Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multidomain geriatric assessment improves prediction of unplanned inpatient admission after day-case ERCP in older adults

Xiaoli Sun[1]
The Second Affiliated Hospital of Harbin Medical University
Xiuli Liu[1]
The Second Affiliated Hospital of Harbin Medical University
Jing Chen[1]
The Second Affiliated Hospital of Harbin Medical University
Jitao Song[1]
The Second Affiliated Hospital of Harbin Medical University
Cuihua Li[1]
The Second Affiliated Hospital of Harbin Medical University
Hao Yang[1]
The Second Affiliated Hospital of Harbin Medical University
Lina Liu[1]
The Second Affiliated Hospital of Harbin Medical University
Lei Zhao[1]
drleizhao@hrbmu.edu.cn

The Second Affiliated Hospital of Harbin Medical University

Article

Keywords: Endoscopic Retrograde Cholangiopancreatography, Geriatric Assessment, Frailty, Patient Discharge, Risk Assessment

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11360444/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Identifying older patients at risk of unplanned inpatient admission after planned day-case endoscopic retrograde cholangiopancreatography (ERCP) remains challenging because conventional clinical factors may overlook discharge-related vulnerabilities. We assessed the incremental predictive value of multidomain geriatric assessment in a retrospective cohort of 1,407 patients aged ≥ 65 years undergoing planned day-case ERCP, comprising a development cohort (n = 1,000) and a subsequent temporal validation cohort (n = 407). Adding frailty, functional independence, nutritional status, and cognition to conventional preprocedural factors increased the temporal-validation AUC from 0.730 to 0.830 (difference, 0.100; 95% CI, 0.039–0.164), improved average precision from 0.395 to 0.552, and reduced the Brier score from 0.133 to 0.118. Models additionally incorporating procedural and early-recovery information available before the final discharge decision achieved temporal-validation AUCs of 0.872 for L2-penalized logistic regression and 0.860 for XGBoost, although calibration deteriorated. These findings demonstrate that multidomain geriatric assessment provides incremental predictive value beyond conventional clinical factors in older adults undergoing planned day-case ERCP. Integrating geriatric assessment with procedural and early-recovery information may support individualized risk stratification, perioperative monitoring, and discharge planning.

Introduction

Endoscopic retrograde cholangiopancreatography (ERCP) is increasingly performed in a day-case setting, with same-day or early discharge offering the potential to reduce unnecessary hospitalization, shorten length of stay, and improve healthcare resource utilization [1, 2].

However, appropriate patient selection is essential to the success of this pathway, particularly in older adults. Despite the growing number of older patients undergoing ERCP, evidence to guide preprocedural assessment for planned day-case ERCP remains limited. Conventional risk assessment commonly emphasizes chronological age, comorbidity burden, clinical and laboratory status, and procedural considerations, yet these factors may not adequately capture the multidimensional vulnerability of older patients [3, 4].

Older adults may differ substantially in frailty, functional independence, nutritional status, cognition, and physiologic reserve, even among individuals of the same chronological age [5, 6]. These differences may influence postprocedural recovery, discharge readiness, and the ability to complete the intended day-case pathway [7]. Consequently, some patients initially considered suitable for day-case ERCP ultimately require unplanned inpatient admission [2]. Multidomain geriatric assessment, integrating frailty, function, nutrition, and cognition, may therefore provide clinically relevant information beyond conventional preprocedural factors and improve identification of patients at greater risk of failing the intended short-stay pathway [6, 8].

Previous ERCP research has focused predominantly on procedure-related adverse events, including post-ERCP pancreatitis, bleeding, cardiopulmonary complications, and mortality [3, 9, 10]. Although frailty and

geriatric vulnerability have been associated with adverse outcomes after ERCP [11, 12], few studies have specifically examined preprocedural risk stratification for unplanned inpatient admission among older adults already selected for planned day-case ERCP.

It therefore remains unclear whether structured multidomain geriatric assessment provides measurable incremental predictive value beyond chronological age and conventional clinical factors [11, 12]. The clinically relevant question is not simply whether complications occur, but whether an older patient can safely complete the planned day-case pathway.

Discharge risk is also dynamic: preprocedural factors inform initial suitability for day-case ERCP, while intraprocedural and early-recovery factors may alter the final discharge decision [2]. Risk assessment may therefore be most informative when performed at two clinically distinct time points: before ERCP, to support patient selection and anticipatory planning, and after ERCP, to reassess discharge suitability using procedural and recovery information. Whether this staged approach improves prediction beyond a static preprocedural assessment remains insufficiently established.

We therefore conducted a single-center retrospective cohort study of patients aged ≥ 65 years undergoing planned day-case ERCP. We sought to address three related questions: whether multidomain geriatric assessment improves prediction of unplanned inpatient admission beyond chronological age and conventional preprocedural factors; whether the incorporation of intraprocedural and early-recovery information further improves risk stratification; and whether model performance is maintained over time. To address these questions, we developed and temporally validated preprocedural and extended prediction models using L2-penalized logistic regression and XGBoost, with model interpretation further explored using SHapley Additive exPlanations (SHAP).

Methods

Study design and setting

This single-center retrospective cohort study was conducted at the Department of Gastroenterology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China. Consecutive patients undergoing ERCP from January 2020 through August 2026 were screened. Eligible patients were aged ≥ 65 years and underwent elective therapeutic ERCP with intended day-case discharge.

Planned day-case discharge was defined as discharge on the day of ERCP or early the following morning within the intended 24-hour day-case pathway. Each patient contributed one eligible episode, and both planned stone extraction and biliary stenting were eligible. Patients with acute cholangitis or biliary pancreatitis were included in the planned day-case pathway only after initial treatment, clinical stabilization, and confirmation during preprocedural assessment that they were suitable for elective therapeutic ERCP.

Patients were excluded if they met any of the following criteria: (1) age < 65 years; (2) emergency or urgent ERCP performed for acute biliary obstruction, uncontrolled acute cholangitis, or other conditions requiring immediate intervention; (3) planned inpatient ERCP or hospitalization required for other medical conditions before the procedure; (4) diagnostic-only ERCP without an intended therapeutic intervention; (5) preprocedural clinical instability, including hemodynamic instability, uncontrolled infection, or acute organ failure requiring inpatient management; (6) planned additional procedures or treatments necessitating hospitalization beyond the intended day-case pathway; or (7) insufficient medical records to establish the intended discharge plan or ascertain the occurrence of unplanned inpatient admission.

The development cohort comprised patients treated from January 2020 through December 2024, whereas the temporal validation cohort comprised those treated from January 2025 through August 2026.

The study was conducted in accordance with the Declaration of Helsinki and approved by the Medical Ethics Committee of The Second Affiliated Hospital of Harbin Medical University on January 12, 2020 (approval No. YJSKY2020-379). Informed consent was waived because of the retrospective design.

Day-case pathway and outcome

Patients underwent preprocedural assessment, therapeutic ERCP, periprocedural monitoring, and structured recovery before final discharge assessment. The intended day-case pathway allowed discharge on the day of ERCP or early the following morning within 24 hours.

The primary outcome was unplanned inpatient admission, defined as an unanticipated conversion from planned day-case management to inpatient care because of clinical concerns, inadequate recovery from anesthesia, persistent or newly developed symptoms, periprocedural abnormalities, or early ERCP-related adverse events requiring continued hospital-based treatment or monitoring beyond the planned discharge pathway. Patients discharged within the intended 24-hour pathway without unplanned conversion to inpatient care were classified as having achieved planned discharge. Discharge delays attributable solely to nonmedical or administrative reasons were not considered unplanned admissions unless an inpatient admission was documented.

Outcome status was ascertained from electronic medical records, including procedural reports, recovery records, admission orders, and discharge summaries. Post-ERCP pancreatitis, bleeding, and perforation were considered adverse outcomes or direct causes of admission rather than candidate predictors to minimize outcome leakage [13]. Extended models incorporated only procedural and early-recovery information available before the final discharge decision.

Candidate predictors

Predictors were prespecified according to clinical relevance and timing of availability and were classified as preprocedural, intraprocedural, or early-recovery variables.

Preprocedural clinical candidates included age, sex, planned procedure type, American Society of Anesthesiologists (ASA) physical status, comorbidity count, acute cholangitis, biliary pancreatitis, creatinine clearance, serum albumin, alanine aminotransferase, previous ERCP, previous pancreatitis, antithrombotic therapy, periampullary diverticulum, largest stone diameter, and stone count [13]. Social and logistical candidates included living alone, adequate 24-hour caregiver availability, and travel time > 1 hour.

Planned procedure type was determined before ERCP. Periampullary diverticulum, largest stone diameter, and stone count were assessed from preprocedural imaging. Stone diameter and stone count were applicable only to patients scheduled for stone extraction. For modeling purposes, these variables were coded as zero in patients not scheduled for stone extraction, with planned procedure type included to distinguish structural non-applicability from observed values.

Multidomain geriatric assessment comprised frailty, functional independence, nutritional status, and cognition. Frailty was assessed using the Clinical Frailty Scale (CFS), with higher scores indicating greater frailty [14]. Functional independence was assessed using the Barthel Index, with lower scores indicating greater dependence [15]. Nutritional status was evaluated using the Mini Nutritional Assessment–Short Form (MNA-SF), with lower scores indicating poorer nutritional status [16]. Cognitive status was assessed using the Mini-Cog, with lower scores indicating poorer cognitive performance [17].

Intraprocedural and early-recovery candidates included difficult biliary cannulation, precut sphincterotomy, procedure duration, and delayed awakening. Difficult cannulation was defined as > 5 papillary contacts, > 5 minutes of cannulation after papillary visualization, or > 1 unintended pancreatic duct cannulation or opacification [18]. Delayed awakening was defined as failure to regain adequate consciousness within 30 minutes after discontinuation of sedation or anesthesia, assessed during early recovery before the final discharge decision.

Sample-size considerations

All consecutive eligible patients during the prespecified study periods were included, and no prospective sample-size calculation was performed. The development and temporal validation cohorts contained 185 and 102 outcome events, respectively. Model complexity was constrained through penalization, feature selection, and nested cross-validation, and uncertainty in predictive performance was quantified using 95% confidence intervals.

Descriptive and explanatory analyses

Continuous variables are reported as median and interquartile range, and categorical variables as number and percentage. Groups were compared using the Mann–Whitney U test, Pearson chi-square test, or Fisher exact test, as appropriate. These comparisons were descriptive and were not used for predictor screening.

Prespecified covariates were entered into multivariable logistic regression to examine adjusted associations with unplanned inpatient admission. The preprocedural explanatory model included age, sex, planned procedure type, ASA physical status, comorbidity count, acute cholangitis, biliary pancreatitis, creatinine clearance, serum albumin, alanine aminotransferase, previous ERCP, previous pancreatitis, antithrombotic therapy, periampullary diverticulum, CFS, Barthel Index, MNA-SF, Mini-Cog, living alone, inadequate 24-hour caregiver availability, and travel time > 1 hour. The extended explanatory model additionally included difficult cannulation, precut sphincterotomy, procedure duration, and delayed awakening.

Adjusted odds ratios (ORs) with 95% confidence intervals (CIs) were reported irrespective of statistical significance. Clinically interpretable increments were used for continuous predictors. Associations were interpreted as conditional associations rather than causal effects.

Incremental predictive value

Three prespecified sequential logistic models were used to quantify the incremental predictive value of multidomain geriatric assessment. M1 included chronological age alone. M2 added conventional preprocedural demographic, clinical, laboratory, and procedural-planning factors. M3 further added CFS, Barthel Index, MNA-SF, and Mini-Cog.

Identical stratified 5-fold partitions were used for all three models. For each model, imputation and any required preprocessing were performed within the training fold and then applied to the corresponding held-out fold. Performance was calculated from pooled out-of-fold predictions using the area under the receiver-operating-characteristic curve (AUC), average precision (AP), and Brier score. Calibration plots and decision-curve analysis were also used to assess model calibration and potential clinical utility. Incremental differences for M2 versus M1 and M3 versus M2 were estimated using 1,500 paired bootstrap replicates.

For temporal validation, M1–M3 were refitted using the complete development cohort, with preprocessing parameters estimated exclusively from that cohort. The fitted models were then locked and applied unchanged to the temporal validation cohort without recalibration or model updating. The primary incremental comparison focused on M3 versus M2 to assess the additional predictive value of multidomain geriatric assessment beyond conventional preprocedural factors.

Preprocessing, feature selection, and model development

The outcome was not imputed. Missing continuous and categorical predictors were imputed using the training-set median and mode, respectively. Missing stone diameter and stone count values among patients scheduled for stone extraction were imputed using medians calculated from stone-extraction patients within the corresponding training partition. Structural non-applicability in other patients was handled separately and was not treated as ordinary missingness. For LASSO and L2-penalized logistic regression, predictors were standardized using means and standard deviations estimated from the training data. XGBoost used imputed, unstandardized predictors. All preprocessing parameters were estimated exclusively within the relevant training partitions and subsequently applied to held-out data.

Predictor selection for the preprocedural and extended models was performed separately using LASSO logistic regression. Within each outer training set, the penalty parameter was selected by inner stratified 5-fold cross-validation using the 1-standard-error rule. Feature selection was therefore repeated independently within each outer training fold.

Each algorithm included a preprocedural model, based only on information available before ERCP, and an extended model that additionally incorporated intraprocedural and early-recovery predictors available before the final discharge decision. The preprocedural model was intended to support initial patient selection and procedural planning, whereas the extended model was designed for dynamic risk reassessment during early postprocedural recovery, using information available before the final discharge decision.

Internal validation used nested stratified 5-fold cross-validation with 5-fold outer and inner loops. The outer loop was reserved for performance estimation, whereas the inner loop was used for feature selection and hyperparameter tuning. Imputation, scaling, feature selection, tuning, and model fitting were completed without access to the corresponding outer validation fold.

After internal validation, feature selection and hyperparameter optimization were repeated using the complete development cohort. Four final models were then refitted and locked: preprocedural and extended L2-penalized logistic regression models and preprocedural and extended XGBoost models. The locked models were applied unchanged to the temporal validation cohort, without predictor reselection, hyperparameter retuning, coefficient updating, recalibration, or other model modification.

Model performance and sensitivity analyses

Internal-validation performance was estimated from pooled outer-fold predictions, whereas temporal-validation performance was based on predictions from the locked models. Discrimination was assessed using AUC, precision-recall performance using AP, and overall prediction error using the Brier score. Calibration was evaluated using calibration plots, calibration intercepts, and calibration slopes. Clinical utility was assessed using decision-curve analysis across threshold probabilities of 0.05–0.35. Ninety-five percent CIs for AUC, AP, and Brier score were estimated using 1,500 bootstrap replicates.

Post hoc sensitivity analyses examined the counterintuitive coefficient directions observed for selected predictors within the preprocedural LASSO framework. Pairwise correlations and variance inflation factors were evaluated for age, ASA physical status, comorbidity count, CFS, Barthel Index, MNA-SF, and Mini-Cog. Two reduced candidate sets were examined: one excluding ASA physical status and comorbidity count, and another additionally excluding age. Predictive performance was compared using paired CIs, and coefficient directions were examined across outer folds.

The locked XGBoost models were interpreted in the temporal validation cohort using SHapley Additive exPlanations (SHAP). Mean absolute SHAP values summarized global feature importance, and beeswarm plots illustrated the direction and magnitude of individual feature contributions. SHAP values were interpreted as model explanations rather than causal effects.

All statistical analyses, model development, validation, and interpretation were performed in Python (version 3.13.5). Scikit-learn (version 1.8.0) was used for preprocessing, cross-validation, penalized logistic regression, and performance evaluation; XGBoost (version 3.1.3) for gradient-boosting models; statsmodels (version 0.14.6) for multivariable regression analyses; and SHAP (version 0.50.0) for model interpretation.

Results

Study population

Among 8,527 patients screened from January 2020 through August 2026, 1,407 eligible patients aged ≥ 65 years were included. The development cohort comprised 1,000 patients, of whom 185 (18.5%) required unplanned inpatient admission, whereas the temporal validation cohort comprised 407 patients, of whom 102 (25.1%) required admission (Fig. 1).

In the development cohort, patients requiring admission had higher CFS scores and lower Barthel Index, MNA-SF, and Mini-Cog scores than those completing the planned day-case pathway. Acute cholangitis and biliary pancreatitis were more frequent among admitted patients, whereas creatinine clearance and serum albumin were lower. Adequate 24-h caregiver availability was less frequent and travel time > 1 h was more common among admitted patients. Difficult biliary cannulation, longer procedure duration, and delayed awakening were also more frequent, whereas age, ASA physical status, and comorbidity count did not differ significantly in unadjusted comparisons (Table 1).

Table 1
Characteristics of the development cohort according to discharge status

Characteristic	Overall (n = 1,000)	Planned discharge (n = 815)	Unplanned admission (n = 185)	P value
Demographic and procedural planning				
Age, y	78.0 (77.0–91.0)	78.0 (77.0–91.0)	77.0 (76.0–92.0)	0.082
Male sex	526 (52.6)	424 (52.0)	102 (55.1)	0.444
Planned stone extraction	500 (50.0)	385 (47.2)	115 (62.2)	< .001
ASA physical status	3.0 (2.0–3.0)	3.0 (2.0–3.0)	3.0 (2.0–3.0)	0.944
Number of comorbidities	2.0 (1.0–3.0)	1.0 (1.0–3.0)	2.0 (1.0–3.0)	0.304
Multidomain geriatric assessment				
CFS score	3.0 (3.0–4.0)	3.0 (3.0–4.0)	4.0 (3.0–5.0)	< .001
Barthel Index	90.0 (85.0–95.0)	90.0 (85.0–95.0)	85.0 (75.0–90.0)	< .001
MNA-SF score	11.0 (10.0–12.0)	12.0 (10.0–13.0)	10.0 (9.0–12.0)	< .001
Mini-Cog score	4.0 (3.0–4.0)	4.0 (3.0–5.0)	3.0 (3.0–4.0)	< .001
Clinical and laboratory characteristics				
Acute cholangitis	279 (27.9)	205 (25.2)	74 (40.0)	< .001
Biliary pancreatitis	189 (18.9)	140 (17.2)	49 (26.5)	0.004
Creatinine clearance, mL/min	61.6 (44.5–75.7)	63.8 (47.3–76.8)	49.7 (36.2–70.8)	< .001

Continuous variables were compared using the Mann–Whitney U test, and categorical variables using the Pearson chi-square test or Fisher exact test, as appropriate. P values are descriptive and were not used for predictor selection or model development. Planned day-case discharge was defined as discharge on the day of ERCP or early the following morning within the intended 24-hour day-case pathway. CFS, Clinical Frailty Scale; MNA-SF, Mini Nutritional Assessment–Short Form; ERCP, endoscopic retrograde cholangiopancreatography. *Stone diameter and stone count were available only for patients scheduled for stone extraction (overall n = 500; planned day-case discharge n = 385; unplanned inpatient admission n = 115).

Characteristic	Overall (n = 1,000)	Planned discharge (n = 815)	Unplanned admission (n = 185)	P value
Demographic and procedural planning				
Serum albumin, g/L	36.2 (31.1–39.6)	36.5 (31.3–40.0)	34.8 (30.1–37.5)	< .001
Alanine aminotransferase, U/L	41.2 (25.6–74.9)	39.4 (24.9–69.7)	52.0 (27.4–83.7)	0.004
Previous ERCP	299 (29.9)	257 (31.5)	42 (22.7)	0.018
Previous pancreatitis	72 (7.2)	62 (7.6)	10 (5.4)	0.296
Antithrombotic therapy	193 (19.3)	159 (19.5)	34 (18.4)	0.725
Periampullary diverticulum	271 (27.1)	227 (27.9)	44 (23.8)	0.261
Largest stone diameter, mm*	8.7 (6.4–11.7)	8.7 (6.4–12.3)	8.5 (6.4–10.3)	0.114
Stone count*	2.0 (1.0–3.0)	2.0 (1.0–3.0)	2.0 (1.0–3.0)	0.472
Social and logistical characteristics				
Living alone	255 (25.5)	203 (24.9)	52 (28.1)	0.367
Adequate 24-h caregiver	776 (77.6)	660 (81.0)	116 (62.7)	< .001
Travel time > 1 h	195 (19.5)	146 (17.9)	49 (26.5)	0.008
Intraprocedural and early-recovery characteristics				
Difficult biliary cannulation	339 (33.9)	246 (30.2)	93 (50.3)	< .001
Precut sphincterotomy	155 (15.5)	119 (14.6)	36 (19.5)	0.099
Procedure duration, min	24.0 (17.0–32.0)	22.0 (16.0–30.0)	30.0 (23.0–38.0)	< .001
Delayed awakening	182 (18.2)	111 (13.6)	71 (38.4)	< .001

Continuous variables were compared using the Mann–Whitney U test, and categorical variables using the Pearson chi-square test or Fisher exact test, as appropriate. P values are descriptive and were not used for predictor selection or model development. Planned day-case discharge was defined as discharge on the day of ERCP or early the following morning within the intended 24-hour day-case pathway. CFS, Clinical Frailty Scale; MNA-SF, Mini Nutritional Assessment–Short Form; ERCP, endoscopic retrograde cholangiopancreatography. *Stone diameter and stone count were available only for patients scheduled for stone extraction (overall n = 500; planned day-case discharge n = 385; unplanned inpatient admission n = 115).

Representative endoscopic and fluoroscopic findings from planned day-case ERCP procedures, including stone extraction, biliary stenting, difficult cannulation, and precut access techniques, are shown in Supplementary Fig. S1.

Multivariable associations

In the preprocedural explanatory model, higher CFS scores (adjusted OR [aOR] per 1-point increase, 4.08; 95% CI, 3.09–5.39), lower Barthel Index scores (aOR per 10-point decrease, 2.22; 95% CI, 1.84–2.69), lower MNA-SF scores (aOR per 1-point decrease, 2.49; 95% CI, 2.07–2.98), and lower Mini-Cog scores (aOR per 1-point decrease, 2.00; 95% CI, 1.62–2.47) were significantly associated with higher odds of unplanned inpatient admission (all $P < 0.001$). Older age, greater comorbidity burden, planned stone extraction, acute cholangitis, biliary pancreatitis, reduced creatinine clearance, lower serum albumin, higher alanine aminotransferase levels, and travel time exceeding 1 hour were also associated with admission. Previous ERCP showed an inverse association (aOR, 0.30; 95% CI, 0.17–0.52; $P < 0.001$), whereas ASA physical status and the remaining preprocedural variables were not statistically significant after adjustment (Fig. 2A).

In the extended explanatory model, difficult cannulation (aOR, 3.52; 95% CI, 2.26–5.48; $P < 0.001$), longer procedure duration (aOR per 10-minute increase, 1.26; 95% CI, 1.03–1.55; $P = 0.027$), and delayed awakening (aOR, 2.95; 95% CI, 1.99–4.36; $P < 0.001$) were independently associated with increased odds of unplanned inpatient admission. By contrast, precut sphincterotomy was not significantly associated with admission after adjustment (aOR, 1.16; 95% CI, 0.74–1.81; $P = 0.522$; Fig. 2B).

Incremental predictive value

In the development cohort, predictive performance improved sequentially across the three models based on pooled out-of-fold predictions (Table 2; Fig. 3A–C). The AUC increased from 0.537 (95% CI, 0.488–0.589) for M1 to 0.773 (95% CI, 0.733–0.815) for M2 and 0.860 (95% CI, 0.824–0.895) for M3. Similarly, average precision (AP) increased from 0.216 to 0.480 and 0.690, respectively, while the Brier score decreased from 0.150 to 0.123 and 0.091. Compared with M1, M2 improved AUC by 0.236 (95% CI, 0.177–0.294) and AP by 0.264 (95% CI, 0.190–0.344), with a Brier score difference of − 0.027 (95% CI, − 0.039 to − 0.020). Further incorporation of multidomain geriatric assessment (M3 vs M2) increased AUC by 0.087 (95% CI, 0.055–0.120) and AP by 0.209 (95% CI, 0.151–0.263), with a Brier score difference of − 0.032 (95% CI, − 0.040 to − 0.021).

Table 2
Incremental predictive value of multidomain geriatric assessment in the development and temporal validation cohorts

Model / comparison	AUC (95% CI)	Average precision (95% CI)	Brier score (95% CI)
Development cohort (n = 1,000)			
Model performance			
M1: Age alone	0.537 (0.488–0.589)	0.216 (0.173–0.249)	0.150 (0.135–0.166)
M2: Age + conventional preprocedural factors	0.773 (0.733–0.815)	0.480 (0.424–0.572)	0.123 (0.108–0.135)
M3: M2 + multidomain geriatric assessment	0.860 (0.824–0.895)	0.690 (0.618–0.762)	0.091 (0.079–0.104)
Incremental comparison			
M2 vs M1	+ 0.236 (+ 0.177 to + 0.294)	+ 0.264 (+ 0.190 to + 0.344)	-0.027 (-0.039 to -0.020)
M3 vs M2	+ 0.087 (+ 0.055 to + 0.120)	+ 0.209 (+ 0.151 to + 0.263)	-0.032 (-0.040 to -0.021)
Temporal validation cohort (n = 407)			
Model performance			
M1: Age alone	0.647 (0.574–0.718)	0.291 (0.220–0.389)	0.146 (0.123–0.170)
M2: Age + conventional preprocedural factors	0.730 (0.664–0.793)	0.395 (0.296–0.517)	0.133 (0.114–0.153)
M3: M2 + multidomain geriatric assessment	0.830 (0.783–0.873)	0.552 (0.444–0.656)	0.118 (0.100–0.136)
Incremental comparison			
M2 vs M1	+ 0.083 (− 0.011 to + 0.177)	+ 0.104 (− 0.009 to + 0.225)	−0.013 (− 0.024 to − 0.002)

M1 included chronological age alone; M2 additionally included conventional preprocedural factors; M3 further incorporated the Clinical Frailty Scale, Barthel Index, Mini Nutritional Assessment–Short Form, and Mini-Cog. Development-cohort performance was estimated from pooled out-of-fold predictions using stratified fivefold cross-validation. Temporal-validation performance was evaluated using models refitted on the full development cohort and applied without recalibration or updating. Differences were calculated as the more comprehensive model minus the preceding model, with 95% CIs estimated from 1,500 paired bootstrap replicates. Higher AUC and AP and lower Brier scores indicate better predictive performance. AUC, area under the receiver operating characteristic curve; AP, average precision; CI, confidence interval.

Model / comparison	AUC (95% CI)	Average precision (95% CI)	Brier score (95% CI)
Development cohort (n = 1,000)			
M3 vs M2	+ 0.100 (+ 0.039 to + 0.164)	+ 0.157 (+ 0.031 to + 0.263)	−0.015 (− 0.026 to − 0.005)

M1 included chronological age alone; M2 additionally included conventional preprocedural factors; M3 further incorporated the Clinical Frailty Scale, Barthel Index, Mini Nutritional Assessment–Short Form, and Mini-Cog. Development-cohort performance was estimated from pooled out-of-fold predictions using stratified fivefold cross-validation. Temporal-validation performance was evaluated using models refitted on the full development cohort and applied without recalibration or updating. Differences were calculated as the more comprehensive model minus the preceding model, with 95% CIs estimated from 1,500 paired bootstrap replicates. Higher AUC and AP and lower Brier scores indicate better predictive performance. AUC, area under the receiver operating characteristic curve; AP, average precision; CI, confidence interval.

In the temporal validation cohort, the locked M2 and M3 models were evaluated to assess the incremental predictive value of multidomain geriatric assessment beyond conventional preprocedural factors (Table 2). Compared with M2, M3 achieved a higher AUC (0.830 [95% CI, 0.783–0.873] vs 0.730 [95% CI, 0.664–0.793]) and AP (0.552 [95% CI, 0.444–0.656] vs 0.395 [95% CI, 0.296–0.517]). The corresponding improvements were 0.100 (95% CI, 0.039–0.164) for AUC and 0.157 (95% CI, 0.031–0.263) for AP. The Brier score decreased from 0.133 to 0.118, with a difference of − 0.015 (95% CI, − 0.026 to − 0.005).

Together, these findings demonstrated that multidomain geriatric assessment provided incremental predictive information beyond chronological age and conventional preprocedural factors, with consistent improvements in discrimination and average precision across the development and temporal validation cohorts.

LASSO feature selection and internal validation

In the refit using the full development cohort, LASSO retained 15 preprocedural predictors at the penalty selected by the one-standard-error rule (λ_1SE): age, planned stone extraction, ASA physical status class, comorbidity count, CFS, Barthel Index, MNA-SF, Mini-Cog, biliary pancreatitis, creatinine clearance, serum albumin, previous ERCP, largest stone diameter, adequate 24-h caregiver availability, and travel time > 1 h (Supplementary Fig. S2A–C). Predictor-selection stability across the outer cross-validation folds is shown in Supplementary Fig. S2D.

Internal-validation AUCs were 0.862 for the preprocedural L2-logistic model, 0.859 for the extended L2-logistic model, 0.885 for the preprocedural XGBoost model, and 0.888 for the extended XGBoost model (Table 3). Corresponding AP values were 0.681, 0.748, 0.828, and 0.829, respectively, and Brier scores were 0.092, 0.084, 0.066, and 0.066. Calibration intercepts ranged from − 0.014 to 0.052 and calibration slopes from 0.949 to 1.148 (Table 3).

Table 3
Internal and temporal validation performance of the final prediction models

Validation setting	Model	AUC (95% CI)	Average precision (95% CI)	Brier score (95% CI)	Calibration intercept	Calibration slope
Internal validation	Preprocedural L2-logistic	0.862 (0.822–0.895)	0.681 (0.607–0.755)	0.092 (0.079–0.105)	−0.010	0.949
Internal validation	Extended L2-logistic	0.859 (0.820–0.893)	0.748 (0.690–0.801)	0.084 (0.073–0.097)	−0.014	0.966
Internal validation	Preprocedural XGBoost	0.885 (0.845–0.919)	0.828 (0.778–0.871)	0.066 (0.055–0.077)	0.052	1.120
Internal validation	Extended XGBoost	0.888 (0.854–0.920)	0.829 (0.780–0.872)	0.066 (0.055–0.077)	0.046	1.148
Temporal validation	Preprocedural L2-logistic	0.844 (0.799–0.886)	0.636 (0.547–0.719)	0.168 (0.147–0.190)	−2.087	0.529
Temporal validation	Extended L2-logistic	0.872 (0.829–0.910)	0.703 (0.613–0.778)	0.132 (0.113–0.153)	−1.736	0.640
Temporal validation	Preprocedural XGBoost	0.784 (0.727–0.838)	0.543 (0.449–0.642)	0.167 (0.145–0.187)	−1.532	0.562
Temporal validation	Extended XGBoost	0.860 (0.812–0.905)	0.722 (0.642–0.796)	0.112 (0.095–0.128)	−1.281	0.739

Internal validation performance was estimated from pooled outer-fold predictions obtained through nested stratified fivefold cross-validation with five outer and five inner folds. The final models were refitted using the complete development cohort and applied unchanged to the temporal validation cohort without further feature selection, hyperparameter tuning, coefficient updating, or recalibration. Perfect calibration corresponds to an intercept of 0 and a slope of 1. Ninety-five percent confidence intervals were estimated using 1,500 bootstrap replicates.

Incorporation of intraprocedural and early-recovery information improved some aspects of predictive performance during internal validation, particularly AP and Brier score for the logistic model, although AUC did not increase consistently across all model classes.

Temporal validation

In the temporal validation cohort, AUCs were 0.844 for the preprocedural L2-logistic model, 0.872 for the extended L2-logistic model, 0.784 for the preprocedural XGBoost model, and 0.860 for the extended XGBoost model (Table 3; Fig. 4A).

The extended L2-logistic model achieved the highest AUC, whereas the extended XGBoost model achieved the highest AP (0.722) and lowest Brier score (0.112) (Table 3; Fig. 4A–C). Compared with their respective preprocedural models, both extended models showed clearer gains in temporal validation, particularly for XGBoost.

Calibration deteriorated relative to internal validation. Calibration intercepts ranged from − 2.087 to − 1.281 and calibration slopes from 0.529 to 0.739, indicating systematic overprediction and overly extreme predicted probabilities (Table 3; Fig. 4C). Decision-curve analysis compared the net benefit of the four locked models across threshold probabilities of 0.05–0.35 (Fig. 4D). The models were applied to the temporal validation cohort without recalibration or model updating.

Sensitivity analyses and SHAP interpretation

Within the preprocedural LASSO sensitivity-analysis framework, pairwise correlations among age, ASA physical status, comorbidity count, CFS, Barthel Index, MNA-SF, and Mini-Cog were generally modest, and variance inflation factors ranged from 1.11 to 2.45, providing no evidence of substantial multicollinearity (Supplementary Fig. S3A).

Excluding ASA physical status and comorbidity count reduced the out-of-fold AUC from 0.850 to 0.824, and further exclusion of age reduced the AUC to 0.808. AP decreased and Brier score increased in parallel (Supplementary Fig. S3B). Although age and comorbidity count showed positive associations with unplanned inpatient admission in the multivariable explanatory analysis (Fig. 2A), their standardized coefficients were negative in the preprocedural LASSO models across the outer cross-validation folds (Supplementary Fig. S3C). ASA physical status also showed negative LASSO coefficients despite its nonsignificant association in the explanatory analysis. These differences in coefficient direction may reflect differences in model specification, predictor adjustment, and regularization and should not be interpreted as evidence of protective effects.

SHAP analyses identified the Barthel Index as the highest-ranked feature by mean absolute SHAP value in both XGBoost models (Fig. 5A,C). In the preprocedural model, the next highest-ranked features were age, creatinine clearance, planned stone extraction, albumin, and MNA-SF (Fig. 5A,B). In the extended model, age remained the second-ranked feature, followed by adequate 24-h caregiver availability, delayed awakening, MNA-SF, and CFS; procedure duration ranked eighth (Fig. 5C,D). After the models had been locked, SHAP analyses were performed using the temporal validation cohort solely for model interpretation. These analyses were not used for model development, feature selection, or performance estimation.

Discussion

Risk assessment for older adults undergoing planned day-case ERCP remains insufficiently characterized [1, 2]. Previous studies have focused primarily on ERCP-related adverse events and mortality [3, 10, 12], or on readmission after inpatient ERCP [11], with limited attention to unplanned

inpatient admission after intended day-case discharge [1, 2]. Chronological age alone cannot adequately capture differences in physiological reserve, functional capacity, and recovery potential among older patients of similar ages [5]. Likewise, conventional clinical indicators, including ASA physical status, summarize medical risk but do not directly measure frailty, functional dependence, nutritional impairment, or cognitive vulnerability, which may contribute additional information relevant to recovery and discharge readiness [4, 6]. This study demonstrates that multidomain geriatric assessment provides substantial predictive information beyond chronological age and conventional clinical factors, supporting a more individualized approach to risk stratification in older adults undergoing planned day-case ERCP.

Chronological age alone showed limited predictive value (AUC, 0.537), whereas the addition of conventional preprocedural factors increased the AUC to 0.773. Further incorporation of frailty, functional status, nutrition, and cognition increased the AUC to 0.860. This incremental value was maintained in temporal validation, with an improvement in AUC of 0.100 (95% CI, 0.039–0.164), an increase in average precision of 0.157 (95% CI, 0.031–0.263), and a reduction in Brier score of 0.015 (95% CI, 0.005–0.026). These consistent improvements across discrimination, precision-recall performance, and overall prediction accuracy reinforce the value of multidomain geriatric assessment beyond conventional risk factors. The findings are consistent with evidence linking frailty and geriatric vulnerability to adverse outcomes in older adults, including those undergoing ERCP [4, 5, 7, 11, 12].

Among the geriatric domains, frailty, functional dependence, nutritional impairment, and cognitive vulnerability were independently associated with unplanned inpatient admission. Higher CFS scores (aOR per 1-point increase, 4.08; 95% CI, 3.09–5.39), lower Barthel Index scores (aOR per 10-point decrease, 2.22; 95% CI, 1.84–2.69), poorer nutritional status (aOR per 1-point MNA-SF decrease, 2.49; 95% CI, 2.07–2.98), and lower cognitive performance (aOR per 1-point Mini-Cog decrease, 2.00; 95% CI, 1.62–2.47) were associated with increased odds of admission. The Barthel Index was also the highest-ranked predictor in both XGBoost models according to SHAP analysis, highlighting the importance of functional independence in discharge readiness. These observations are consistent with studies linking functional performance and geriatric assessment domains to hospital outcomes and discharge-related measures in other older patient populations [19–21], although those studies did not specifically examine day-case ERCP.

Age and comorbidity burden showed positive associations with unplanned inpatient admission in the explanatory regression but negative coefficients in the preprocedural LASSO models. These differences may reflect variations in model specification and regularization rather than substantial multicollinearity. Therefore, negative LASSO coefficients should not be interpreted as evidence that older age or greater comorbidity burden reduces admission risk.

The two-stage risk-assessment framework further highlights the value of updating discharge risk as procedural and recovery information becomes available. Procedural complexity and sedation-related complications are established considerations in ERCP safety assessment [10, 13, 18]. In our cohort,

difficult biliary cannulation (aOR, 3.52; 95% CI, 2.26–5.48), longer procedure duration (aOR per 10-minute increase, 1.26; 95% CI, 1.03–1.55), and delayed awakening (aOR, 2.95; 95% CI, 1.99–4.36) were associated with increased odds of unplanned admission. Incorporating these factors improved predictive performance in temporal validation, with AUC increasing from 0.844 to 0.872 for L2-penalized logistic regression and from 0.784 to 0.860 for XGBoost. The extended XGBoost model also achieved the highest average precision (0.722) and lowest Brier score (0.112). These findings support a dynamic assessment strategy that combines preprocedural patient selection with reassessment during early postprocedural recovery.

The clinical significance of this study lies in improving the selection of older adults suitable for planned day-case ERCP and identifying patients who may benefit from individualized perioperative management and closer postprocedural observation. Incorporating multidomain geriatric assessment into routine preprocedural evaluation may help identify vulnerable patients whose risks are not adequately captured by conventional clinical indicators, thereby supporting more informed patient selection and procedural planning. Further integration of intraprocedural and early-recovery information allows dynamic reassessment of discharge suitability and may help optimize postoperative monitoring and discharge decisions. These findings provide a clinically relevant framework for integrating comprehensive geriatric assessment into the day-case ERCP pathway.

This study is strengthened by its focus on an underexplored clinical outcome, incorporation of multidomain geriatric assessment, nested cross-validation, and temporal validation of locked models. Although the retrospective single-center design limits generalizability, the consistent incremental predictive value observed in temporal validation supports the relevance of multidomain geriatric assessment. Calibration deterioration, including systematic overprediction, indicates the need for recalibration before predicted probabilities are applied to individual discharge decisions [22–25]. Further external validation will help establish the applicability of these models across different clinical settings.

In conclusion, multidomain geriatric assessment meaningfully improved prediction of unplanned inpatient admission after planned day-case ERCP in older adults beyond chronological age and conventional preprocedural factors. Incorporating procedural complexity and early-recovery information further enhanced risk stratification in temporal validation. These findings support a personalized, staged approach to identifying older adults suitable for day-case ERCP and optimizing postprocedural monitoring and discharge planning.

Declarations

Competing interests

The authors declare no competing interests.

Funding

The authors declare that no funding was received for the conduct of this study or the preparation of this manuscript.

Author Contribution

Xiaoli Sun and Lei Zhao conceived and designed the study. Xiaoli Sun performed the primary data analysis, prepared the figures and tables, and drafted the manuscript. Xiuli Liu contributed to clinical data collection and data curation. Jing Chen contributed to data collection and verification. Jitao Song contributed to clinical data compilation, statistical analysis, and validation of the results. Cuihua Li contributed to clinical data acquisition and data management. Hao Yang contributed to methodology, data verification, and visualization. Lina Liu contributed to data curation and critical revision of the manuscript. Lei Zhao supervised the study, provided methodological and clinical oversight, and critically reviewed and revised the manuscript. All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Data Availability

The datasets generated and analysed during the current study are not publicly available due to patient privacy and institutional data protection requirements. De-identified data may be made available from the corresponding author upon reasonable request, subject to approval by the relevant institutional authorities and applicable ethical and data-sharing regulations.

References

1. Freeman, M. L. et al. Same-day discharge after endoscopic biliary sphincterotomy: observations from a prospective multicenter complication study. *Gastrointest. Endosc*. **49**, 580–586. https://doi.org/10.1016/S0016-5107(99)70385-8 (1999).
2. Sperna Weiland, C. J. et al. Endoscopic retrograde cholangiopancreatography discharge tool combined with rapid trypsinogen-2 test to predict same-day discharge: a prospective cohort study. *Eur. J. Gastroenterol. Hepatol.* **37**, 1206–1212. https://doi.org/10.1097/MEG.0000000000003014 (2025).
3. Day, L. W., Lin, L. & Somsouk, M. Adverse events in older patients undergoing ERCP: a systematic review and meta-analysis. *Endosc Int. Open.* **2**, E28–E36. https://doi.org/10.1055/s-0034-1365281 (2014).
4. Ramai, D. et al. Influence of frailty in patients undergoing endoscopic retrograde cholangiopancreatography for biliary stone disease: a nationwide study. *Dig. Dis. Sci.* **68**, 3605–3613. https://doi.org/10.1007/s10620-023-08013-0 (2023).

5. Clegg, A., Young, J., Iliffe, S., Rikkert, O., Rockwood, K. & M. G. M. & Frailty in elderly people. *Lancet* **381**, 752–762. https://doi.org/10.1016/S0140-6736(12)62167-9 (2013).
6. Veronese, N. et al. Comprehensive geriatric assessment in older people: an umbrella review of health outcomes. *Age Ageing*. **51**, afac104. https://doi.org/10.1093/ageing/afac104 (2022).
7. Lin, H. S., Watts, J. N., Peel, N. M. & Hubbard, R. E. Frailty and post-operative outcomes in older surgical patients: a systematic review. *BMC Geriatr.* **16,** 157. https://doi.org/10.1186/s12877-016-0329-8 (2016).
8. Ellis, G. et al. Comprehensive geriatric assessment for older adults admitted to hospital. Cochrane Database Syst. Rev. CD006211 (2017). (2017). https://doi.org/10.1002/14651858.CD006211.pub3
9. Arabpour, E. et al. A network meta-analysis of optimal strategies for preventing post-endoscopic retrograde cholangiopancreatography pancreatitis. *Sci. Rep.* **15**, 13702. https://doi.org/10.1038/s41598-025-98969-y (2025).
10. Prosenz, J. et al. A prospective study on incidence of desaturations in ERCP with non-anesthesiologist sedation and adverse event awareness of endoscopists. *Sci. Rep.* **15**, 22781. https://doi.org/10.1038/s41598-025-04922-4 (2025).
11. Le, K. H. N. et al. The hospital frailty risk score as a predictor of readmission after ERCP. *Surg. Endosc*. **38**, 260–269. https://doi.org/10.1007/s00464-023-10531-x (2024).
12. Suzuki, M. et al. Association of the Geriatric-8 Score With In-Hospital Mortality in Older Patients Undergoing Endoscopic Retrograde Cholangiopancreatography. *Geriatr. Gerontol. Int.* **26**, e70609. https://doi.org/10.1111/ggi.70609 (2026).
13. Dumonceau, J. M. et al. ERCP-related adverse events: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. *Endoscopy* **52**, 127–149. https://doi.org/10.1055/a-1075-4080 (2020).
14. Rockwood, K. et al. A global clinical measure of fitness and frailty in elderly people. *CMAJ* **173**, 489–495. https://doi.org/10.1503/cmaj.050051 (2005).
15. Mahoney, F. I. & Barthel, D. W. Functional evaluation: the Barthel Index. *Md. State Med. J.* **14**, 61–65 (1965).
16. Kaiser, M. J. et al. Validation of the Mini Nutritional Assessment short-form (MNA-SF): a practical tool for identification of nutritional status. *J. Nutr. Health Aging*. **13**, 782–788. https://doi.org/10.1007/s12603-009-0214-7 (2009).
17. Borson, S., Scanlan, J. M., Chen, P. & Ganguli, M. The Mini-Cog as a screen for dementia: validation in a population-based sample. *J. Am. Geriatr. Soc.* **51**, 1451–1454. https://doi.org/10.1046/j.1532-5415.2003.51465.x (2003).
18. Testoni, P. A. et al. Papillary cannulation and sphincterotomy techniques at ERCP: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline. *Endoscopy* **48**, 657–683. https://doi.org/10.1055/s-0042-108641 (2016).
19. Ocagli, H., Cella, N., Stivanello, L., Degan, M. & Canova, C. The Barthel index as an indicator of hospital outcomes: a retrospective cross-sectional study with healthcare data from older people. *J. Adv. Nurs.* **77**, 1751–1761. https://doi.org/10.1111/jan.14708 (2021).

20. D'Souza, A. N. et al. Predicting discharge destination in older people from acute general medical wards: a systematic review of the psychometric properties of 23 assessment tools. *J. Geriatr. Phys. Ther.* **47**, E109–E123. https://doi.org/10.1519/JPT.0000000000000401 (2024).
21. Bruijnen, C. P., van Harten-Krouwel, D. G., Koldenhof, J. J., Emmelot-Vonk, M. H. & Witteveen, P. O. Predictive value of each geriatric assessment domain for older patients with cancer: a systematic review. *J. Geriatr. Oncol.* **10**, 859–873. https://doi.org/10.1016/j.jgo.2019.02.010 (2019).
22. Van Calster, B. et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.* **74**, 167–176. https://doi.org/10.1016/j.jclinepi.2015.12.005 (2016).
23. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931. https://doi.org/10.1093/eurheartj/ehu207 (2014).
24. Zhang, J. & Li, T. Discrimination stability and calibration of cardiovascular risk prediction models in the Framingham baseline cohort. *Sci. Rep.* **16**, 23965. https://doi.org/10.1038/s41598-026-54869-3 (2026).
25. Binuya, M. A. E., Engelhardt, E. G., Schats, W., Schmidt, M. K. & Steyerberg, E. W. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med. Res. Methodol.* **22**, 316. https://doi.org/10.1186/s12874-022-01801-8 (2022).

Figures

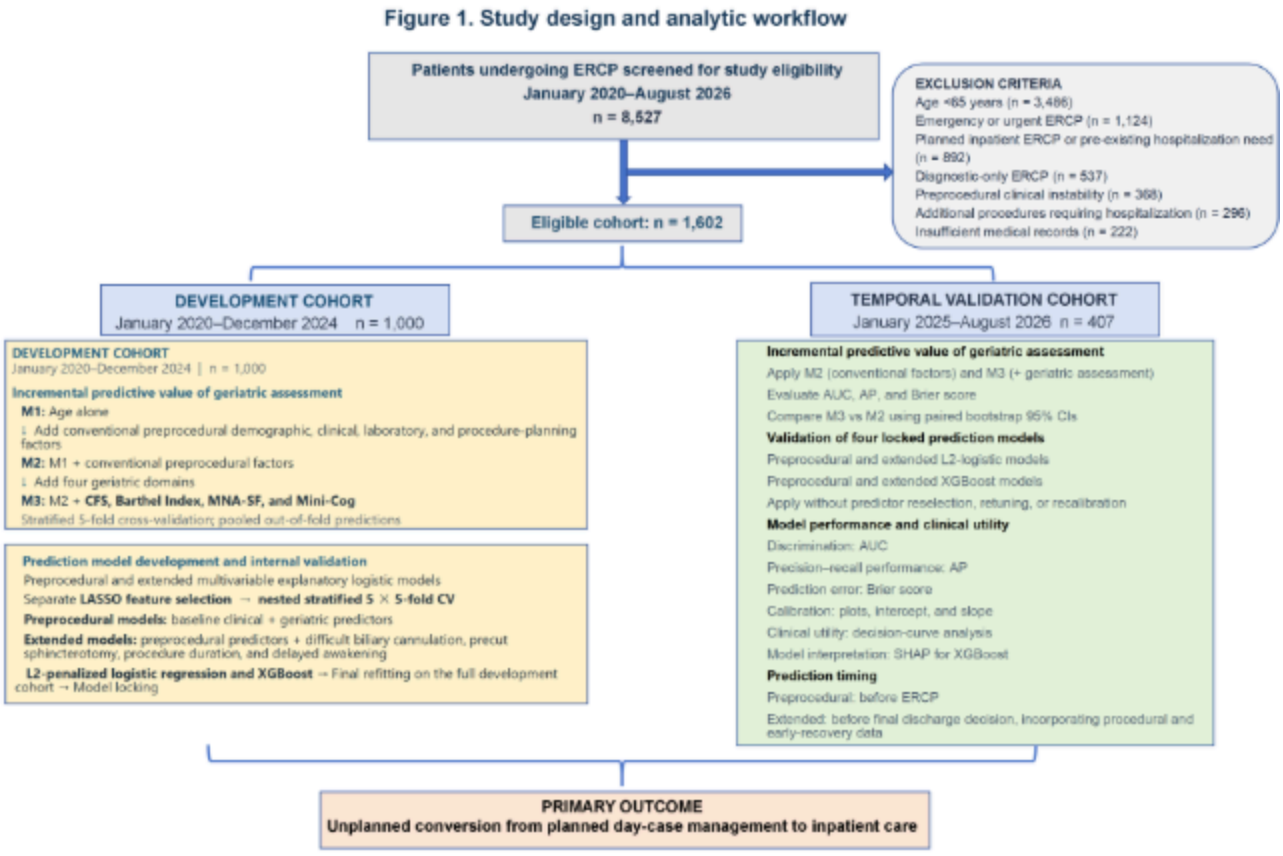
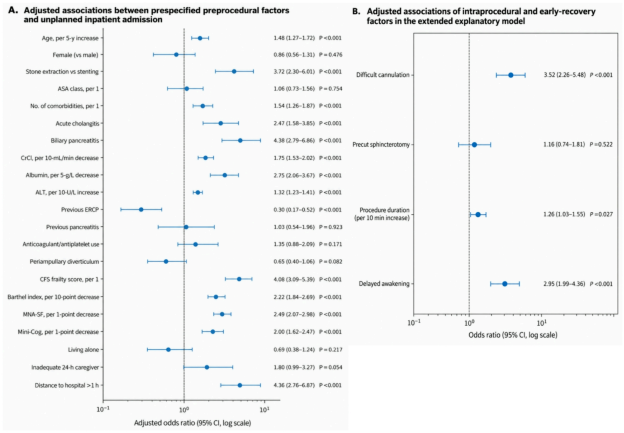


Figure 1

Study design and analytic workflow.



(A) (B)

s) s)

Figure 2

Multivariable clinical factors associated with unplanned inpatient admission following planned day-case ERCP.

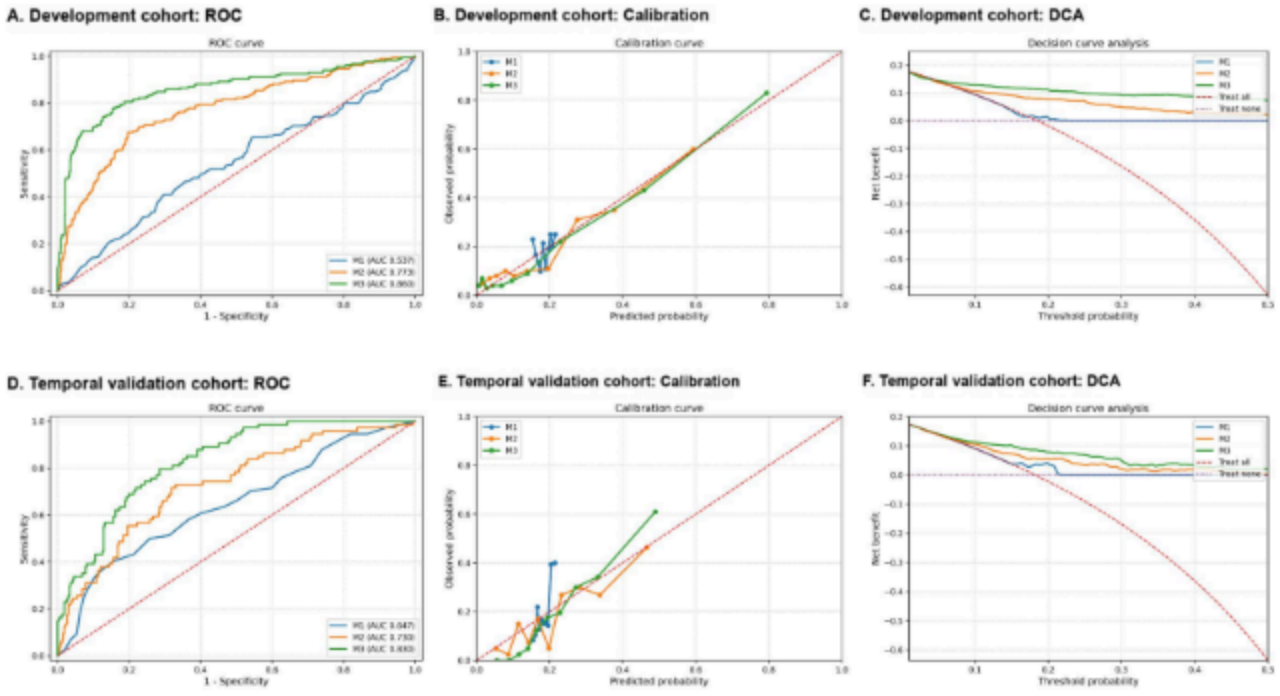


Figure 3

Discrimination, calibration, and decision curve analysis of prediction models in the development and temporal validation cohorts.

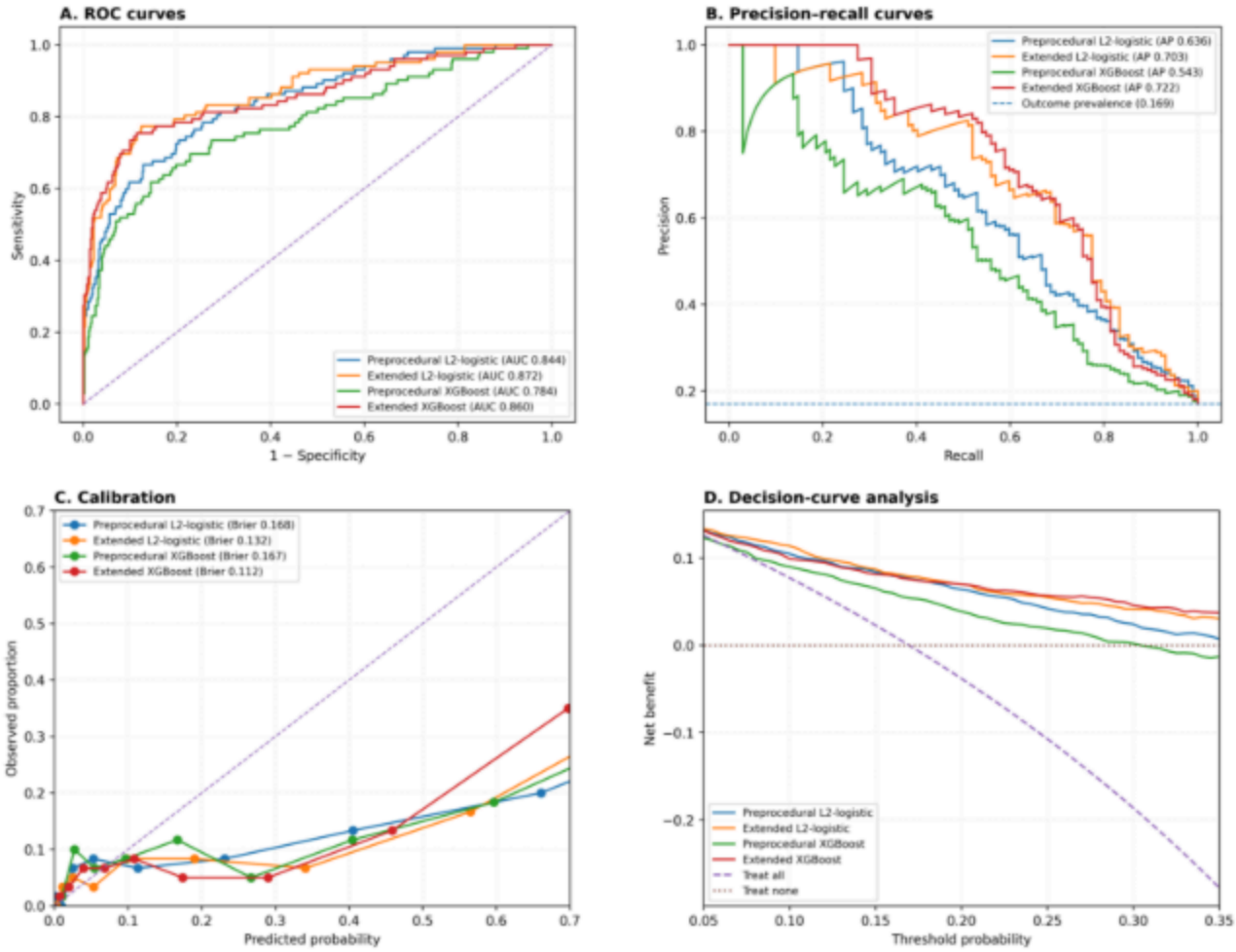


Figure 4

Temporal validation of the locked prediction models.

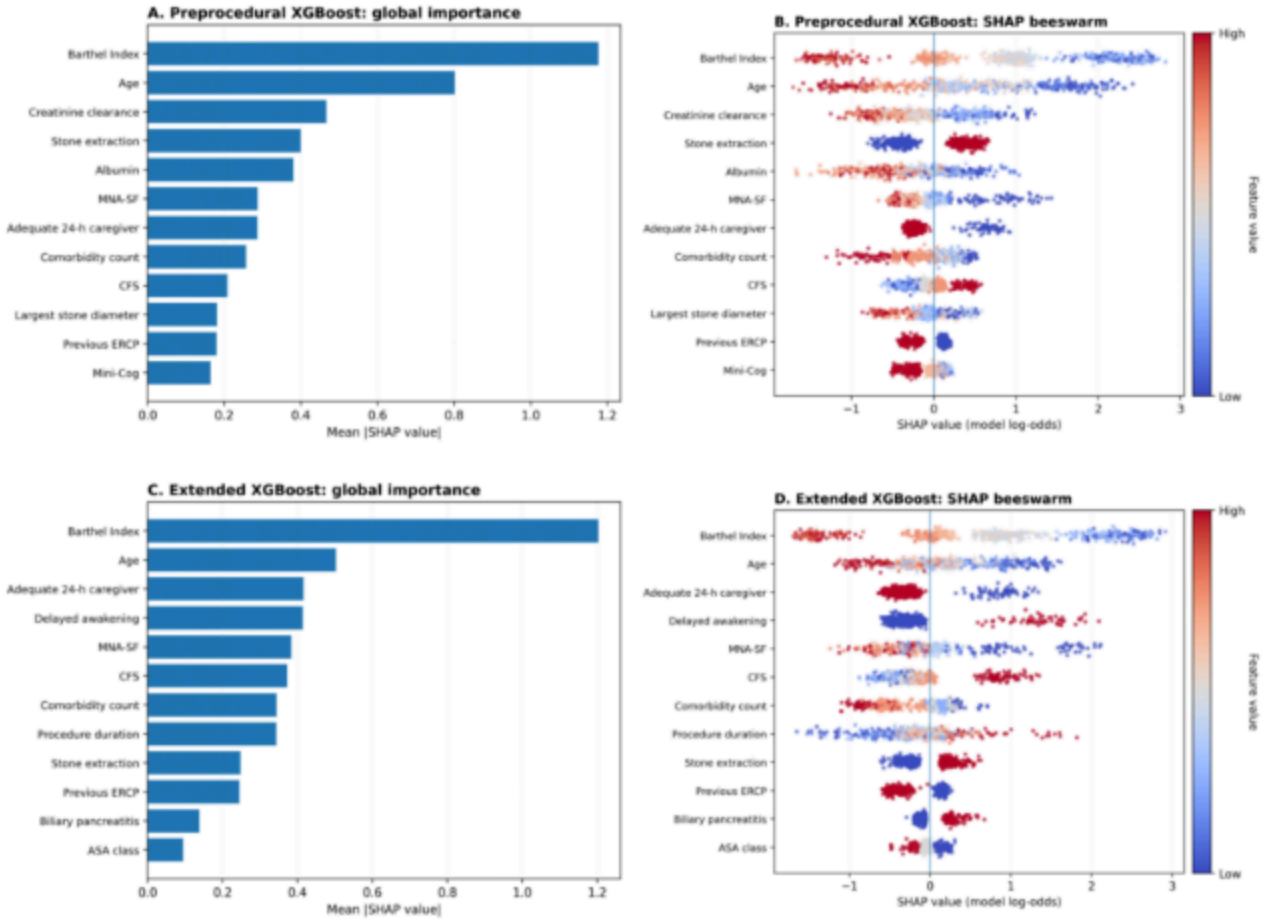


Figure 5

SHAP-based interpretation of the locked XGBoost models in the temporal validation cohort.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx