

Development and internal validation of a machine-learning-informed nomogram for predicting severe acute pancreatitis in the intensive care unit: a retrospective cohort study using the MIMIC-IV database

Gaozhi Zhang

¹Nanjing Medical University

Xufeng Chen

cx-fyx@163.com

²Jiangsu Provincial People's Hospital

Research Article

Keywords: acute pancreatitis, severe acute pancreatitis, prediction model, nomogram, machine learning, MIMIC-IV, organ failure, intensive care

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11061587/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: Most deaths from acute pancreatitis (AP) occur in the one in five patients who develop persistent organ failure, that is severe acute pancreatitis (SAP), and existing severity scores discriminate modestly. We developed and internally validated a model for SAP from routinely available first-24-hour data and benchmarked it against machine learning.

Methods: Retrospective cohort study using MIMIC-IV version 2.2. Adults admitted to an intensive care unit (ICU) with a first hospitalisation for AP (ICD-9 577.0 or ICD-10 K85.x) and an ICU stay of at least 24 hours were included. SAP was defined per the revised Atlanta classification as persistent organ failure beyond 48 hours, operationalised with a daily modified Marshall score. 62 first-24-hour candidate predictors were considered. Missing data were handled by multiple imputation by chained equations, excluding the outcome. Predictors were selected by least absolute shrinkage and selection operator (LASSO) within each imputation, retaining variables selected in at least three of five; a multivariable logistic model was pooled by Rubin's rules and presented as a nomogram. Performance was assessed by 10-fold cross-validation and bootstrap, with calibration and decision curve analysis, and benchmarked against eight machine learning algorithms. Reporting followed TRIPOD+AI.

Results: Of 5,881 AP hospitalisations, 734 patients formed the analytic cohort and 290 (39.5%) developed SAP; hospital mortality was 23.4% versus 6.1%. LASSO selected 21 predictors. The model achieved a cross-validated area under the receiver operating characteristic curve of 0.877 (95% CI 0.852-0.903), a calibration slope of 0.884 (95% CI 0.755-1.013) and a Brier score of 0.136; bootstrap-corrected AUC was 0.889. Discrimination exceeded the modified BISAP score (AUC 0.693; $p < 0.001$), with a net reclassification improvement of 0.600. Across eight machine learning models (test AUC 0.818-0.850), none of which significantly outperformed logistic regression ($p = 0.494$). SHAP ranked minimum Glasgow Coma Scale, creatinine, mechanical ventilation and blood urea nitrogen highest. Removing organ-support variables reduced AUC to 0.816 ($p = 0.016$); adding five sparse biomarkers did not improve it ($p = 0.143$).

Conclusions: A first-24-hour nomogram predicted severe acute pancreatitis with good discrimination and calibration, matching eight machine learning models while remaining transparent and usable at the bedside. External validation is required before clinical adoption.

2. Background

Acute pancreatitis (AP) is among the most common gastrointestinal causes of acute hospitalisation, with an incidence that has risen steadily over recent decades [1, 2]. Most patients have a mild, self-limiting course, but approximately one in five develops severe acute pancreatitis (SAP), a syndrome characterised by persistent organ failure and associated with mortality of 20% to 30% [3, 4]. The clinical trajectory of SAP is established early: organ failure that persists beyond 48 hours, rather than organ failure at any single time point, is the principal determinant of outcome [5]. This makes the first two days of admission a decisive window, because interventions that alter outcome, including goal-directed fluid

therapy, early enteral nutrition and admission to a higher level of care, are most effective when applied before organ failure becomes established [6, 7].

Several severity scores are in routine use, including the Acute Physiology and Chronic Health Evaluation (APACHE) II score, the Bedside Index for Severity in Acute Pancreatitis (BISAP), together with the organ-failure scores used to define severity, namely the modified Marshall score and the Sequential Organ Failure Assessment (SOFA) score [3, 8–11]. Their performance is nonetheless modest. BISAP, which was designed for early bedside use, typically achieves an area under the receiver operating characteristic curve (AUC) of approximately 0.70 to 0.80 for severe disease, and APACHE II requires 24 hours of data collection and a large number of variables [8, 9]. More importantly, these scores were derived in mixed or historical populations and were not developed with the statistical machinery now expected of a clinical prediction model, so their calibration in contemporary intensive care populations is uncertain.

Machine learning has been proposed as a route to better risk stratification in acute pancreatitis, and a substantial number of models have been published from critical care databases [12, 13]. The evidence that these methods deliver materially better discrimination than a carefully built logistic model is, however, weak. Many reports compare machine learning against univariable or unpenalised regression and do not report calibration, and some achieve implausibly high AUCs that suggest information leakage between predictors and outcome [14]. There is a corresponding risk that complexity is adopted without benefit to the clinician, who must weigh the loss of interpretability against a gain that may not exist.

We therefore developed and internally validated a prediction model for SAP among patients with AP admitted to an intensive care unit (ICU), using only data available within the first 24 hours, and set it against eight competing machine learning algorithms. Our specific objectives were, first, to derive and internally validate a parsimonious and interpretable model presented as a nomogram; second, to determine whether any machine learning algorithm materially outperforms it; and third, to test two threats to validity directly, namely whether the model is merely re-encoding organ failure through organ-support variables, and whether sparsely measured biomarkers add incremental value. Reporting follows the TRIPOD + AI statement [15, 16].

3. Methods

3.1. Study design and data source

This was a retrospective cohort study using MIMIC-IV version 2.2, a publicly available de-identified electronic health record dataset from a single tertiary academic medical centre in the United States, covering the period 2008 to 2019 [17]. MIMIC-IV is distributed through PhysioNet under a credentialed data use agreement; the institutional review board of the Beth Israel Deaconess Medical Center approved its collection and waived the requirement for individual informed consent. All analyses used only de-identified data. We followed the TRIPOD+AI checklist for prediction model studies [15]; the completed checklist is provided as Additional file 1.

3.2. Cohort definition

We identified all hospitalisations with a diagnosis of AP using International Classification of Diseases, Ninth and Tenth Revision codes (ICD-9 577.0; ICD-10 K85.x and its subcodes). We excluded hospitalisations with a diagnosis of chronic pancreatitis (ICD-9 577.1; ICD-10 K86.0-K86.1) or pancreatic malignancy (ICD-9 157.x; ICD-10 C25.x), because these conditions have a different natural history and are not the population to which the revised Atlanta classification applies. We further restricted the cohort to patients aged 18 years or older who were admitted to an ICU and whose ICU stay lasted at least 24 hours after ICU admission; the latter requirement ensured that the 24-hour predictor window was complete. Where a patient had more than one eligible hospitalisation, only the first was retained.

The requirement of an ICU stay warrants explanation, because it determines both the size and the severity composition of the cohort. Assessment of persistent organ failure requires organ-function variables, including arterial blood gases and hourly haemodynamics, that are recorded only in the ICU in this dataset. Patients with AP who are never admitted to an ICU are therefore not eligible for the outcome definition applied here, and the resulting cohort is enriched for severe disease relative to a general AP population.

3.3. Outcome definition

The outcome was SAP, defined according to the revised Atlanta classification as persistent organ failure lasting more than 48 hours [3]. Organ failure was operationalised with the modified Marshall score, computed for each of the first seven days after ICU admission, using the respiratory (PaO₂/FiO₂ ratio), renal (serum creatinine) and cardiovascular (systolic blood pressure and arterial pH) domains. A score of 2 or more in any single organ system constituted organ failure. Because the classification cannot be finalised before 48 hours have elapsed, and to prevent overlap with the 24-hour predictor window, persistence was assessed from day 2 to day 7 and required a score of 2 or more in the same organ system on two consecutive days.

The cardiovascular domain deserves particular comment. The modified Marshall definition distinguishes systolic blood pressure below 90 mmHg that responds to fluid resuscitation (grade 1, not organ failure) from pressure that does not respond (grade 2, organ failure), and specifies that patients receiving inotropic support are assigned grade 2 or higher irrespective of the blood pressure achieved. Fluid responsiveness is not recorded in a form that permits this distinction retrospectively, so we assigned grade 1 to a systolic pressure below 90 mmHg and grade 2 or above to any patient receiving vasopressors, with grades 3 and 4 further assigned for arterial pH below 7.3 and below 7.2 respectively in the presence of vasopressor support. Because a single dichotomous outcome can be derived in more than one defensible way, and the choice materially affects the apparent prevalence, we prespecified five candidate operationalisations and compared them before adopting one; the comparison is reported in the Results and in Table 4.

Patients whose ICU stay was shorter than 48 hours could not be assessed for persistence and were classified as not having SAP by construction. The consequences of this are examined in a sensitivity analysis and in the Discussion.

3.4. Candidate predictors

62 candidate predictors were extracted for the first 24 hours after ICU admission. These comprised demographic characteristics (age, sex), aetiology indicators (biliary, alcohol-related), eleven comorbid conditions, routine vital signs (heart and respiratory rate, minimum systolic, diastolic and mean arterial pressure, minimum oxygen saturation, maximum temperature, minimum Glasgow Coma Scale), laboratory measurements (white cell count, platelet count, haematocrit, red cell distribution width, glucose, blood urea nitrogen, creatinine, calcium, sodium, potassium, chloride, aspartate and alanine aminotransferase, alkaline phosphatase, bilirubin, lactate, arterial pH, PaO₂, base excess, international normalised ratio, partial thromboplastin time), derived ratios (blood urea nitrogen to creatinine, aspartate to alanine aminotransferase), organ-support and output variables (mechanical ventilation, vasopressor use, renal replacement therapy, 24-hour urine output), and several sparsely measured biomarkers retained for the purposes of the sensitivity analysis described below. Variables were extracted at their worst value within the window, reflecting the convention used by severity scores. Race and admission type were recorded for descriptive purposes but were not offered as predictors, because race is not a biological risk factor and admission type is a care-process variable rather than a clinical variable.

3.5. Missing data

Variables missing in more than 35% of patients were excluded from the candidate pool before imputation, because imputation of a variable absent in a third or more of patients is unlikely to recover the underlying distribution reliably. 49 of the 62 candidates were retained on this rule. The remainder of the missing data was handled by multiple imputation by chained equations using the mice package, with five imputed datasets and ten iterations [18]. The outcome was deliberately excluded from the imputation model so that outcome information could not migrate from the test set into the training data. Imputation was performed on the whole cohort before any splitting, so that training and test sets share a single imputation model, and analyses were pooled across the five imputations according to Rubin's rules [19].

3.6. Sample size

The minimum sample size was calculated with the pmsampsize approach of Riley and colleagues, assuming a conservative Cox-Snell R-squared of 0.35, a shrinkage factor of 0.9 and the observed outcome prevalence [20, 21]. For a model with 21 parameters this required 450 patients and 178 events,

corresponding to 8.5 events per parameter. The analytic cohort of 734 patients with 290 events exceeded this requirement.

3.7. Model development

The cohort was divided into a training set (70%) and a held-out test set (30%), stratified by outcome. Predictor selection used the least absolute shrinkage and selection operator (LASSO) with a logistic loss function, implemented with 10-fold cross-validation and the lambda minimising deviance [22, 23]. Selection was performed independently within each of the five completed datasets, and a variable was retained if it was selected in at least three of the five. This majority rule is more stable than relying on a single imputation and avoids the anticonservative behaviour of selecting on pooled statistics.

The selected variables were entered into a multivariable logistic regression model, fitted in each completed dataset and pooled by Rubin's rules to give final odds ratios and confidence intervals. Predictors were retained on the clinical scale, without standardisation, so that the coefficients correspond to the units in which clinicians measure them and the resulting nomogram is directly usable. The nomogram is drawn from the pooled coefficients themselves, so that the figure, the odds ratios in Table 2 and the reported performance all describe one model. A nomogram can represent a pooled model exactly because it adds up a linear predictor and applies a single logistic transform; a patient's predicted risk is therefore the mean linear predictor across the five completed datasets, which is the quantity the nomogram's probability scale is drawn to match. Averaging the five predicted probabilities instead would be equally reasonable as a prediction but would correspond to no set of coefficients, and the probability read off a printed nomogram would not then be the probability the model reports.

3.8. Model evaluation

Discrimination was quantified by the AUC with DeLong confidence intervals and compared between models with the DeLong test for correlated curves [24]. Calibration was assessed by the calibration slope and intercept, the Brier score, and the Hosmer-Lemeshow test, and displayed graphically [25]. Because the Hosmer-Lemeshow test is sensitive to grouping and sample size and has been criticised as a summary of calibration, we report it alongside the calibration slope and intercept and interpret the three together [26]. Clinical utility was assessed by decision curve analysis [27]. The incremental value of the model over the modified BISAP score was quantified by the net reclassification improvement and the integrated discrimination improvement [28]; BISAP was placed on the probability scale by fitting it as a single-predictor logistic model in the training set, so that the comparison was not flattered by using an uncalibrated count.

Internal validation used two complementary approaches. First, a 10-fold cross-validation was performed across the whole cohort, refitting the entire modelling procedure within each fold and pooling the out-of-fold predictions; this estimates the performance of the modelling strategy rather than of one particular

split and is the estimate we regard as primary. Second, a 1000-replicate bootstrap was used to correct the apparent AUC for optimism [29]. A single 70/30 split is reported for transparency and for comparison with the machine learning models, which were evaluated on the same split, but it rests on only 221 patients and 87 events and is correspondingly imprecise.

3.9. Machine learning comparison

Eight algorithms were fitted to the same LASSO-selected predictors, the same imputed datasets and the same training and test split: logistic regression, random forest, support vector machine with a radial kernel, gradient boosting, extreme gradient boosting, naive Bayes, a multilayer perceptron and k-nearest neighbours. Hyperparameters were tuned by repeated 5-fold cross-validation with three repeats, using the caret framework [30, 31]. Continuous predictors were standardised on training-set moments for the scale-sensitive algorithms. Predictions were averaged across the five completed datasets, and classification thresholds were derived from the training set by the Youden index so that the test set was never used to choose an operating point.

No class-imbalance correction was applied. The outcome was present in 39.5% of the cohort, which is close to balanced, and we preferred to leave the predicted probabilities unadjusted so that they remain interpretable as risks; resampling or weighting would have distorted calibration without improving discrimination. Sensitivity and specificity are reported at a threshold chosen by the Youden index rather than at the default of 0.5, and the decision curve analysis in Fig. 4 allows the reader to examine any threshold of interest.

3.10. Model interpretation

SHAP values, which attribute each prediction to individual predictors on an additive scale, were computed for the extreme gradient boosting model using the TreeSHAP algorithm [32, 33]. This provides a variable-importance ranking that does not depend on the scale of the predictors and permits comparison with the direction of effect in the logistic model.

3.11. Sensitivity analyses

Three prespecified sensitivity analyses were performed. First, the sparsely measured biomarkers excluded by the missingness filter (albumin, lactate dehydrogenase, lipase, and the neutrophil-to-lymphocyte and platelet-to-lymphocyte ratios) were re-offered to the selection procedure with the missingness threshold relaxed to 50%, to test whether they carry incremental information that the filter discarded. Second, because SAP is by definition a state of organ failure, first-24-hour organ-support variables (mechanical ventilation, vasopressor use and renal replacement therapy) may partly encode the outcome rather than predict it; the model was therefore refitted with these three variables removed

from the candidate pool. Third, the model was refitted in the subset of patients whose ICU stay was at least 48 hours, in whom the persistence criterion could be applied to every patient rather than being assumed absent.

3.12. Software

All analyses were performed in R version 4.5.3 (R Foundation for Statistical Computing, Vienna, Austria) [34], using the packages glmnet, mice, pROC, rms, dcurves, caret, ranger, xgboost and tableone. Two-sided p values below 0.05 were considered statistically significant, and no adjustment was made for multiplicity in the sensitivity analyses, which are interpreted as exploratory.

4. Results

4.1. Cohort

Of 5,881 hospitalisations coded for acute pancreatitis, exclusion of chronic pancreatitis (1,621) and pancreatic malignancy (112) left 4,148. Of these, 928 involved an ICU admission, 780 had an ICU stay of at least 24 hours, and 734 remained after retaining only the first eligible admission per patient. These 734 patients formed the analytic cohort (Fig. 1). The median age was 59 years (interquartile range 46 to 72) and 422 patients (57.5%) were men. The most common first ICU was a mixed medical and surgical unit (321 patients, 43.7%).

4.2. Outcome and its derivation

The five prespecified operationalisations of SAP produced prevalences ranging from 27.4% to 48.1% (Table 4). The definition originally considered, which counted any organ failure across days 1 to 7 and treated a systolic pressure below 90 mmHg as cardiovascular grade 2, gave a prevalence of 48.1%. We judged this implausibly high and, on review, identified three problems: it ignored fluid responsiveness in the cardiovascular domain, it allowed persistence to be satisfied by different organ systems on consecutive days, and its day-1 inclusion overlapped the predictor window, creating a degree of circularity.

The adopted definition, which assesses persistence from day 2 to day 7 within the same organ system and assigns cardiovascular grade 1 to hypotension without vasopressor support, identified 290 patients with SAP, a prevalence of 39.5%. Supporting this choice, the adopted definition produced the largest separation in hospital mortality between groups among the definitions tested and an odds ratio for death of 4.73. Hospital mortality was 23.4% (68 of 290) in patients with SAP compared with 6.1% (27 of 444) in those without, a gradient consistent with the published literature and considerably steeper than the 21.2% versus 5.2% yielded by the original definition. A further 186 patients had transient organ failure that did not persist, and 258 had no organ failure. Among the three organ systems, respiratory failure

was the most frequent driver of persistent organ failure (227 patients), followed by renal (148) and cardiovascular (134).

Of the 220 patients whose ICU stay was shorter than 48 hours and who were therefore classified as not having SAP by construction, 22 (10.0%) died in hospital.

4.3. Baseline characteristics

Patients who developed SAP were similar in age and sex to those who did not, but differed markedly across physiological and laboratory variables (Table 1). The largest standardised differences were for mechanical ventilation within 24 hours (75.5% versus 26.6%, standardised mean difference 1.12), minimum Glasgow Coma Scale (7.4 versus 12.3, SMD 1.18), vasopressor use (53.4% versus 14.9%, SMD 0.89), minimum systolic blood pressure (87.4 versus 100.1 mmHg, SMD 0.67) and blood urea nitrogen (32.5 versus 17.0 mg/dL, SMD 0.62). Serum lipase and amylase did not differ between groups (SMD 0.02 and 0.06 respectively). Patients with SAP had substantially longer ICU stays (median 244.8 versus 49.5 hours) and hospital stays (19.9 versus 8.2 days).

Several variables were available in only part of the cohort, including PaO₂ (513 of 734, 69.9%), lactate (551, 75.1%), lipase (423, 57.6%), albumin (417, 56.8%) and the neutrophil-to-lymphocyte ratio (385, 52.5%). The number of patients with each variable recorded is given in Table 1 so that the precision of each comparison can be judged.

4.4. Predictor selection and the final model

Of the 49 candidates that passed the missingness filter, LASSO selected 21 variables in at least three of the five imputations (Table 2). Fourteen were selected in all five: liver disease, heart rate, respiratory rate, minimum oxygen saturation, maximum temperature, minimum Glasgow Coma Scale, blood urea nitrogen, creatinine, lactate, arterial pH, international normalised ratio, mechanical ventilation, vasopressor use and 24-hour urine output.

In the pooled multivariable model, the strongest independent associations with SAP were maximum temperature (odds ratio 2.48, 95% CI 1.72-3.57 per degree Celsius, $p < 0.001$), vasopressor use (OR 2.60, 95% CI 1.32-5.15, $p = 0.006$), mechanical ventilation (OR 2.39, 95% CI 1.07-5.34, $p = 0.034$), blood urea nitrogen (OR 1.02, 95% CI 1.01-1.04 per mg/dL, $p = 0.004$), respiratory rate (OR 1.10, 95% CI 1.03-1.18 per breath per minute, $p = 0.007$), 24-hour urine output (OR 0.96, 95% CI 0.94-0.99 per 100 mL, $p = 0.002$), minimum Glasgow Coma Scale (OR 0.88, 95% CI 0.80-0.97 per point, $p = 0.010$) and heart rate (OR 1.02, 95% CI 1.00-1.04 per beat per minute, $p = 0.042$). Thirteen of the 21 variables were retained despite a p value above 0.05. This is expected and appropriate: LASSO selects variables for their contribution to prediction rather than for statistical significance, and selecting or discarding predictors on the basis of p

values in a prediction model is known to harm performance and is discouraged by reporting guidance [15, 35].

The nomogram translating the model into a bedside tool is shown in Fig. 2.

4.5. Model performance

On the held-out test set of 221 patients, the model achieved an AUC of 0.840 (95% CI 0.786-0.895) with a Brier score of 0.157. Because that estimate rests on only 87 events and one particular split, we regard the cross-validated result as primary: across 10-fold cross-validation of the whole cohort the AUC was 0.877 (95% CI 0.852-0.903), the calibration slope was 0.884 (95% CI 0.755-1.013), the calibration intercept was -0.039 and the Brier score was 0.136 (Table 3, Fig. 3). The bootstrap-corrected AUC was 0.889, against an apparent AUC of 0.910, indicating modest optimism of 0.021. The corrected calibration slope from the bootstrap was 0.845 (95% CI 0.667-1.012), consistent with the cross-validated estimate.

Calibration was acceptable in the centre of the risk distribution and at the extremes of the observed range, with one exception. The Hosmer-Lemeshow test was statistically significant (chi-square 20.67, 8 degrees of freedom, $p=0.008$), but inspection of the deciles of predicted risk localises the discrepancy almost entirely to the highest decile, where the model predicted a mean risk of 97.0% and 66 events were observed against 71.8 expected; this single decile contributed 15.5 of the 20.7 chi-square statistic, or 75.1% of it. In the remaining nine deciles the agreement between predicted and observed risk was close, and the overall calibration slope was consistent with perfect calibration. The practical interpretation is that the model is somewhat overconfident for the small group of patients it considers almost certain to develop SAP, and that outside that group its probabilities can be taken at face value.

Decision curve analysis showed that the model provided a higher net benefit than treating all patients or none across the full range of threshold probabilities examined, and greater net benefit than the modified BISAP score (Fig. 4). Against modified BISAP (AUC 0.693, 95% CI 0.627-0.760), the model improved discrimination significantly (DeLong $p<0.001$), with a net reclassification improvement of 0.600 and an integrated discrimination improvement of 0.300.

4.6. Comparison with machine learning models

The eight algorithms achieved test-set AUCs ranging from 0.818 for k-nearest neighbours to 0.850 for random forest, a spread of only 0.032, with overlapping confidence intervals throughout (Table 5). Logistic regression achieved 0.840 and was not significantly different from the best-performing model (DeLong $p=0.494$; Fig. 6). Several algorithms showed clear overfitting: random forest reached an apparent training AUC of 1.000 against a test AUC of 0.850, the multilayer perceptron 0.976 against 0.825, and gradient boosting 0.961 against 0.843. No machine learning model achieved a test AUC that was both higher than logistic regression and statistically distinguishable from it.

4.7. Model interpretation

SHAP analysis of the extreme gradient boosting model ranked minimum Glasgow Coma Scale (mean absolute SHAP value 0.591), creatinine (0.351), mechanical ventilation (0.340), blood urea nitrogen (0.301), maximum temperature (0.251) and 24-hour urine output (0.236) as the strongest contributors (Fig. 5). The ranking agrees with the logistic model in identifying neurological, renal and respiratory status as the dominant domains, and confirms that the discriminative signal is concentrated in a small number of variables rather than distributed evenly across the 21 selected predictors.

4.8. Sensitivity analyses

Relaxing the missingness threshold to 50% brought albumin, lactate dehydrogenase, lipase and the neutrophil-to-lymphocyte and platelet-to-lymphocyte ratios into the candidate pool. None was selected by LASSO, and discrimination was unchanged (AUC 0.832 versus 0.840, $p=0.143$) (Table 6). These markers therefore added no incremental information beyond the routinely available variables already in the model, which supports their exclusion on grounds of missingness.

Removing the three organ-support variables reduced the test AUC from 0.840 to 0.816 (DeLong $p=0.016$) and increased the Brier score from 0.157 to 0.168. The model thus retains most of its discrimination from physiological and laboratory variables alone, indicating that it is not simply re-encoding organ failure through its treatment variables, although the 0.025 decrement confirms that organ-support status does carry genuine predictive information.

Restricting the analysis to the 155 patients whose ICU stay was at least 48 hours, in whom the persistence criterion could be applied to every patient, gave a prevalence of 54.8% and a test AUC of 0.823, with a calibration slope of 0.743 that was closer to ideal than in the primary analysis. The higher prevalence reflects the fact that patients with SAP almost by definition have prolonged ICU stays, so removing short stays preferentially removes patients without SAP. This analysis is limited by its sample size: 27 variables were selected for 155 patients, which exceeds the conventionally recommended number of events per parameter and makes the selection unstable.

6. Discussion

6.1. Principal findings

In a cohort of 734 patients with acute pancreatitis admitted to an intensive care unit, a model built from 21 variables available within the first 24 hours predicted persistent organ failure with a cross-validated AUC of 0.877 and a calibration slope of 0.884, indicating both good discrimination and calibration that is close to ideal. The model substantially outperformed the modified BISAP score and provided greater net benefit across the range of threshold probabilities relevant to clinical decisions. Its most striking feature, however, is comparative rather than absolute: eight machine learning algorithms, tuned and evaluated

under identical conditions, all landed within 0.032 AUC of one another, and none significantly outperformed logistic regression. The interpretable nomogram is therefore not a concession to simplicity but the equal of the alternatives on the evidence available.

Two further findings bear on how the model should be understood. First, removing mechanical ventilation, vasopressor use and renal replacement therapy from the candidate pool reduced discrimination only modestly, from 0.840 to 0.816, indicating that the model draws most of its signal from physiological and laboratory derangement rather than from the fact that organ failure was already being treated. Second, five sparsely measured biomarkers, including lipase and albumin, added nothing when offered to the selection procedure with a relaxed missingness threshold, which is consistent with the observation in Table 1 that lipase and amylase did not differ between patients who did and did not develop SAP.

6.2. Comparison with existing literature

Previous models for severe acute pancreatitis derived from MIMIC-IV have reported broadly similar discrimination. A nomogram based on persistent inflammation, immunosuppression and catabolism syndrome achieved an AUC of 0.730 in a derivation cohort of 4,930 patients [12], and a machine learning model for acute respiratory distress syndrome complicating acute pancreatitis reported a C-index of 0.818 in training and 0.799 in testing [13]. Our cross-validated AUC of 0.877 sits at the upper end of this range. Part of that advantage is likely to reflect cohort composition rather than model quality: by requiring an ICU admission we selected a population with a higher prevalence of severe disease and a wider spread of physiological derangement, both of which tend to raise discrimination. Direct comparison of AUCs across cohorts with different case mixes is unreliable, and we caution against reading our figure as evidence of superiority.

Our finding that machine learning does not outperform logistic regression is consistent with a substantial body of methodological work in clinical prediction, in which the advantage of flexible models over a well-specified regression model is typically small or absent once sample size is modest and the signal is carried by a few strong predictors [35]. It is worth stating plainly that several of the machine learning models here overfit conspicuously, with random forest separating its training data perfectly. Reporting only the apparent performance of such a model would have produced a spuriously impressive result, and the practice of comparing flexible models against unpenalised or univariable regression likely explains some of the more spectacular AUCs in this literature [14, 35].

6.3. Interpretation of predictors

The variables driving prediction are clinically coherent. Minimum Glasgow Coma Scale was the strongest contributor by SHAP analysis and one of the strongest in the logistic model, which is consistent with neurological dysfunction being a recognised marker of severe systemic inflammation in acute pancreatitis, although sedation in ventilated patients is an important confounder and the variable should not be read as a purely neurological measurement. Renal variables, blood urea nitrogen and creatinine, featured prominently, in keeping with the established prognostic importance of haemoconcentration and

early renal impairment. Blood urea nitrogen in particular is a component of the BISAP score and has repeatedly been shown to predict mortality in acute pancreatitis [9]. Maximum temperature was the strongest single association in the logistic model with an odds ratio of 2.48 per degree Celsius; this is a larger effect than would be anticipated from the classical literature and may partly reflect the fact that temperature and organ failure share inflammatory antecedents, as well as the influence of a small number of patients with high fever. The absence of lipase and amylase from the model is expected, since these enzymes establish the diagnosis rather than the severity.

6.4. Strengths and limitations

The principal strengths of this study are methodological. The outcome was derived explicitly from the revised Atlanta definition and, because more than one defensible operationalisation exists, five candidate definitions were compared and reported rather than one being chosen silently. The adopted definition avoided two specific pitfalls, namely treating fluid-responsive hypotension as organ failure and allowing persistence to be satisfied across different organ systems on consecutive days, and it was anchored from day 2 onwards so that it does not overlap the predictor window. Imputation was performed without the outcome, preventing leakage. Predictor selection used a majority rule across imputations rather than a single dataset, and the nomogram was drawn from the pooled coefficients rather than from one completed dataset, so that the figure, the regression table and the reported performance describe the same model. Internal validation carried the same multiple imputation as the final model rather than validating one completed dataset in isolation. It used both cross-validation and bootstrap, and we report a single train/test split alongside the much more precise cross-validated estimate rather than presenting the split as though it were definitive.

Several limitations should be weighed. The cohort was drawn from a single academic medical centre and the model has not been externally validated; the performance reported here is internal and, as with any model, is likely to be optimistic in a new population. Restricting the cohort to patients with an ICU stay of at least 24 hours excluded patients managed on general wards, so the model applies to patients sick enough to reach intensive care and should not be used to triage patients outside that setting. Two hundred and twenty patients (30%) had an ICU stay shorter than 48 hours and could not be assessed for persistence; they were classified as not having SAP by construction, and although only 22 of them died in hospital, this introduces some misclassification that a sensitivity analysis restricted to longer stays did not fully resolve because the resulting sample was small and the selection was non-random.

The respiratory domain of the modified Marshall score requires knowledge of whether PaO₂ was measured during mechanical ventilation, and ventilator settings are not recorded in a form that could be extracted for this cohort without reprocessing the entire chart dataset. We therefore used all available PaO₂/FiO₂ pairs rather than restricting the ratio to ventilated patients, which is the more usual convention and will have raised the apparent prevalence of respiratory organ failure relative to published estimates; the sensitivity analysis that restricted the ratio in this way yielded a lower prevalence of 28.1%, and the choice between the two definitions is reported transparently in Table 4. Cardiovascular organ failure required a further approximation, because fluid responsiveness is not recorded

retrospectively and vasopressor administration was used as its proxy. Some variables were available in only half to three quarters of patients, and although multiple imputation is appropriate under a missing-at-random assumption, that assumption cannot be verified and informative missingness, in which the sickest patients have the most missing data, would bias the model. Finally, the model predicts a composite outcome of persistent organ failure in any of three systems and does not distinguish which system will fail, which limits its usefulness for anticipating specific complications.

Fairness was not formally assessed, and this is a limitation. We did not estimate model performance separately by sex, race, ethnicity or socioeconomic status, so we cannot exclude differential performance across groups; the single-centre source and the small number of patients within some categories would in any case have left such subgroup estimates imprecise. Race and admission type were deliberately withheld from the predictor pool, so the model does not encode them directly, but a predictor such as creatinine can carry differential meaning across groups through differences in muscle mass. TRIPOD + AI identifies fairness as a reporting requirement, and we record its absence here rather than leaving it unstated.

6.5. Implications

The practical value of this work lies less in the nomogram itself than in the evidence that an interpretable model is sufficient. Clinicians considering whether to adopt a prediction tool in acute pancreatitis should not assume that a machine learning model will outperform a transparent one; in this cohort it did not. The variables in the final model are those of a routine first-day ICU assessment, suggesting that the model could be computed automatically from data already recorded, which is the condition for any prediction tool to be used in practice. Before that, external validation in an independent cohort is required, along with an assessment of whether acting on the model's predictions changes management and improves outcomes, since good discrimination does not by itself demonstrate clinical benefit.

7. Conclusions

A nomogram constructed from 21 routinely available first-24-hour variables predicted severe acute pancreatitis among intensive care patients with acute pancreatitis with a cross-validated AUC of 0.877 and a calibration slope of 0.884, exceeding the modified BISAP score. Eight machine learning algorithms performed comparably and none was significantly better than logistic regression. The model requires external validation before clinical use.

Abbreviations

ALT, alanine aminotransferase.

AP, acute pancreatitis.

APACHE II, Acute Physiology and Chronic Health Evaluation II.

AST, aspartate aminotransferase.

AUC, area under the receiver operating characteristic curve.

BISAP, Bedside Index for Severity in Acute Pancreatitis.

CI, confidence interval.

CV, cross-validation.

GBM, gradient boosting machine.

ICD, International Classification of Diseases.

ICU, intensive care unit.

IDI, integrated discrimination improvement.

INR, international normalised ratio.

IQR, interquartile range.

KNN, k-nearest neighbours.

LASSO, least absolute shrinkage and selection operator.

LR, logistic regression.

MIMIC-IV, Medical Information Mart for Intensive Care IV.

MLP, multilayer perceptron.

NB, naive Bayes.

NRI, net reclassification improvement.

OR, odds ratio.

RF, random forest.

SAP, severe acute pancreatitis.

SD, standard deviation.

SHAP, Shapley additive explanations.

SMD, standardised mean difference.

SOFA, Sequential Organ Failure Assessment.

SpO₂, peripheral oxygen saturation.

SVM, support vector machine.

TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, artificial intelligence extension.

XGBoost, extreme gradient boosting.

Declarations

8.1. Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. The MIMIC-IV database was approved by the Institutional Review Board of the Beth Israel Deaconess Medical Center, which waived the requirement for individual informed consent for the collection and de-identified sharing of these data. All analyses used de-identified data. Access was obtained under the PhysioNet credentialed data use agreement: the credential holder (GZ) completed the CITI 'Data or Specimens Only Research' course (completion record 79656399, completed 13 September 2026) and signed the MIMIC-IV data use agreement on PhysioNet. As this study involved only retrospective analysis of a publicly available de-identified database, no further ethical approval was required.

8.2. Consent for publication

Not applicable.

8.3. Availability of data and materials

This study is a secondary analysis of an existing database. No new data were generated or collected; what the analysis produced is a derived analytic cohort, a fitted model and the tables and figures reported here. MIMIC-IV version 2.2 [17] is distributed by PhysioNet at <https://physionet.org/content/mimiciv/2.2/> under the MIMIC-IV data use agreement. Access is granted to any user who completes the required training and signs that agreement, so the source data are not restricted to the author. The record-level data are not redistributed here, and the derived cohort is not separately deposited, because the agreement prohibits sharing record-level extracts; the cohort can be reconstructed exactly by any credentialed user who runs the analysis code against MIMIC-IV version 2.2. The analysis code is available from the corresponding author on reasonable request. All coefficients needed to apply the model are given in Table 2.

8.4. Competing interests

The author declares that they have no competing interests.

8.5. Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

8.6. Authors' contributions

Gaozhi Zhang conceived and designed the study, extracted and cleaned the data, performed the statistical analysis and model development, interpreted the results, and drafted the manuscript. The author read and approved the final manuscript.

8.7. Protocol

No study protocol was prepared or registered before this analysis; the analysis plan was specified by the author and the sensitivity analyses were prespecified before the models were fitted.

8.8. Registration

This study was not registered. It is a secondary analysis of an existing de-identified database and did not prospectively enrol participants.

8.9. Patient and public involvement

There was no patient or public involvement in the design, conduct, reporting, interpretation or dissemination of this study. The analysis used an existing de-identified database and no participants were recruited.

8.10. Acknowledgements

Not applicable.

References

1. Peery AF, Crockett SD, Murphy CC, Lund JL, Dellon ES, Williams JL, et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2018. *Gastroenterology*. 2019;156(1):254–e27211.
2. Yadav D, Lowenfels AB. The epidemiology of pancreatitis and pancreatic cancer. *Gastroenterology*. 2013;144(6):1252–61.
3. Banks PA, Bollen TL, Dervenis C, Gooszen HG, Johnson CD, Sarr MG, et al. Classification of acute pancreatitis – 2012: revision of the Atlanta classification and definitions by international consensus. *Gut*. 2013;62(1):102–11.
4. Forsmark CE, Vege SS, Wilcox CM. Acute pancreatitis. *N Engl J Med*. 2016;375(20):1972–81.
5. Johnson CD, Abu-Hilal M. Persistent organ failure during the first week as a marker of fatal outcome in acute pancreatitis. *Gut*. 2004;53(9):1340–4.
6. Tenner S, Baillie J, DeWitt J, Vege SS. American College of Gastroenterology guideline: management of acute pancreatitis. *Am J Gastroenterol*. 2013;108(9):1400–15.
7. Working Group IAP/APA Acute Pancreatitis Guidelines. IAP/APA evidence-based guidelines for the management of acute pancreatitis. *Pancreatology*. 2013;13(4 Suppl 2):e1–15.
8. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med*. 1985;13(10):818–29.
9. Wu BU, Johannes RS, Sun X, Tabak Y, Conwell DL, Banks PA. The early prediction of mortality in acute pancreatitis: a large population-based study. *Gut*. 2008;57(12):1698–703.
10. Marshall JC, Cook DJ, Christou NV, Bernard GR, Sprung CL, Sibbald WJ. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. *Crit Care Med*. 1995;23(10):1638–52.
11. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Med*. 1996;22(7):707–10.
12. Liu D, Liu H, Yao J, Chen G, Wang W, Cao B, et al. Development and validation of a nomogram model for assessing the severity of acute pancreatitis from the perspective of PICS. *Front Nutr*. 2025;12:1611501.
13. Hu H, Li D, et al. Clinical use of nomogram based on machine learning for diagnosis prediction of acute respiratory distress syndrome in patients with acute pancreatitis. *Mediators Inflamm*. 2025;2025:5610316.
14. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
15. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
16. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*.

2015;350:g7594.

17. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1.
18. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1–67.
19. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
20. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
21. Ensor J, Martin EC, Riley RD. pmsampsize: calculates the minimum sample size required for developing a multivariable prediction model. R package version 1.1.3. 2022.
22. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. 1996;58(1):267–88.
23. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–45.
25. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken: Wiley; 2013.
26. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–76.
27. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565–74.
28. Pencina MJ, D'Agostino RB, Sr, Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157–72.
29. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54(8):774–81.
30. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1–26.
31. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2016. pp. 785–794.
32. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–74.
33. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.
34. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2025.

35. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Cham: Springer; 2019.

Tables

Tables 1 to 6 are available in the supplementary files section.

Figures

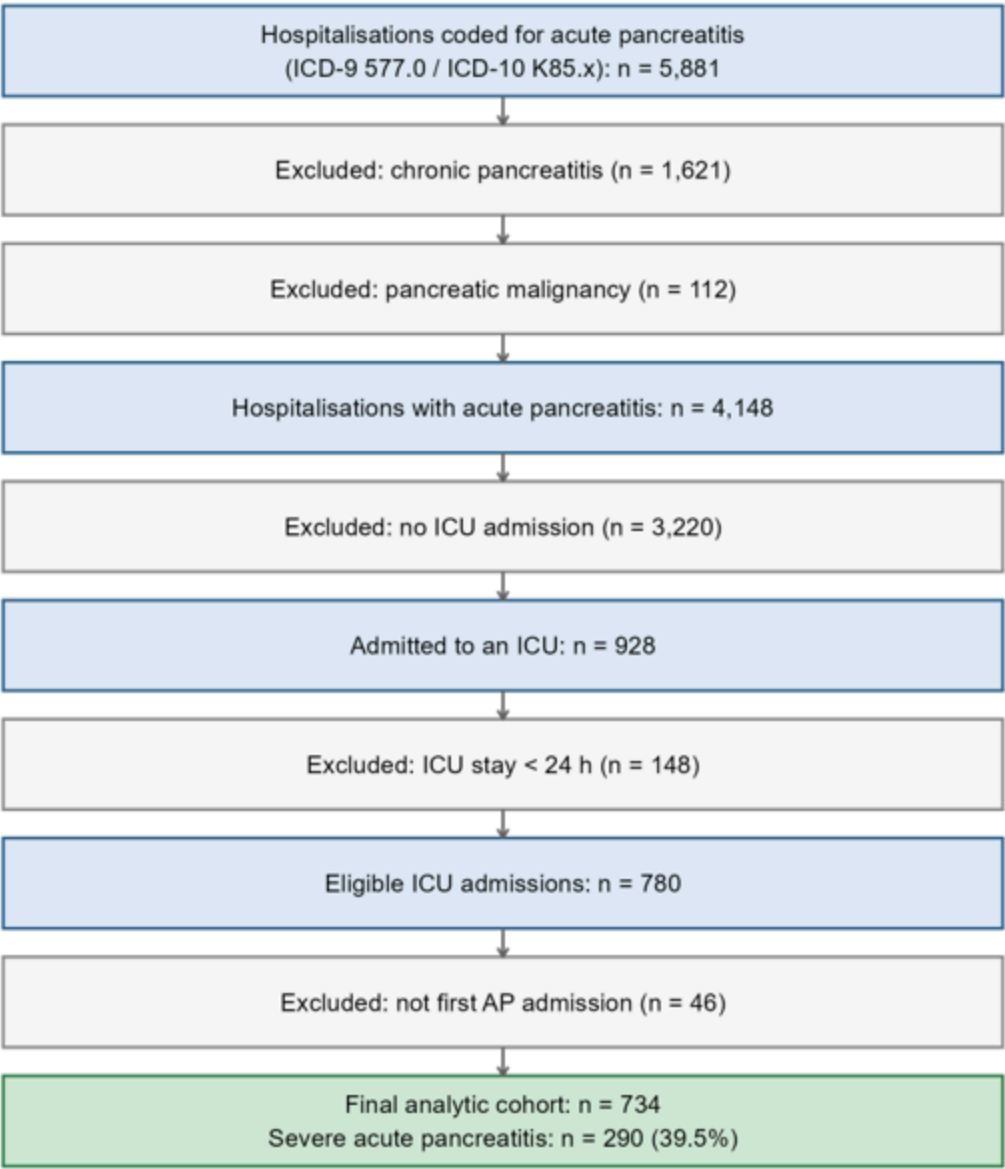


Figure 1

Cohort derivation

AP, acute pancreatitis; ICU, intensive care unit.

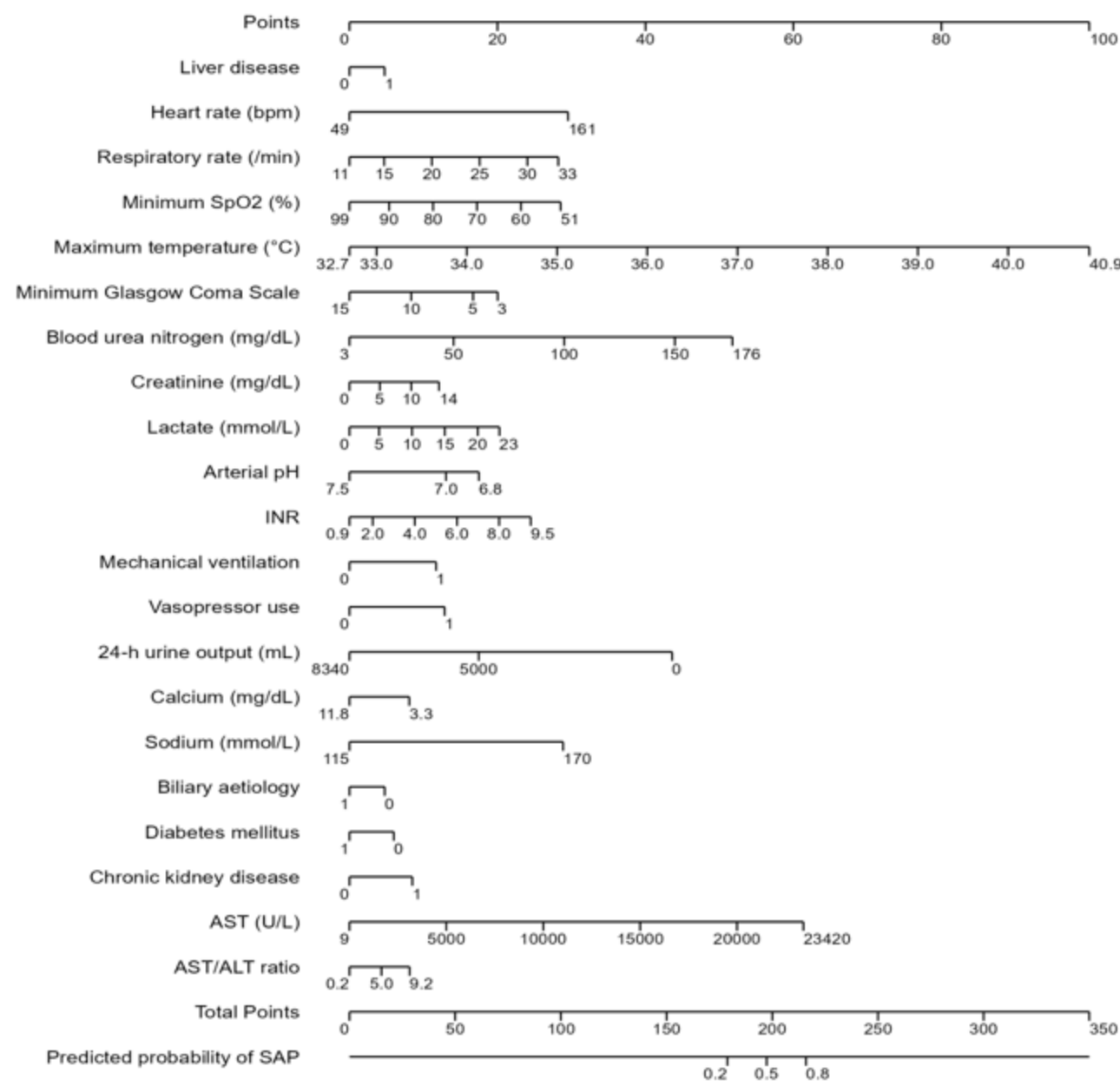


Figure 2

Nomogram for predicting severe acute pancreatitis

Individual patient values are located on each variable axis, points are read from the top scale and summed, and the total is mapped to the predicted probability of severe acute pancreatitis on the bottom scale. Total points are shown over the range reached by the cohort; the theoretical maximum is higher because it sums every variable at its riskiest value. The model is linear in each variable on the scale shown. The AST axis is compressed by a small number of very high values: its interquartile range (41 to 200 U/L) occupies only the first half point of that axis, so AST should be read against its printed ticks

rather than by eye, and its contribution to the total score is correspondingly small. The calcium and 24-hour urine axes are short for the same reason.

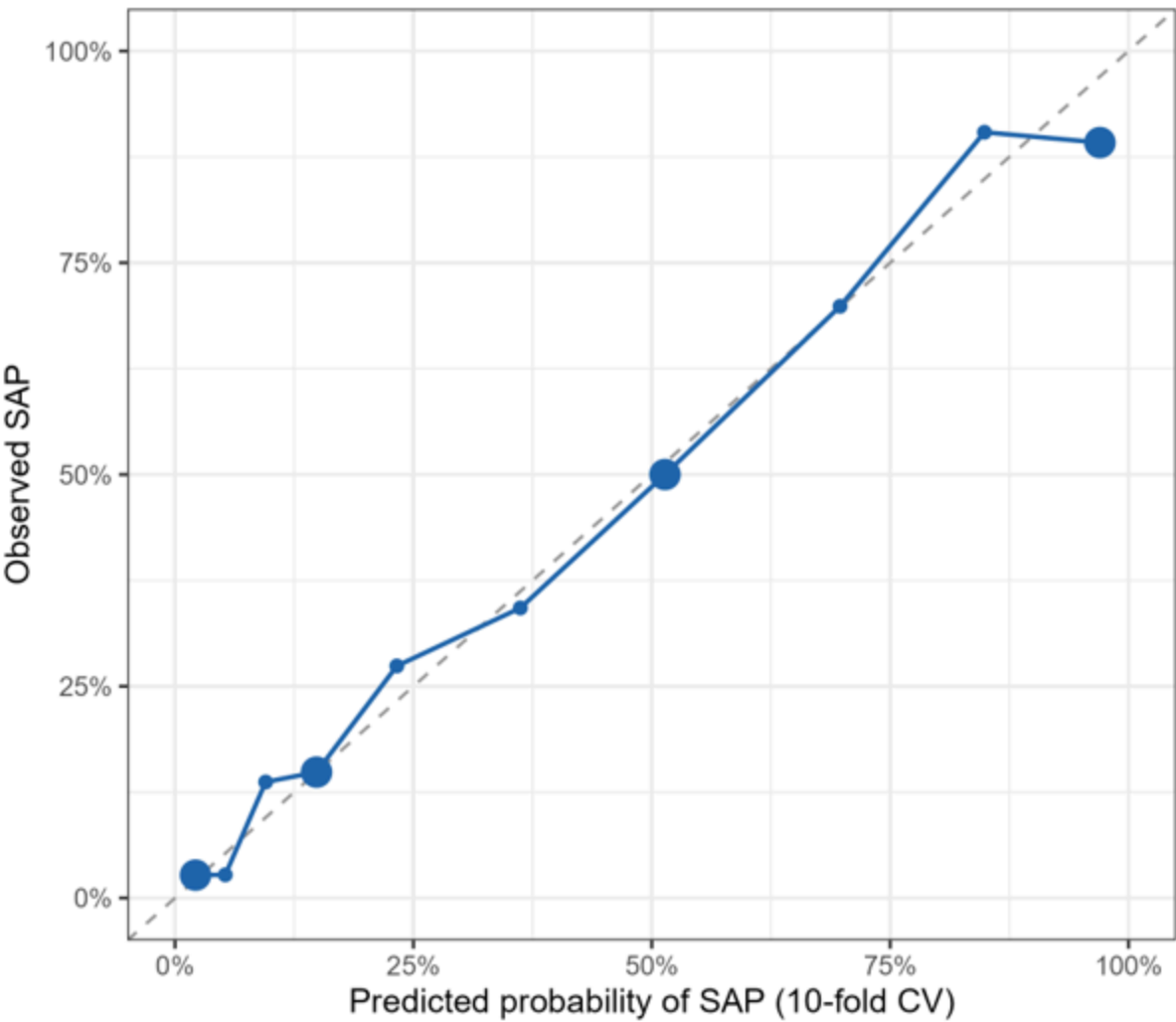


Figure 3

Cross-validated calibration of the nomogram

Points show observed against predicted risk by decile of predicted risk, on pooled out-of-fold predictions from 10-fold cross-validation; point size is proportional to the number of patients in the decile. The dashed line is perfect calibration.

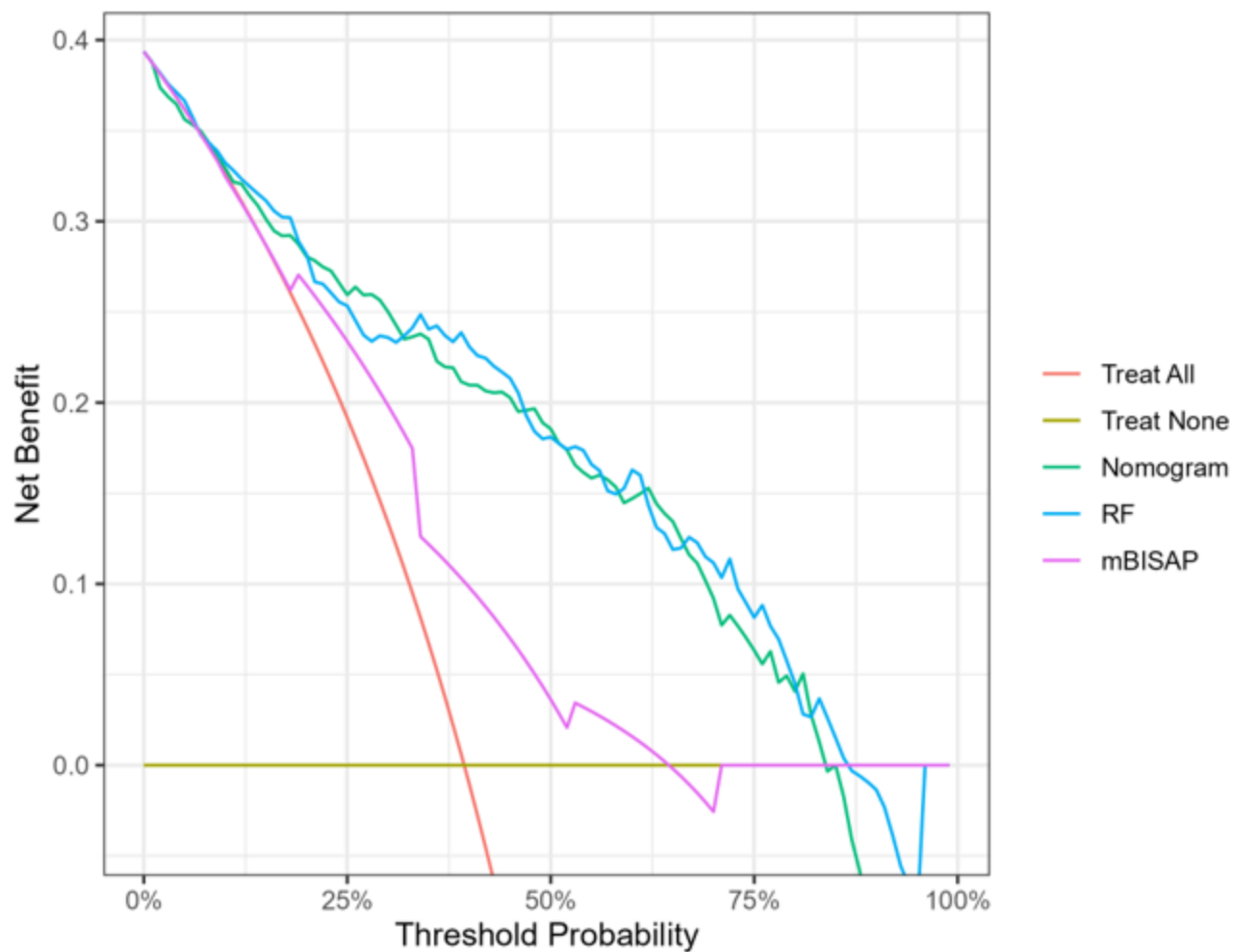


Figure 4

Decision curve analysis

Net benefit is plotted against threshold probability for the nomogram, the best-performing machine learning model and the modified BISAP score, with the strategies of treating all and treating no patients as references. BISAP, Bedside Index for Severity in Acute Pancreatitis.

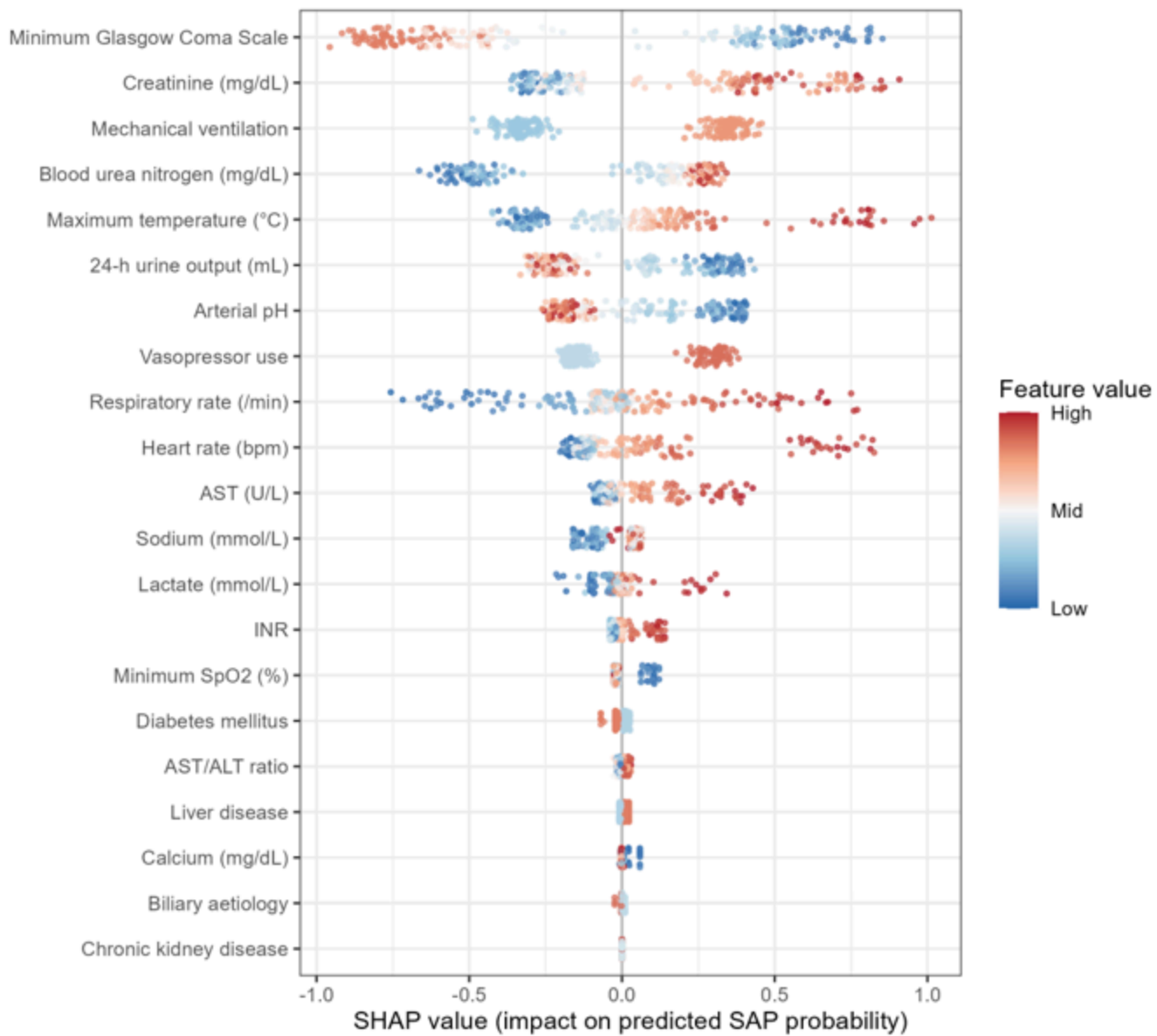


Figure 5

SHAP summary plot for the extreme gradient boosting model

Each point is one patient; the horizontal position is the SHAP value, indicating the direction and magnitude of that variable's contribution to the predicted probability of severe acute pancreatitis, and colour encodes the variable value from low to high. SHAP, Shapley additive explanations.

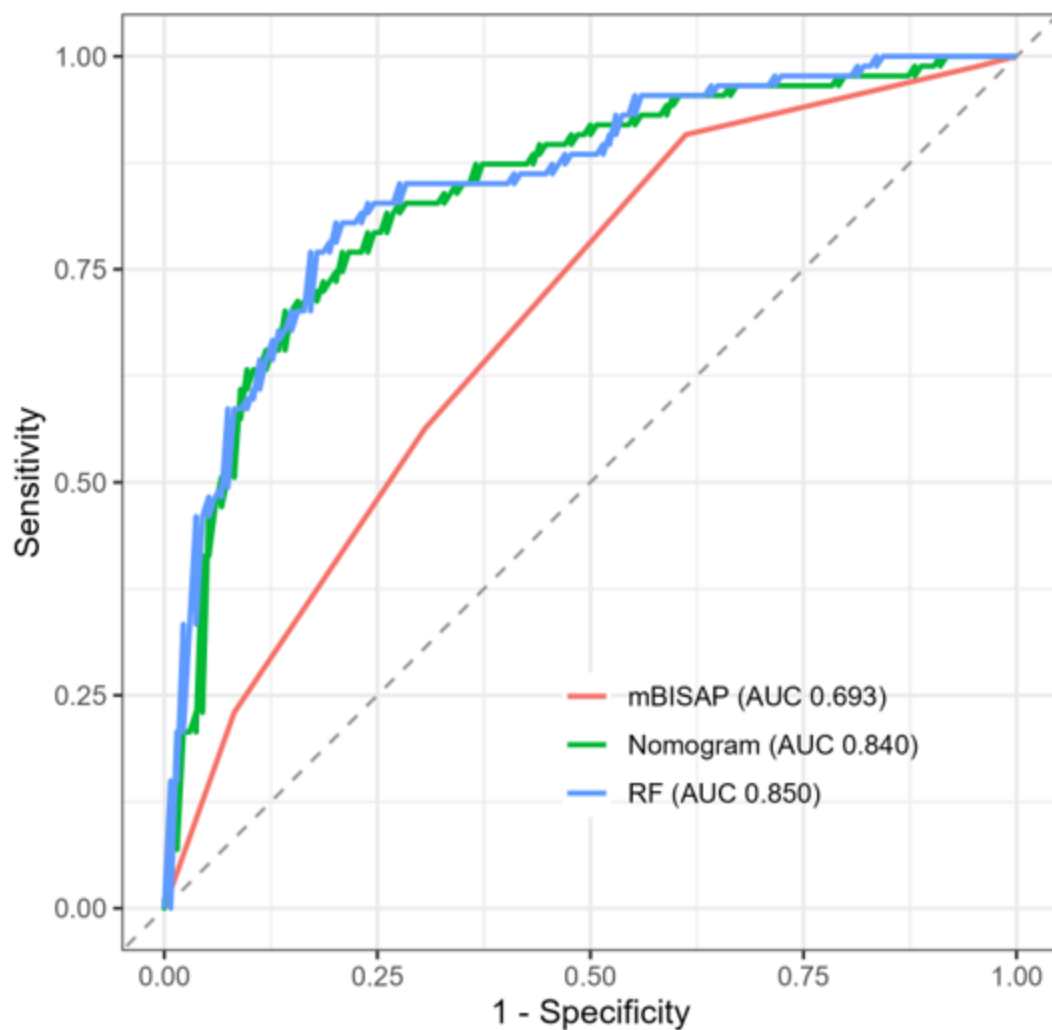


Figure 6

Receiver operating characteristic curves on the test set

Curves are shown for the nomogram, the best-performing machine learning model and the modified BISAP score, with the line of no discrimination as reference. AUC, area under the receiver operating characteristic curve; RF, random forest.

Supplementary Files

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

- [Tables.docx](#)