLSR	0.849	0.162	0.503	-0.606	2.504
GBM	0.868	0.129	0.396	0.032	1.357
ANN	0.800	0.221	0.702	-0.400	0.338
Naive Bayes	0.761	0.218	0.869	0.052	0.206
LDA	0.849	0.158	0.469	0.154	0.534
CatBoost	0.855	0.134	0.406	0.246	0.870
LightGBM	0.845	0.159	0.502	0.411	0.462

Performance metrics were computed on the strictly isolated held-out validation cohort (n=89, 24 events) following a single evaluation pass. Perfect calibration corresponds to a calibration intercept of 0 and a calibration slope of 1.0. The pre-locked LASSO-LR model confirmed excellent generalizability with an AUC of 0.877 and exceptional calibration alignment (intercept=0.006). ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; CatBoost, categorical boosting; GBM, gradient boosting machine; k-NN,

k-nearest neighbors; LASSO-LR, least absolute shrinkage and selection operator–logistic regression; LDA, linear discriminant analysis; LightGBM, light gradient boosting machine; PLSR, partial least squares regression; RBF-SVM, radial basis function support vector machine; XGBoost, extreme gradient boosting.

Supplementary Table 4 Bootstrap intervals for selected-model probabilistic performance

Metric	Development estimate	Development 95%CI	Validation estimate	Validation 95%CI
AUC	0.849	0.791-0.900	0.877	0.798-0.939
Brier	0.138	0.113-0.163	0.134	0.098-0.171
Log loss	0.432	0.375-0.492	0.417	0.335-0.499
Calibration intercept	0.022	-0.288 to 0.318	0.006	-0.543 to 0.467
Calibration slope	1.748	1.365-2.315	1.939	1.353-3.107

Performance metrics correspond to the pre-specified, locked LASSO-logistic regression model. Development estimates reflect aggregated OOF predictions (5-fold×3 repeats, n=211), whereas validation estimates reflect the strictly isolated held-out cohort (n=89). 95%CIs were estimated *via* 1,000 non-parametric bootstrap iterations within each cohort. A calibration intercept of 0 and slope of 1.0 indicate perfect calibration alignment. AUC, area under the receiver operating characteristic curve; CI, confidence interval; LASSO-LR, least absolute shrinkage and selection operator–logistic regression; OOF, out-of-fold.

Supplementary Table 5 Model-rank uncertainty and outer-fold stability

Model	Bootstrap rank-first count	Bootstrap probability	Mean outer-fold AUC rank	Outer folds ranked first
Random forest	5	0.005	5.667	2
XGBoost	57	0.057	4.933	0
RBF-SVM	0	0.000	6.200	2
LASSO-LR	920	0.920	2.233	6
k-NN	0	0.000	9.933	0
PLSR	0	0.000	7.600	0
GBM	10	0.010	5.567	1
ANN	0	0.000	7.900	0
Naive Bayes	3	0.003	8.167	0
LDA	1	0.001	5.800	1
CatBoost	4	0.004	6.667	0
LightGBM	0	0.000	7.333	1

Model-rank uncertainty was evaluated using 1,000 non-parametric bootstrap resamples of the development cohort (n=211). Bootstrap probability represents the proportion of iterations in which a given model achieved the highest AUC. Outer-fold stability metrics summarize model rankings across all 15 outer folds generated by repeated cross-validation (5-fold×3 repeats). LASSO-LR demonstrated decisive dominance, achieving rank-first status in 92.0% of bootstrap resamples. ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; CatBoost, categorical boosting; GBM, gradient boosting machine; k-NN, k-nearest neighbors; LASSO-LR, least absolute shrinkage and selection operator–logistic regression; LDA, linear discriminant analysis; LightGBM, light gradient boosting machine; PLSR,

partial least squares regression; RBF-SVM, radial basis function support vector machine; XGBoost, extreme gradient boosting.