

Dynamic prediction of severe complications after pancreatic resection: incremental value of the latest laboratory results and longitudinal history

Jingyun Ning

ningjy5@mail2.sysu.edu.cn

Sun Yat-Sen Memorial Hospital

Ruiqi Li

Sun Yat-Sen Memorial Hospital

Huihuang Yu

Sun Yat-Sen Memorial Hospital

Jianfeng Huang

Sun Yat-Sen Memorial Hospital

Haoming Lin

Sun Yat-Sen Memorial Hospital

Rui Zhang

Sun Yat-Sen Memorial Hospital

Research Article

Keywords: Pancreatic resection, severe complications, electronic health records, dynamic prediction, landmark analysis, informative missingness, temporal validation

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11194132/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

Background

Whether complete laboratory histories improve postoperative risk prediction beyond the latest results requires direct evaluation. We compared laboratory representations for predicting severe complications after pancreatic resection within a common prediction and validation framework to inform feature selection for clinical information systems.

Methods

This exploratory retrospective study included 1,423 patients undergoing pancreatic resection in 2012–2026; 207 had Clavien–Dindo $\geq$ IIIa events within 30 days. End-of-day landmarks at postoperative days 0, 1, 3, 5, and 7 predicted first events through day 30. Static features (M0), latest laboratory values (M1), additional historical summaries (M2), and daily histories (M3) were evaluated using L2-regularized logistic regression and XGBoost. Patient-grouped 5×4-fold nested cross-validation assessed stratified c-statistic differences for M3 – M0 and M3 – M1. Temporal validation, calibration, and missingness analyses were performed. 95% confidence intervals used 2,000 paired patient-cluster bootstrap samples of fixed out-of-fold predictions.

Results

The prediction pool comprised 1,420 patients and 204 unique events. XGBoost stratified c-statistics for M0–M3 were 0.563, 0.611, 0.610, and 0.597. The exploratory secondary difference between latest-value and static models was 0.048 (95% CI 0.007 to 0.091). Primary M3 – M0 and M3 – M1 differences were 0.034 (– 0.017 to 0.087) and – 0.015 (– 0.061 to 0.033). At POD3, the M3 – M0 average-precision difference was 0.047 (0.010 to 0.104). In the 2024 and 2025-onward temporal tests, M1 c-statistics were 0.654 and 0.693; M3 – M1 differences were 0.020 (– 0.087 to 0.113) and – 0.102 (– 0.208 to – 0.004). The two primary logistic-regression contrasts were not estimated under the project-specific probability-dispersion rule.

Conclusions

Laboratory information provided incremental predictive signals beyond static features, supporting latest values as a useful comparator. Daily histories did not show consistent additional gains. These findings inform representation selection and support evaluating historical-feature increments, temporal stability, and clinical utility in independent data.

1 Background

The risk of complications after pancreatic resection evolves during recovery, making newly available clinical information relevant to repeated risk assessment. Laboratory results accumulate during routine care and can be incorporated into electronic health record prediction workflows. For clinical information systems, the question concerns both whether postoperative tests should be included and how they should be represented: are the latest results sufficient, or should historical levels, trends, and daily measurements also be retained? Quantifying these increments can help prioritize subsequent model development and validation.

Postoperative dynamic prediction has an established research basis. Varghese et al. used electronic medical record information from postoperative days 0–3 to predict early discharge after pancreatic surgery and also examined 90-day readmission [1]. Guillen-Ramirez et al. incorporated postoperative laboratory values, change features, and missing-data strategies into infection prediction across surgical procedures [2]. These studies support the value of updating risk estimates from postoperative records and show that combining dynamic laboratory data with machine learning is already an active research area. For a composite endpoint defined by complication severity, the additional information in daily histories beyond the latest laboratory values still warrants direct assessment at a clearly specified prediction cutoff.

In postoperative intensive care, Sheikhalishahi et al. compared progressively longer observation windows and different prediction windows for transfusion demand, showing that information accumulation and prediction timing affect model performance [3]. Our comparisons at shared landmarks further distinguish the value of newly available measurements from that of retaining earlier history.

Previous studies also suggest that model complexity does not have a fixed relationship with predictive performance. Ruan et al. compared GRU-D architectures using electronic health records from colorectal surgery and found that gains varied by outcome and prediction time, with no general advantage for more complex architectures [4]. In a multicenter pancreatoduodenectomy study, Ingwersen et al. found no superiority of the evaluated machine learning methods over logistic regression [5]. Evaluating longitudinal data therefore requires clinically relevant comparators and separate consideration of information representation and algorithm choice.

Whether a laboratory test is ordered and the interval between repeat measurements may also reflect clinical concern. Methods such as GRU-D use missingness indicators and elapsed times to represent irregular sequences [6]. Missing-data strategies for prediction must consider the information carried by measurement behavior and the conditions under which it can be used [7, 8]. Measurement workflows can change over a long study period, making temporal validation an important complement to internal evaluation across the full cohort.

Using longitudinal laboratory data from electronic health records, Sun et al. found that the performance of missing-data strategies depended on temporal representation and model architecture; carrying the last observation forward did not have the same effect in cross-sectional and sequential models [9]. This

supports evaluating the incremental contribution of observation indicators and measurement intervals alongside the measured values.

Using Clavien–Dindo $\geq$ IIIa events after pancreatic resection as the clinical setting, we compared static features, the latest laboratory values, historical summaries, and daily histories. Within common landmark risk sets and patient-grouped validation, we assessed incremental information alongside calibration, internal temporal validation, and explicit missingness representations to identify features warranting further evaluation. Outcomes were defined by events captured in the available clinical records and adjudication process.

2 Methods

2.1 Study design and cohort

This single-center retrospective cohort study used an existing cohort of index pancreatic resections at Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China. The final cohort included operations performed in 2012–2026 and comprised 1,423 patients, 1,423 index operations, and the corresponding 1,423 index encounters. The cohort was constructed in two stages. In the first stage, 1,860 candidate surgical records were screened, retaining each patient's first eligible operation to form an initial index cohort of 1,247 patients. Records requiring verification were reviewed, eligible records were reinstated, and the first eligible operation was selected across the reinstated and previously included records. This produced a cohort of 1,320 patients, a net increase of 73 over the initial cohort. In the second stage, an audit of date completeness recovered 116 eligible surgical records. After these records were combined with the existing cohort, the first-operation selection was repeated and 13 records of non-first operations were removed, yielding the final index cohort of 1,423 patients. Screening and revision records were retained separately for the two stages. The present prediction analysis used this cohort, verified the correspondence among patients, index encounters, and operations, and restricted laboratory data to the index encounter. Patients were not reselected, nor was the clinical outcome changed, on the basis of model performance.

The index operation was the unit of observation, and the patient was the clustering unit for validation grouping and resampling. Although each patient currently contributed only one index operation, records from the same patient at multiple landmarks remained repeated observations. Postoperative length of stay and discharge date were used for cohort description or selection of the inpatient subset. Outcome fields were used to define target labels and landmark risk sets. None of these fields was used directly as a model predictor.

The original analysis plan and subsequent revision records were retained separately. The information representations, primary comparisons, and internal temporal validation windows were revised and finalized on September 10, 2026, after earlier results were available; the study was therefore positioned as an exploratory methodological comparison. Reporting was informed by TRIPOD + AI, with PROBAST +

AI used to consider potential bias in data sources, outcome measurement, and performance evaluation [10, 11].

2.2 Outcome definition and clinical adjudication

The target outcome was the first severe complication of Clavien–Dindo grade $\geq$ IIIa within 30 days after the index operation, graded according to the Clavien–Dindo classification [12]. The investigator performed manual clinical adjudication using medical orders and relevant clinical records, with programmatic support for organizing evidence, extracting labels, and checking date consistency. The present prediction analysis used the existing outcome version produced by this process.

Event dates were anchored according to the project rules: the procedure date was used for discrete therapeutic interventions; 48 hours after initiation was used for ventilation; 24 hours after first use was used for vasopressors; and the initiation time was used for renal replacement therapy. ICU admission alone did not constitute a grade IV event. Purely diagnostic angiography or endoscopy did not constitute a grade III event, and feeding-tube placement was classified as grade II under the existing rules. These date definitions were used to maintain consistency between outcome timing and prediction cutoffs.

The outcome observation window extended to 30 days after the index operation. Medical orders and clinical records in the research database were used to identify severe complications and assign outcome labels. The study reports events identified and adjudicated within the scope of these records. Event dates were used to construct each landmark risk set and its subsequent prediction outcome.

2.3 Landmark times and available information

A landmark approach was used to update risk at specified times among patients who had not yet experienced the target event [13]. Landmarks were defined at the end of calendar postoperative days (PODs) 0, 1, 3, 5, and 7. Only patients who remained free of the target event through the end of the corresponding day were retained. Each prediction window extended from after that landmark to the common endpoint of POD30; patients with an event on the landmark day were excluded from that risk set. Each prediction used only information reported by the end of the landmark day. The prediction windows were not extended by a further 30 days from each landmark.

Laboratory timing was calculated as the difference between the report date and the operation date, at calendar-day resolution rather than in exact hours from completion of surgery. The primary risk sets retained patients who had been discharged but remained free of the target event. A sensitivity analysis was conducted among patients still hospitalized at the end of the landmark day; patients were considered discharged by the end of their discharge day. Because the remaining populations and event rates differed across landmarks, changes in absolute performance across landmarks were not interpreted as paired information gains in the same population.

2.4 Laboratory data extraction and information representations

Laboratory data were obtained only from records linking each patient to the exact index encounter, within the window from POD – 7 to the end of each landmark day. Test names were harmonized using the existing project dictionary, and specimen and test exclusion rules were applied. Multiple values for the same test on the same calendar day were summarized by their median. After excluding N-terminal pro-B-type natriuretic peptide, which had previously been designated for exclusion from modeling, 53 laboratory tests were used. The latest value was the daily value from the most recent observed date within the window. Historical summaries included the mean, minimum, maximum, and least-squares slope of daily values against POD; calculation of a slope required observations on at least two distinct dates.

All models shared age, sex, procedure type, surgical approach, and concomitant major surgery fields, together with the same categorical encoding of landmark time. The four information representations were expanded sequentially, as shown in Table 1. M3 flattened the 15-day grid from POD – 7 to POD7 into tabular features, with all cells after the relevant landmark set to missing. The latest-value summary did not imply forward-filling the entire daily grid. M3 was not a neural sequence model such as GRU-D or a Transformer. Category vocabularies were learned only within the corresponding training portion, and categories unseen during training were handled by the existing encoding procedure.

Table 1
Information representations and feature composition

Feature set	Information added to the preceding level	Numeric features	Purpose
M0	Age, concomitant major surgery, and shared categorical fields	2	Static features with landmark encoding
M1	Latest value of each of 53 laboratory tests	55	Increment from current laboratory information
M2	Mean, minimum, maximum, and slope for each test	267	Increment from historical summaries
M3	15-day daily-value grid for 53 tests	1062	Increment from daily histories
Note: Numeric feature counts exclude one-hot encoded columns for sex, procedure type, surgical approach, and landmark; the final number of encoded columns depends on the categories present in the training data. M3 includes all M2 features, and daily-grid cells after the landmark contain no future observations from that patient.			

To account for previously flagged platform changes, procalcitonin and drain-fluid amylase were divided into the periods 2012–2017, 2018–2022, and 2023 onward and scaled using the mean and standard deviation from the corresponding training portion. If a period contained no training observations, the overall training statistics for that test were used as the sole fallback; test data did not contribute to estimation. Laboratory values were extracted on their source numeric scales. The current implementation did not complete test-specific unit conversion and harmonization checks or preserve modifiers indicating values below or above assay limits. Period-specific scaling is a statistical procedure

and does not replace verification of laboratory-unit consistency. Further details of preprocessing and model implementation are provided in Additional file 1, Supplementary methods.

2.5 Algorithms and processing within training data

Each representation was evaluated using L2-regularized logistic regression (hereafter Ridge) and XGBoost, yielding 8 primary combinations. Ridge used median imputation and standardization based on the training portion and the lbfgs solver. XGBoost retained missing values for native handling and used the histogram-based tree method. Neither algorithm family used class weights.

The primary analysis used patient-grouped nested cross-validation with 5 outer folds and 4 inner folds. All landmark records from a given patient remained in the same fold, and all representations and algorithms shared the outer-fold assignments. Hyperparameters were selected independently within each outer training set. Categorical encoding, period-specific scaling, imputation, and other preprocessing were restricted to the corresponding training portion. Ridge evaluated 6 values of the regularization parameter C, ranging from 0.0001 to 10. XGBoost evaluated a fixed set of 20 candidate configurations with 150, 250, or 350 trees. The outer test sets were not used for early stopping or expansion of the search. Inner-loop hyperparameter selection used the stratified c-statistic, based on comparisons within the same landmark.

After hyperparameter selection, models were refitted to the corresponding outer training data and used to predict outcomes in the outer test patients. XGBoost used seed 42 for inner-loop candidates, while final outer-fold predictions were the mean probabilities across seeds 42–46. The deterministic Ridge models were not ensembled by repeating identical predictions. Raw outer-fold held-out probabilities were used for the primary results. Supplementary calibration used the Platt method: the calibrator was fitted to inner out-of-fold probabilities from the outer training patients without using outer test outcomes. For XGBoost, the calibrator was fitted to single-seed inner-loop probabilities and then applied to the outer-fold five-seed mean. The evaluation in this study reflects that implemented procedure. The candidate and selected hyperparameters are reported in Additional file 1, Tables S2 and S3.

2.6 Performance evaluation and statistical analysis

The primary discrimination metric was the stratified c-statistic, defined as the weighted mean of landmark-specific AUROCs, with weights equal to the number of case–non-case pairs at each landmark; only within-landmark pairs contributed. M3 – M0 and M3 – M1 within each algorithm family were designated as the four primary information contrasts. POD3 was selected as the main presentation time point. Landmark-specific AUROC, non-interpolated average precision (AP), Brier score, calibration-in-the-large, and the jointly estimated calibration slope were also reported. The reference for AP was the event rate in the corresponding risk set. AP was distinguished from the area under the precision–recall curve calculated by trapezoidal integration [14]. Calibration assesses agreement between predicted probabilities and observed outcomes and cannot be replaced by AUROC [15].

The existing project rule for prediction dispersion was retained. If the standard deviation of predicted logits for a model at a landmark was below 0.30, that model–landmark combination was excluded from the corresponding paired inference. If any constituent landmark was ineligible, the overall stratified c-statistic was retained only as a descriptive estimate, and the related overall contrasts were not estimated under the rule, rather than being mathematically incalculable. Calibration did not restore eligibility lost on the basis of raw predictions; eligibility after calibration was recorded separately. This threshold was an operational diagnostic criterion for the project, not a universal boundary for clinical validity.

95% confidence intervals were obtained from 2,000 paired patient-cluster bootstrap samples drawn with replacement, retaining all landmark records from the same patient and the contributions of repeatedly sampled patients. These intervals were conditional on fixed outer-fold held-out predictions and did not capture the full development uncertainty arising from repeated hyperparameter selection and model fitting. Differences were calculated as the model minus the reference. Higher stratified c-statistic, AUROC, and AP values indicate better performance, whereas lower Brier scores indicate better performance. The intervals were not adjusted for multiple comparisons. The four primary information contrasts were reported separately from other exploratory comparisons.

2.7 Temporal validation and supplementary analyses

Two temporal windows from the same center were used: training through 2023 with testing in 2024, and training through 2024 with testing in 2025 onward. Each window compared M1 with M3 in both algorithm families. Preprocessing, inner-loop hyperparameter selection, and calibration used only historical training data. These windows were selected after earlier results were available and were treated as retrospective internal temporal validations to examine model performance in later patients.

The primary M3 did not additionally include explicit missingness indicators or time since the last observation, although XGBoost's native missing-value branches and Ridge imputation could still carry information about missingness. Sensitivity analyses separately added daily observation indicators (mask) and observation indicators plus time intervals (mask + elapsed). The interval represented the number of days since the most recent observation as of the relevant day. When no prior observation was available within the window, the start of the window was used as the fallback; channels after the landmark carried no future information. Additional comparisons used an M2 version with unscaled values for the two tests subject to period-specific scaling and the subset of patients still hospitalized at the end of the landmark day.

Supplementary evaluations included diagnostic analyses using 20 patient-level permutations of event times under a fixed configuration, an independent repeat run using the same configuration, descriptions of annual missingness, and assessment of the stability of feature attributions across outer folds. Risk sets were reconstructed after permutation, but the full candidate search was not repeated; no confirmatory P value for predictive signal was calculated from this procedure. SHAP was used to describe outer-fold prediction attributions for specified members of the raw XGBoost models. Decision

curves were used to explore net benefit at threshold probabilities of 2%, 5%, 10%, 15%, and 20% [16]. Supplementary methods and diagnostic results are provided in Additional file 1.

3 Results

3.1 Cohort and landmark risk sets

The median age was 59 years (quartiles: 49 and 67), and 798 patients (56.1%) were male. There were 207 CD $\geq$ IIIa events within 30 postoperative days (14.5%), including 22 recorded deaths. Procedure type and surgical approach were unknown in 33 and 31 patients, respectively. Cohort characteristics are shown in Table 2. The 3 patients with a target event on POD0 were excluded from prediction beyond that day's end; all 3 were retained in the full-cohort description. Missing and unknown cohort characteristics are summarized in Additional file 1, Table S1.

Table 2
Characteristics of the full cohort and groups defined by severe events

Characteristic	Full cohort (n = 1423)	No recorded event (n = 1216)	Recorded event (n = 207)
Age, years, median [Q1, Q3]	59 [49, 67]	59 [48, 66]	61 [52, 69]
Male, n	798	678	120
Female, n	625	538	87
Pancreatoduodenectomy (PD), n	755	633	122
Distal pancreatectomy with splenectomy (DPS), n	321	285	36
Distal pancreatectomy without splenectomy (DP), n	111	101	10
Unknown procedure, n	33	29	4
Laparoscopic approach, n	742	660	82
Open approach, n	504	410	94
Conversion to open surgery, n	96	76	20
Robotic approach, n	50	43	7
Unknown approach, n	31	27	4
Postoperative stay, days, median [Q1, Q3]	14 [10, 22]	13 [9, 19]	24 [15, 35.5]
Recorded deaths by POD30, n	22	0	22
Note: The listed procedures are the main categories and do not exhaust all procedure types. PD includes pylorus-preserving pancreatoduodenectomy under the current coding scheme. DPS denotes distal pancreatectomy with splenectomy; DP denotes distal pancreatectomy without splenectomy. Postoperative length of stay was calculated as discharge POD + 1 and was used for outcome-related description, not as a predictor.			

The POD0 risk set contained 1,420 patients and 204 subsequent events; corresponding counts were 1,385 and 169 at POD1, 1,323 and 107 at POD3, 1,317 and 101 at POD5, and 1,302 and 86 at POD7. Event proportions were 14.37%, 12.20%, 8.09%, 7.67%, and 6.61%, respectively. Across all landmarks, there were 6,747 rows and 667 event rows, representing 1,420 unique patients eligible for prediction and 204 unique events. The cohort flow and landmark risk sets are shown in Fig. 1.

Verification of data extraction excluded 2,368 laboratory records within the time window from other encounters and 536 records that did not meet the existing specimen or test rules. At POD3, corrections to input selection affected 1,744 test-day cells across 202 operations. These counts document the consequences of restricting extraction to index encounters; they are not a controlled estimate of the effect of temporal leakage on predictive performance.

3.2 Primary discrimination and incremental information

Stratified c-statistics and eligibility for inference for the eight main models are presented in Table 3. For XGBoost, extending static features to include the latest laboratory values increased the c-statistic from 0.563 to 0.611. Historical summaries and daily grids did not produce a further stepwise increase. Ridge M2 and M3 met the comparison criteria, whereas M0 and M1 did not meet the project's probability-dispersion threshold at any of the five landmarks. Their descriptive c-statistics were retained, but the corresponding primary paired contrasts were not estimated.

Table 3

Stratified c-statistics from uncalibrated predictions and eligibility for inference in the main analysis

Feature set	Ridge stratified c (95% CI)	Ridge status	XGBoost stratified c (95% CI)	XGBoost status
M0	0.581(0.538, 0.623)	Descriptive only	0.563(0.519, 0.606)	Eligible
M1	0.622(0.585, 0.659)	Descriptive only	0.611(0.572, 0.651)	Eligible
M2	0.615(0.577, 0.654)	Eligible	0.610(0.568, 0.651)	Eligible
M3	0.623(0.580, 0.665)	Eligible	0.597(0.553, 0.641)	Eligible
Note: The stratified c-statistic uses case–noncase pairs within the same landmark. The 95% confidence intervals are conditional on fixed outer-fold predictions. “Descriptive only” indicates that at least one constituent landmark did not meet the project's rule for the standard deviation of predicted logits; it does not indicate a zero estimate or general algorithmic failure.				

All four primary contrasts are presented in Table 4. XGBoost M3 – M0 was 0.034 (95% CI – 0.017 to 0.087), and M3 – M1 was – 0.015 (– 0.061 to 0.033); neither interval excluded zero. The two Ridge contrasts were not estimated under the project rule and should not be interpreted as two findings of “no gain.” The primary intervals allow both positive and negative differences of some magnitude and therefore do not establish equivalence.

Table 4
Four primary information contrasts and within-center temporal-validation contrasts

Analysis	Algorithm	Contrast	Difference in stratified c (95% CI)
Main nested validation	Ridge	M3 – M0	Not estimated under the rule
Main nested validation	Ridge	M3 – M1	Not estimated under the rule
Main nested validation	XGBoost	M3 – M0	0.034(–0.017, 0.087)
Main nested validation	XGBoost	M3 – M1	–0.015(–0.061, 0.033)
2024 temporal test	Ridge	M3 – M1	Not estimated under the rule
2024 temporal test	XGBoost	M3 – M1	0.020(–0.087, 0.113)
2025-onward temporal test	Ridge	M3 – M1	Not estimated under the rule
2025-onward temporal test	XGBoost	M3 – M1	–0.102(–0.208, – 0.004)
Note: The first four rows are the primary contrasts specified in the revised analysis plan; the last four are separate within-center temporal validations. Models were trained and evaluated separately for each analysis population, and their sample sizes and intervals were not pooled. Differences are M3 minus the comparator; positive values favor M3. Ridge contrasts were not estimated because at least one model had a constituent landmark that was ineligible for inference.			

In secondary comparisons between adjacent information levels, XGBoost stratified c-statistic differences were 0.048 (0.007 to 0.091) for M1 – M0, – 0.001 (– 0.039 to 0.037) for M2 – M1, and – 0.013 (– 0.049 to 0.023) for M3 – M2. The positive interval for the latest values versus static features suggests that gains may be concentrated in a relatively simple expansion of information. These comparisons were not adjusted for multiplicity and do not replace the four primary contrasts or constitute independent confirmatory findings. Paired differences for the primary and adjacent information-level contrasts are shown in Fig. 2.

Algorithm comparisons using the same representation also showed no clear advantage. XGBoost-minus-Ridge differences in the stratified c-statistic were – 0.005 (– 0.047 to 0.038) for M2 and – 0.026 (– 0.076 to 0.024) for M3. Cross-family contrasts for M0 and M1 were not estimated because of the Ridge eligibility restriction.

3.3 POD3 performance and calibration

The POD3 risk-set event proportion was 8.09%, providing the reference level for AP at that landmark. Uncalibrated AUROCs for XGBoost M0, M1, M2, and M3 were 0.572, 0.600, 0.618, and 0.611; AP values were 0.093, 0.130, 0.140, and 0.140; and Brier scores were 0.075, 0.075, 0.075, and 0.074. For M3, AUROC was 0.611 (0.553 to 0.670), AP was 0.140 (0.103 to 0.205), and the Brier score was 0.074 (0.062 to 0.086). Paired AUROC differences were 0.039 (– 0.025 to 0.109) for M3 – M0 and 0.011 (– 0.054 to 0.075) for M3 – M1, with uncertainty similar to that of the overall primary contrasts.

For the exploratory supporting metric, XGBoost M3 had a positive AP increment over M0 at POD3: 0.047 (0.010 to 0.104). The M3 – M1 difference was 0.009 (– 0.029 to 0.055). These findings suggest complementary predictive information beyond static features, whereas the AP increment from adding daily histories to the latest-value comparator remained uncertain. POD3 ROC and precision–recall curves are shown in Fig. 3.

At POD3, the jointly estimated calibration slopes for the four uncalibrated XGBoost representations ranged from 0.400 to 0.500. The M3 slope was 0.485 (0.246 to 0.752) before calibration and 0.976 (0.590 to 1.392) after Platt calibration within the training data, bringing the point estimate closer to the ideal value of 1. The Brier score changed from 0.074 to 0.073, and calibration-in-the-large was – 0.137 (– 0.355 to 0.052). These descriptive changes support further evaluation of calibration; no significance claim was made for a paired improvement. Some Ridge combinations remained ineligible after calibration, and their original primary contrasts remained unestimated. Curves before and after calibration within training data are shown in Fig. 4. Detailed POD3 estimates are provided in Additional file 1, Table S4.

3.4 Internal temporal validation

The POD0 risk sets contained 172 patients and 27 events in the 2024 test window and 245 patients and 29 events in the 2025-onward window. XGBoost M1 stratified c-statistics were 0.654 (0.568 to 0.738) and 0.693 (0.602 to 0.773), indicating that the latest-value model retained some discrimination in both later periods. Corresponding M3 estimates were 0.674 (0.580 to 0.757) and 0.590 (0.479 to 0.698). M3 – M1 differences were 0.020 (– 0.087 to 0.113) and – 0.102 (– 0.208 to – 0.004), respectively (Table 4); M1 outperformed M3 in the latter window. Event counts were limited in both windows, and differences should be interpreted alongside their intervals.

Ridge M1 and M3 each had at least one ineligible landmark in both temporal windows. Their stratified c-statistics were therefore descriptive, and paired differences were not estimated. Temporal validation assessed historical training data followed by later cases from the same center and does not replace external validation at an independent institution. Discrimination and paired differences in the two temporal windows are shown in Fig. 5. Additional temporal-validation estimates are reported in Additional file 1, Table S5.

3.5 Missingness representations and other supplementary findings

For XGBoost, adding mask to M3 yielded a stratified c-statistic difference of – 0.009 (– 0.023 to 0.002) relative to the main M3 model. Adding elapsed time to the mask version yielded 0.028 (0.005 to 0.052), indicating an increment for that particular comparison. The mask-plus-elapsed version versus the main M3 model yielded 0.019 (– 0.006 to 0.046). Corresponding Ridge differences were – 0.009 (– 0.036 to 0.018), – 0.005 (– 0.030 to 0.019), and – 0.013 (– 0.052 to 0.024). The contribution of explicit

missingness representations thus depended on the comparator, with no clear overall advantage over the main M3 model. The full sensitivity comparisons are provided in Additional file 1, Table S6.

When era-specific scaling was disabled for the two tests, the unscaled M2 version differed from the main M2 model by -0.001 (-0.002 to 0.001) for Ridge and 0.004 (-0.008 to 0.015) for XGBoost, with no clear difference. The sensitivity analysis restricted to patients still hospitalized at the end of POD3 included 1,303 patients and 104 events. This restriction changed the risk population but did not establish a subgroup with complete follow-up. Inpatient-subset results are reported in Additional file 1, Table S7.

The five outer folds of the specified uncalibrated M3 XGBoost ensemble member generated 10 pairwise attribution comparisons. The median Spearman correlation of feature rankings was 0.391, and the median overlap among the top 20 features was 0.30. Attributions varied across folds and were used primarily to describe the information used by the model. Decision curves were exploratory analyses at specified thresholds and were not used to select a clinical intervention threshold. Repeated-execution and permutation diagnostics and feature-ranking stability are reported in Additional file 1, Tables S8 and S9.

4 Discussion

4.1 Principal findings

This study found incremental predictive signal from postoperative laboratory information beyond static data, with the gain most apparent when the latest laboratory values were added. The XGBoost stratified c-statistic increased from 0.563 for M0 to 0.611 for M1, with an exploratory paired difference of 0.048 (0.007 to 0.091). In the two later temporal windows, the M1 stratified c-statistics were 0.654 and 0.693, respectively. These findings support using the latest laboratory values as a basic reference for dynamic prediction and refine the question for future work: which historical information can provide reproducible gains beyond this reference?

The intervals for both primary contrasts comparing daily history with static data and with the latest values included zero. The study therefore did not establish an additional discrimination advantage for M3, nor did it establish equivalence between the models. At POD3, the AP comparison suggested a positive increment for M3 over static data, but the increment over the latest values remained uncertain; M3 performed less well than M1 in the later temporal window. The two primary Ridge contrasts were not estimated under the project rule, limiting conclusions across algorithms. Overall, the findings support establishing the reference value of laboratory information and using matched comparisons to assess the conditional gains from more detailed history.

4.2 Relationship to previous postoperative prediction studies

The pancreatic surgery model developed by Varghese et al. primarily identified candidates for early discharge, whereas Guillen-Ramirez et al. studied postoperative infections and incorporated laboratory changes and measurement behavior [1, 2]. These studies share the concept of updating risk using postoperative electronic records, but their prediction outcomes, populations at risk, and comparators differ. The specific question in the present study was whether historical representations add information beyond the latest values when predicting severe complications during the remaining observation period. AUROCs across studies depend on the population, outcome, and assessment time and should not be used directly to rank performance.

In their study of complications after colorectal surgery, Ruan et al. found that the gains from temporal architectures varied with the outcome and observation conditions, and that greater model complexity did not necessarily improve performance [4]. The pancreatoduodenectomy study by Ingwersen et al. also supported retaining conventional regression as a reference [5]. The present study specified information representations and algorithm families separately and compared their increments within common risk sets. The M3 representation evaluated here was a daily-value grid, not a neural sequence model such as GRU-D.

Evidence from pancreatic surgery also addresses an outcome similar to ours. In a study of Clavien–Dindo grade $\geq$ IIIa complications after pancreatoduodenectomy, van Dongen et al. compared CRP and white blood cell counts measured before complications and performed validation in an external cohort; changes in these markers had poorer diagnostic performance than their absolute values [17]. This finding provides clinical support for current marker levels as a comparator. The present study extends the evaluation to multiple laboratory tests and information representations within a common validation framework.

Recent research across hospital sites further suggests that the incremental value of history depends on the prediction time. Ruan et al. evaluated GRU-D for prediction of postoperative ileus after colorectal surgery, using logistic regression and random forest models with last-observation-carried-forward handling as references. GRU-D generally performed better at later postoperative times, whereas logistic regression was superior at some early times; transfer across electronic medical record systems also did not consistently improve performance [18]. Thus, the limited increment from the daily grid in the present study should not be generalized to all temporal models.

Conversely, Kato et al. combined serum pancreatic amylase on POD1 with subsequent drain-fluid amylase and CRP for dynamic risk stratification of pancreatic fistula after pancreatoduodenectomy [19], suggesting that specific early markers may provide complementary information for later assessment. Their outcome, information representation, and sample differed from the composite severe outcome and study conditions evaluated here. Future studies could assess targeted historical information based on explicit clinical mechanisms and prediction times, without assuming that retaining all daily values will improve performance.

4.3 Implications for information selection and temporal stability

The latest laboratory results may already summarize a patient's current status, while means, extremes, and slopes compress part of the preceding history. A daily grid retains a more complete temporal distribution but also introduces sparse, correlated features. M3 contained 1,062 numeric features, while stacking landmarks did not increase the number of independent patients or events. The balance between additional information and estimation difficulty may explain why a more elaborate representation of history did not produce a corresponding improvement in performance; this mechanism requires dedicated investigation.

The Clavien–Dindo grade $\geq$ IIIa composite outcome encompasses different clinical processes, and a trajectory useful for one complication may not contribute equally to all severe events. Changes in laboratory platforms, units, and measurement practices over time may also affect temporal performance. M1 retained discrimination signal in both temporal windows, whereas the relative performance of M3 changed, illustrating the practical importance of assessing temporal stability alongside average internal performance. With limited event counts in the available windows, the mechanisms underlying these performance changes remain uncertain.

Stability in the intended application also involves training variability and probability accuracy. Olza et al. assessed random-seed variability, calibration, and performance in subsequent years when predicting unplanned hospitalization [20]. Chen and He recently reported prediction error and calibration in a temporal test cohort for length of stay after total knee arthroplasty [21]. These studies support evaluating representations across several dimensions; the highest point estimate from a single fit does not establish a stable advantage.

For future clinical information system development, these findings provide a clear sequence of comparisons: first quantify the gain from the latest laboratory values over static data, then determine whether specific historical indicators, summaries, or sequences add useful information. The practical value of the study lies in identifying representations and reference comparisons that warrant further validation. Data-processing costs, workload, and patient benefit require separate measurement and cannot be inferred directly from the number of features.

4.4 Missingness information and predicted probabilities

The interval for the increment from adding observation intervals to explicit missingness indicators lay above zero, whereas the increment over the primary M3 remained uncertain. Native missing-value branches in the primary M3 XGBoost model and imputation in Ridge could also carry information about missingness. These analyses therefore assessed the value of additional explicit channels, rather than fully isolating the presence versus absence of missingness information. Whether a test is ordered and when it is repeated may be influenced by both physiological changes and clinical workflows [7, 8]; this study did not separate their causal contributions.

Predicted probabilities and risk ranking should be assessed separately [15]. For XGBoost M3 at POD3, the raw calibration slope was 0.485 and increased to 0.976 after Platt calibration fitted within the training data, suggesting that calibration is an aspect of this prediction workflow worth further refinement. Given the wide intervals, probability accuracy still requires independent validation. The exploratory decision curves provide a starting point for further evaluation [16]. Actual intervention thresholds should be determined in relation to specific actions, the consequences of false positives and false negatives, and clinical validation.

4.5 Limitations

First, this was a retrospective study, with outcomes based on clinical records and manual adjudication by the investigator. Outcome capture remained dependent on the coverage of the source records, particularly for events after discharge or at other institutions. Event dates were anchored using the existing clinical rules and handled at calendar-day resolution. The study therefore evaluated risk updates at the end of the day and does not establish the ability to provide early warning at an hourly resolution.

Second, the single-center retrospective cohort spanned a long period and included a limited number of independent events. Cohort selection and date recovery were documented in two stages, but the established data-availability conditions may still have affected population representativeness. Laboratory extraction did not yet include complete test-specific unit harmonization or retention of modifiers for values outside assay limits. Period-specific scaling cannot replace verification of laboratory units. These factors may affect estimates and their applicability across periods and centers.

Third, the research question and some analytical details were revised after earlier results were available, so all findings represent exploratory evidence. Patient-grouped nested validation controlled information sharing between the current training and test processes but did not remove the influence of research-question selection. Resampling fixed held-out predictions did not encompass the full uncertainty from repeated hyperparameter selection and model fitting. Multiple comparisons were not adjusted for, and positive secondary findings require independent validation.

Fourth, a standard deviation of 0.30 for predicted logits was an operational project rule, not a universal threshold for the calculability of discrimination metrics or clinical validity. A narrow distribution of predicted probabilities may still retain ranking information. Under this rule, the two primary Ridge contrasts were not estimated, limiting the extent to which conclusions about incremental information apply across algorithms. Descriptive values were retained, and this eligibility restriction was not interpreted as failure of the regression algorithm.

Finally, the temporal test data came from the same center; data from independent institutions are needed to assess generalizability. Feature attributions varied across folds and should not be interpreted as stable etiological mechanisms. The study has not yet undergone prospective workflow evaluation, and performance metrics alone do not establish improvements in clinical management or patient outcomes.

5 Conclusions

In this exploratory study of dynamic prediction of severe complications after pancreatic resection, the latest laboratory values provided incremental discrimination signal over static data and retained some predictive ability in two internal temporal windows. The primary comparisons did not show a consistent additional benefit from daily history, which performed less well than the latest-value model in the later temporal window. This study provides empirical evidence to guide selection of laboratory information: more complete historical representations should demonstrate their incremental value against an explicit latest-value reference, with subsequent application assessed through calibration, temporal stability, and independent validation.

Abbreviations

AP, average precision; AUROC, area under the receiver operating characteristic curve; CD, Clavien-Dindo classification; CI, confidence interval; DP, distal pancreatectomy without splenectomy; DPS, distal pancreatectomy with splenectomy; EHR, electronic health record; ICU, intensive care unit; PD, pancreatoduodenectomy; POD, postoperative day; PPPD, pylorus-preserving pancreatoduodenectomy; Ridge, L2-regularized logistic regression; SHAP, Shapley additive explanations; XGBoost, extreme gradient boosting.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. The study and the application for waiver of informed consent were approved by the Medical Ethics Committee of Sun Yat-sen Memorial Hospital, Sun Yat-sen University (approval number SYSKY-2026-589-01; 13 July 2026).

Consent for publication

Not applicable. This manuscript reports aggregate findings and contains no identifiable individual participant information.

Availability of data and materials

Individual-level clinical data are not publicly available because of patient confidentiality and institutional data-governance restrictions. Supplementary methods and aggregate results are provided in Additional file 1. Requests for analysis code or further research materials should be directed to the corresponding authors and are subject to institutional ethics and data-governance requirements.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (grant no. 82273476); the Guangdong Basic and Applied Basic Research Foundation (grant no. 2025A1515220082); the Sun Yat-sen Pilot Scientific Research Fund (grant no. SYSQH-II-2024-03); and the Beijing Xisike Clinical Oncology Research Foundation (grant no. Y-QLQN2025-0297).

Authors' contributions

Jingyun Ning led the study design, developed the methodology, conducted the research and data analysis, and drafted the manuscript. Ruiqi Li contributed to verification and interpretation of the results and manuscript revision. Huihuang Yu and Jianfeng Huang contributed to data curation and manuscript review. Haoming Lin and Rui Zhang contributed to study conceptualization and design, supervised the research, and critically revised the manuscript. Jingyun Ning and Ruiqi Li contributed equally to this work.

Declaration of generative AI use

OpenAI Chatgpt assisted with manuscript drafting, translation, language revision, consistency checks. The authors take responsibility for the accuracy and integrity of the final manuscript.

References

1. Varghese C, Choudhary A, Starlinger P, et al. Dynamic prediction of early discharge after major pancreatic surgery: Derivation and cross-validation of an automatable electronic medical record-based model. *Surgery*. 2026;195:109976. <https://doi.org/10.1016/j.surg.2025.109976>.
2. Guillen-Ramirez H, Sanchez-Taltavull D, Perrodin S, et al. Prediction of postoperative infections by strategic data imputation and explainable machine learning. *J Am Med Inf Assoc*. 2025;32(11):1706–17. <https://doi.org/10.1093/jamia/ocaf145>.
3. Sheikhalishahi S, Goss S, Seidlmayer LK, et al. Predicting blood transfusion demand in intensive care patients after surgery by comparative analysis of temporally extended data selection. *BMC Med Inf Decis Mak*. 2024;24:397. <https://doi.org/10.1186/s12911-024-02800-z>.
4. Ruan X, Fu S, Storlie CB, Mathis KL, Larson DW, Liu H. Real-time risk prediction of colorectal surgery-related post-surgical complications using GRU-D model. *J Biomed Inf*. 2022;135:104202. <https://doi.org/10.1016/j.jbi.2022.104202>.

5. Ingwersen EW, Stam WT, Meijs BJV, Dutch Pancreatic Cancer Group, et al. Machine learning versus logistic regression for the prediction of complications after pancreatoduodenectomy. *Surgery*. 2023;174(3):435–40. <https://doi.org/10.1016/j.surg.2023.03.012>.
6. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci Rep*. 2018;8:6085. <https://doi.org/10.1038/s41598-018-24271-9>.
7. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol*. 2020;125:183–7. <https://doi.org/10.1016/j.jclinepi.2020.03.028>.
8. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res*. 2020;4:8. <https://doi.org/10.1186/s41512-020-00077-0>.
9. Sun M, Engelhard MM, Bedoya AD, Goldstein BA. Incorporating informatively collected laboratory data from EHR in clinical prediction models. *BMC Med Inf Decis Mak*. 2024;24:206. <https://doi.org/10.1186/s12911-024-02612-1>.
10. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>.
11. Moons KGM, Damen JAA, Kaul T, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. <https://doi.org/10.1136/bmj-2024-082505>.
12. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg*. 2004;240(2):205–13. <https://doi.org/10.1097/01.sla.0000133083.54934.ae>.
13. van Houwelingen HC. Dynamic Prediction by Landmarking in Event History Analysis. *Scand J Stat*. 2007;34(1):70–85. <https://doi.org/10.1111/j.1467-9469.2006.00529.x>.
14. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 2015;10(3):e0118432. <https://doi.org/10.1371/journal.pone.0118432>.
15. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. <https://doi.org/10.1186/s12916-019-1466-7>.
16. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565–74. <https://doi.org/10.1177/0272989X06295361>.
17. van Dongen JC, Smits FJ, van Santvoort HC, et al. C-reactive protein is superior to white blood cell count for early detection of complications after pancreatoduodenectomy: a retrospective multicenter cohort study. *HPB (Oxford)*. 2020;22(10):1504–12. <https://doi.org/10.1016/j.hpb.2020.02.005>.
18. Ruan X, Fu S, Jia H, et al. Gated recurrent unit with decay has real-time capability for postoperative ileus surveillance and offers cross-hospital transferability. *Commun Med*. 2025;5:331.

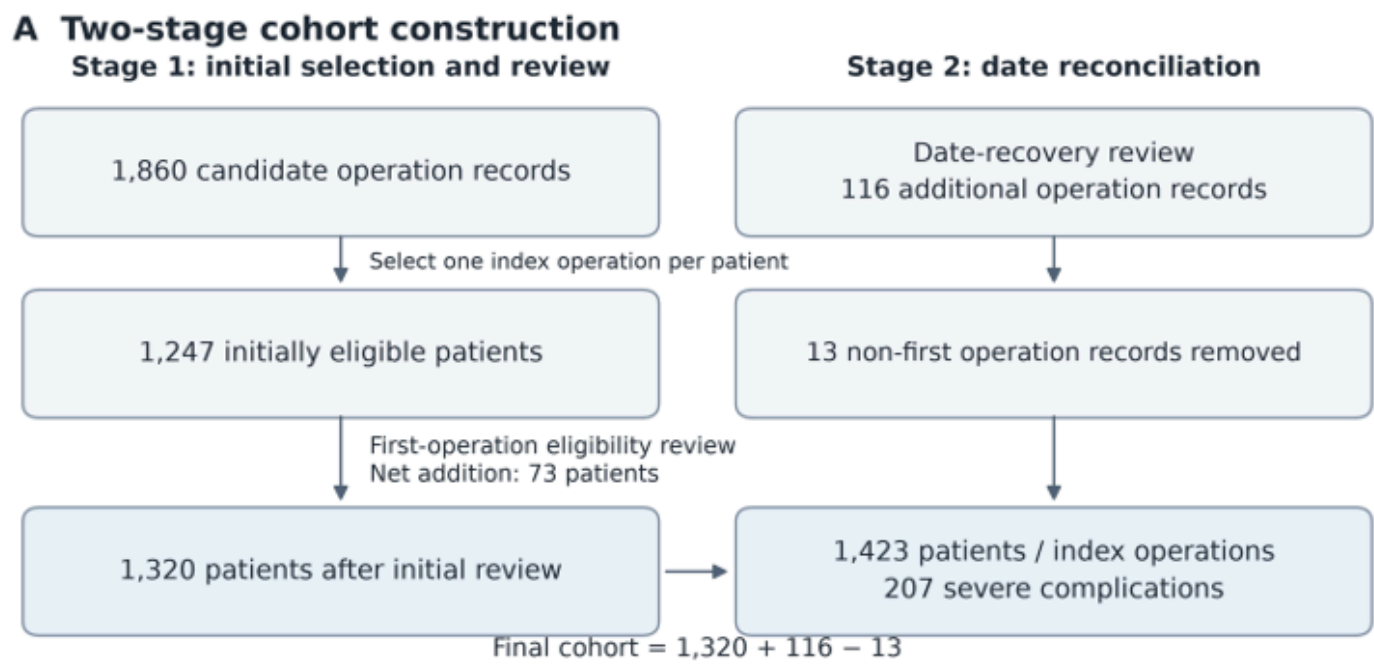
<https://doi.org/10.1038/s43856-025-01053-9>.

19. Kato T, Murase Y, Baba Y, et al. Impact of dynamic risk stratification for postoperative pancreatic fistula using serum pancreatic amylase on the day after pancreatoduodenectomy. *Surgery*. 2025;184:109454. <https://doi.org/10.1016/j.surg.2025.109454>.

20. Olza A, Millán E, Rodríguez-Álvarez MX. Development and validation of predictive models for unplanned hospitalization in the Basque Country: analyzing the variability of non-deterministic algorithms. *BMC Med Inf Decis Mak*. 2023;23:152. <https://doi.org/10.1186/s12911-023-02226-z>.

21. Chen S, He W. Interpretable ensemble learning for hospital prediction duration following total knee arthroplasty: a retrospective temporal validation study. *BMC Med Inf Decis Mak*. 2026. <https://doi.org/10.1186/s12911-026-03825-2>. Advance online publication, 9 September 2026.

Figures



B Event-free landmark risk sets

POD0: exclude 3 events on or before the cutoff → 1,420 patients

	POD0	POD1	POD3	POD5	POD7
At risk	1,420	1,385	1,323	1,317	1,302
Events*	204	169	107	101	86
Event rate	14.37%	12.20%	8.09%	7.67%	6.61%

* First severe complication after the landmark cutoff and through POD30.
Risk sets overlap; totals across landmarks are not independent patients.

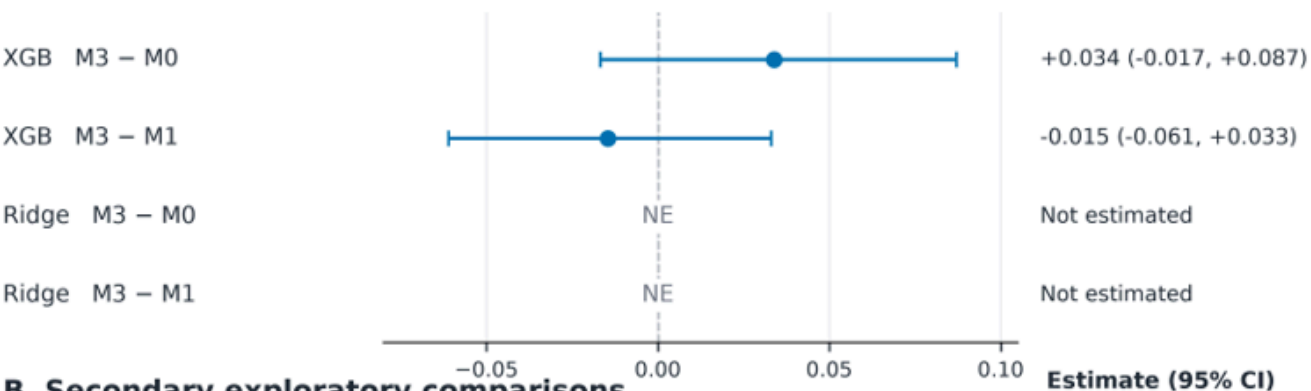
Figure 1

Cohort construction and landmark risk sets

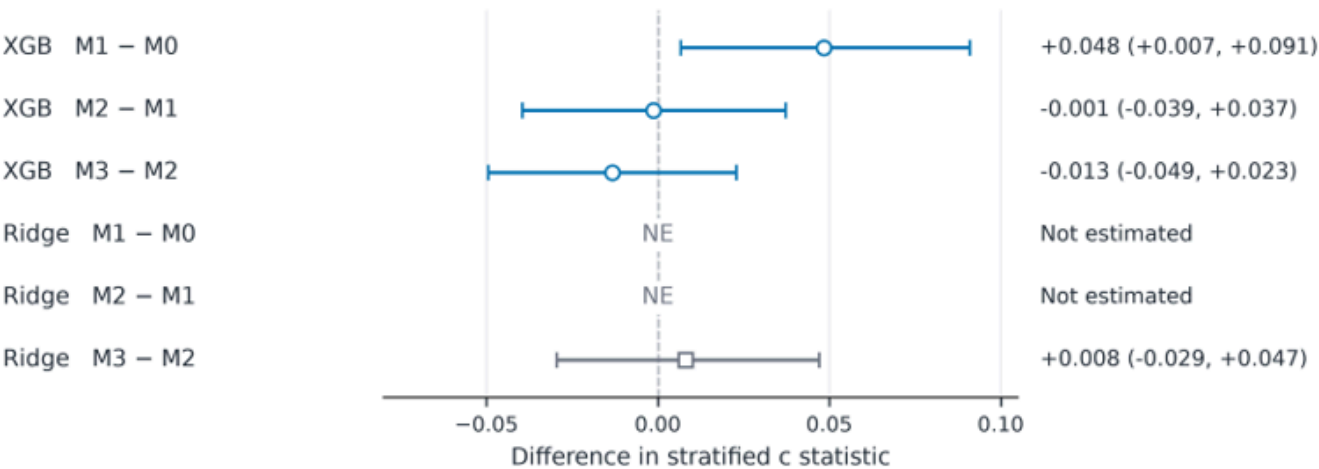
(A) Initial selection and review of first-operation eligibility yielded 1,320 patients. Date reconciliation recovered 116 operation records, and 13 non-first operation records were removed, resulting in 1,423 patients with their index operations. Candidate-stage counts refer to operation records; counts after index selection refer to patients. (B) Each landmark includes patients free of a first severe complication at the end of that postoperative day. Events are first Clavien–Dindo grade $\geq$ IIIa complications after the landmark and through postoperative day 30. Landmark risk sets overlap and must not be summed as independent patients. Figure prepared from the study results using plotting code assisted by OpenAI Codex.

Incremental discrimination across all five landmarks

A Primary comparisons



B Secondary exploratory comparisons



M0: static; M1: + latest values; M2: + summaries; M3: + daily time grid.
NE: not estimated under the study criterion because a component model had $SD(\text{logit } p) < 0.30$.

Figure 2

Incremental discrimination by information representation

Points show differences in the stratified c statistic across five landmarks; horizontal lines show 95% confidence intervals from 2,000 paired patient-cluster bootstrap resamples of existing out-of-fold predictions. (A) Four primary comparisons specified in the revised analysis plan. (B) Secondary exploratory comparisons between adjacent representations. No comparisons were adjusted for multiplicity. NE indicates that a contrast was not estimated under the study $SD(\logit p) < 0.30$ criterion; it does not imply mathematical non-estimability or an absence of discrimination. M0: static information; M1: additional latest laboratory values; M2: additional summary statistics; M3: additional daily time-grid information. Figure prepared from the study results using plotting code assisted by OpenAI Codex.

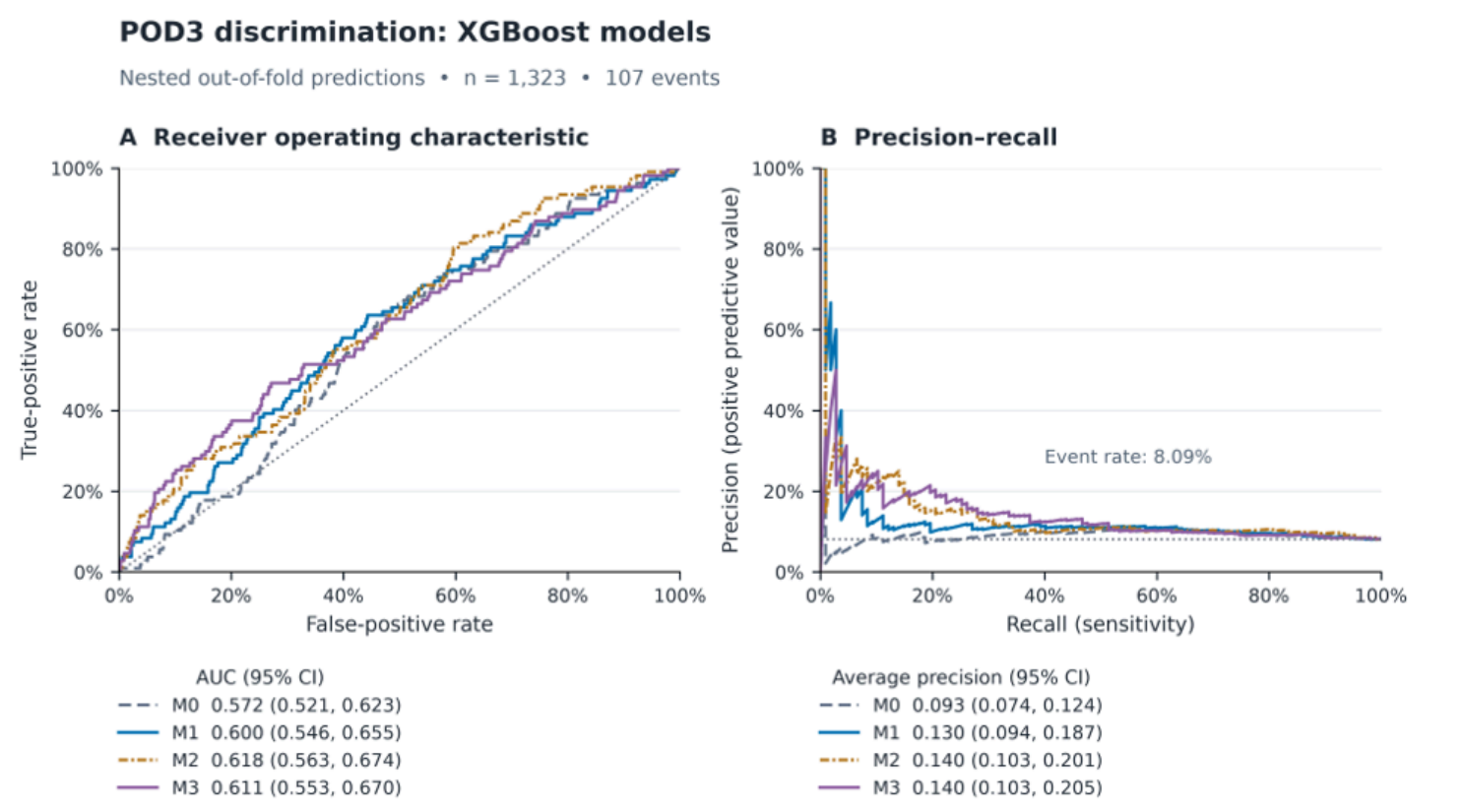


Figure 3

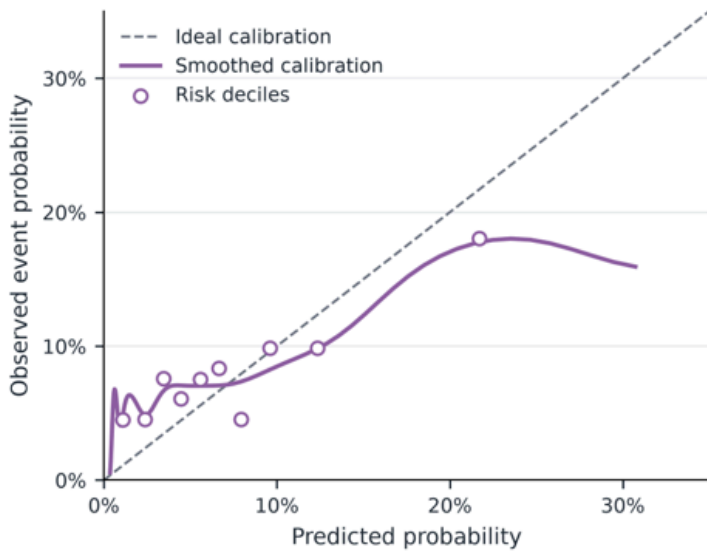
Discrimination at postoperative day 3

The risk set contains 1,323 patients and 107 subsequent severe complications (event rate, 8.09%). Curves use out-of-fold predictions from patient-grouped nested cross-validation. Legends report AUC or non-interpolated average precision with existing patient-cluster bootstrap 95% confidence intervals. The ROC reference line represents chance discrimination; the horizontal precision–recall reference line represents the risk-set event rate. Intervals refer to the summary metrics in the legends; curves are shown without confidence bands. Figure prepared from the study results using plotting code assisted by OpenAI Codex.

POD3 calibration: M3 XGBoost

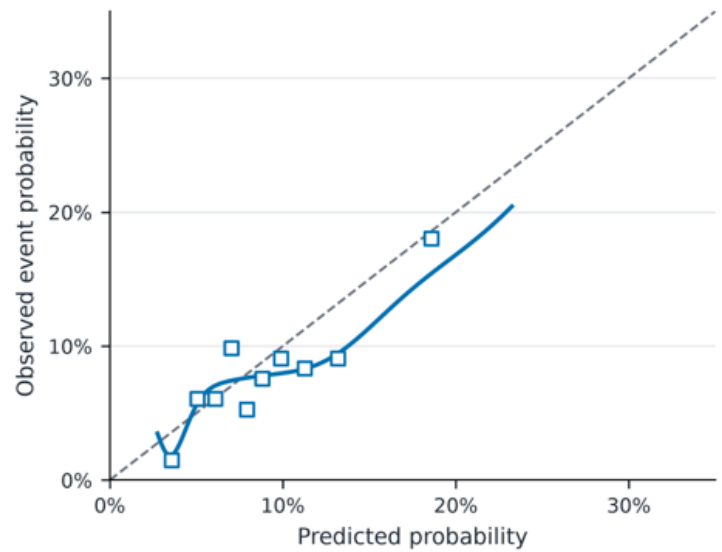
Nested out-of-fold predictions • n = 1,323 • 107 events

A Original probabilities



Calibration slope: 0.485 (0.246, 0.752)
Brier score: 0.074 (0.062, 0.086)

B Training-only Platt recalibration



Calibration slope: 0.976 (0.590, 1.392)
Brier score: 0.073 (0.061, 0.085)

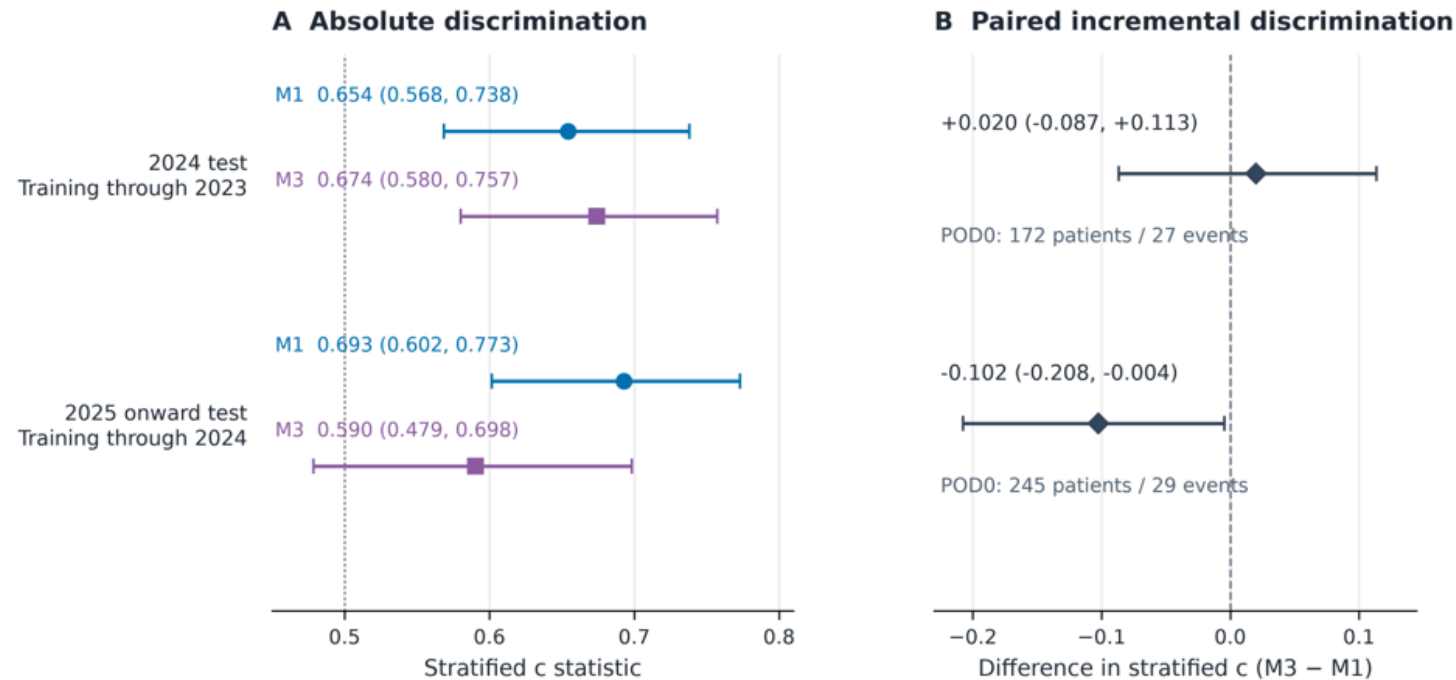
Figure 4

Calibration of M3 XGBoost at postoperative day 3

(A) Original predicted probabilities. (B) Out-of-fold probabilities after Platt recalibration fitted using training data only. Points represent mean predicted probabilities and observed event proportions within the existing risk deciles; solid lines are the saved smooth calibration curves, and dashed lines indicate ideal calibration. Calibration slopes and Brier scores are shown with existing 95% confidence intervals. Both panels use identical axis ranges, and intervals refer to summary metrics. Calibration changes are descriptive. Figure prepared from the study results using plotting code assisted by OpenAI Codex.

Temporal validation: latest-value and time-grid XGBoost models

Five landmarks per window; raw held-out probabilities



Points and horizontal lines show estimates and patient-cluster bootstrap 95% confidence intervals. POD0 counts describe the initial risk set; later landmarks have smaller overlapping risk sets.

Figure 5

Temporal validation of XGBoost M1 and M3

(A) Stratified c statistics across five landmarks. (B) Paired M3–M1 differences within each window. Horizontal lines show existing 95% confidence intervals from 2,000 patient-cluster bootstrap resamples; intervals for differences are based on paired comparisons. At POD0, the 2024 test window contains 172 patients and 27 subsequent events; the 2025-onward window contains 245 patients and 29 events. Later landmark risk sets are smaller and overlap. The two temporal windows are reported separately. Figure prepared from the study results using plotting code assisted by OpenAI Codex.

Supplementary Files

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

- [Additionalfile1Supplementarymaterial.docx](#)