Research Square

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

Development of a machine learning-based prediction model for objective indications for renal replacement therapy after stage 3 acute kidney injury

Chaodong Chen
Beijing Chao-Yang Hospital, Capital Medical University
PengChao Tian
Beijing Chao-Yang Hospital, Capital Medical University
Xin Chen
University College London
FangXing Guo
Beijing Chao-Yang Hospital, Capital Medical University
WenXiong Li
liwenxiong@mail.ccmu.edu.cn
Beijing Chao-Yang Hospital, Capital Medical University

Article

Keywords:

Posted Date: June 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10020586/v1

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

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on September 20th, 2026. See the published version at https://doi.org/10.1038/s41598-026-72758-7.

Abstract

Renal replacement therapy (RRT) is considered when severe acute kidney injury (AKI) is complicated by severe electrolyte disturbance, acid-base derangement, fluid overload with respiratory compromise, or azotemia, but actual RRT initiation is influenced by clinician and center practice. We developed and externally validated machine-learning models to predict objective indications for RRT within 24–72 h after creatinine- or urine-output-defined stage 3 AKI. The outcome was defined as new severe hyperkalemia, severe metabolic acidosis, fluid overload with hypoxemia, or severe azotemia within 24–72 h, corresponding to serum potassium > 6.0 mmol/L, pH ≤ 7.20 with bicarbonate ≤ 12 mmol/L, cumulative fluid balance > 10% of body weight with PaO_2/FiO_2 ≤200, or blood urea nitrogen > 112 mg/dL. MIMIC-IV was used for model development and internal validation, and eICU-CRD was used for external validation. Predictors were routinely collected variables measured during the first 24 h after the stage 3 AKI index time. Objective RRT indications occurred in 185 of 2,126 MIMIC-IV patients (8.7%) and 121 of 1,916 eICU-CRD patients (6.3%). The primary Gradient Boosting model achieved AUROC/AUPRC/Brier scores of 0.848/0.459/0.062 internally and 0.764/0.181/0.065 externally, although it overestimated absolute risk in external validation. The model may support short-term risk stratification and structured reassessment after stage 3 AKI, but local recalibration and prospective validation are required before clinical implementation.

Introduction

Acute kidney injury is common among critically ill patients and is associated with excess mortality, prolonged organ support, and long-term kidney dysfunction[1, 2]. Stage 3 AKI represents the most severe category of AKI, but it remains clinically heterogeneous. Some patients recover without kidney support, whereas others develop hyperkalemia, metabolic acidosis, severe azotemia, or fluid overload that prompts consideration of renal replacement therapy (RRT)[3].

Although RRT can correct life-threatening metabolic and fluid disturbances, it also exposes patients to vascular access complications, hemodynamic instability, anticoagulation-related bleeding, infection, and potential overtreatment. Randomized trials comparing earlier with delayed or standard RRT initiation strategies have not established a single timing rule that applies to all critically ill patients with severe AKI[4–7].

However, delayed or standard strategies in these trials were not strategies of indefinite deferral. They incorporated prespecified rescue or safety criteria, including severe hyperkalemia, severe metabolic acidosis, fluid overload or pulmonary edema with respiratory compromise, and severe azotemia, beyond which continued observation without RRT was no longer clinically appropriate. These criteria provide a pragmatic basis for defining objective RRT indications that are less dependent on local treatment practice than actual RRT initiation[4–7].

Existing prediction studies have examined actual RRT initiation or continuous RRT requirement in patients with sepsis, rhabdomyolysis, and broader ICU cohorts[8–11]. These studies show that routinely

collected clinical variables can support risk stratification. However, actual RRT initiation is a treatment decision influenced by clinician preference, local practice, resource availability, vascular-access feasibility, hemodynamic tolerance, and goals of care. Models trained to predict actual treatment use may therefore capture practice patterns as well as biological deterioration, and may not directly identify patients who are about to cross objective safety thresholds for RRT consideration.

We therefore developed and externally validated machine-learning models to predict objective renal replacement therapy need (objective RRT-NEED), defined as trial-informed objective indications for RRT, within 24–72 h after creatinine- or urine-output-defined stage 3 AKI. Predictors were restricted to routinely collected variables measured during the first 24 h after the stage 3 AKI index time. Model development and internal validation were performed in MIMIC-IV, with external validation in the eICU Collaborative Research Database (eICU-CRD).

Methods

Data sources and study design

We performed a retrospective prediction-model development and external validation study using two public critical care databases. MIMIC-IV was used for model development and internal validation, and the eICU-CRD was used only for external validation[12, 13]. No eICU-CRD data were used for imputation fitting, feature selection, hyperparameter tuning, or model training. Both databases contain de-identified ICU electronic health record data and have existing institutional approvals.

Reporting guideline adherence and risk-of-bias assessment

The study was reported with reference to TRIPOD + AI and internally assessed using PROBAST + AI[14, 15]. The analytic pipeline was designed to avoid information leakage: imputation, feature selection, hyperparameter tuning, and model training were performed within the MIMIC-IV training data, and eICU-CRD was used only for external validation.

Study population and time zero

Adult patients were eligible if they had a first ICU stay of at least 24 h and developed stage 3 AKI by Kidney Disease: Improving Global Outcomes-compatible creatinine or urine-output criteria[3]. RRT-based stage 3 AKI was not used to define the index time. The stage 3 AKI index time was the first time at which the creatinine or urine-output criterion for stage 3 AKI was met.

We excluded patients with chronic kidney disease or end-stage renal disease, patients who had received RRT before the stage 3 AKI index time, patients who received RRT during the first 24 h after the index time, and patients who had already met objective RRT-NEED during the first 24 h after the index time. Patients with invalid body weight or insufficient records for predictor or outcome ascertainment were

also excluded. Documented body weight values outside clinically plausible ranges were treated as invalid before calculating fluid balance relative to body weight.

Prediction time, predictor window, and outcome window

The prediction time was 24 h after the stage 3 AKI index time. Candidate predictors were restricted to variables measured from 0–24 h after the index time. The outcome was assessed from 24–72 h after the index time. Variables measured after the prediction time were not used as predictors.

Outcome definition and intercurrent RRT

The primary outcome was the development of objective indications for RRT within 24–72 h after the stage 3 AKI index time. Objective RRT-NEED was defined as the occurrence of at least one prespecified biochemical or physiological criterion: serum potassium > 6.0 mmol/L; pH ≤ 7.20 together with bicarbonate ≤ 12 mmol/L; cumulative fluid balance > 10% of documented body weight together with PaO_2/FiO_2 ≤200; or blood urea nitrogen > 112 mg/dL. These components were selected to reflect rescue or safety criteria used in major randomized trials of RRT timing in severe AKI, covering electrolyte, acid-base, fluid overload with respiratory compromise, and azotemic indications.

Actual RRT initiation was not included as an outcome component because it represents a treatment decision that may be influenced by local practice, resource availability, perceived reversibility, and goals of care. Death without recorded objective criteria was also not counted as objective RRT-NEED. Patients who initiated actual RRT during the 24- to 72-h outcome window before meeting any recorded objective criterion were excluded from the primary analysis, because RRT can modify subsequent potassium, acid-base status, blood urea nitrogen, fluid balance, and oxygenation, potentially preventing later observation of objective criteria.

Candidate predictors and missing data

Candidate predictors were routinely available variables measured during the first 24 h after the stage 3 AKI index time. Variables included demographics, AKI phenotype, severity of illness, organ support, vital signs, renal and fluid variables, electrolyte and acid-base variables, and hematological variables. Variables with more than 30% missingness in the MIMIC-IV training set or more than 50% missingness in eICU-CRD were excluded from the primary model.

Multiple imputation was performed after cohort splitting. The imputation model was fitted in the MIMIC-IV training set only and then applied to the MIMIC-IV internal test set and the eICU-CRD external validation cohort, thereby avoiding the use of validation-set information during preprocessing. Five imputed datasets were generated. Model development and evaluation were performed across the imputed datasets, and pooled performance estimates were reported. Imputation diagnostics,

missingness profiles, and completed-data stability checks are provided in the Supplementary Information.

Model development and validation

The MIMIC-IV cohort was split into training and internal test sets using a 7:3 stratified split by objective RRT-NEED. Feature selection was conducted within the imputed MIMIC-IV training set using recursive feature elimination. Feature-selection stability across imputed datasets was summarized as the frequency with which each predictor was selected. Pairwise correlations among candidate predictors were inspected to assess collinearity. Highly correlated variables were not automatically removed when they represented clinically distinct extrema or trajectories within the 24-h predictor window, but collinearity was considered when interpreting feature importance and SHAP values. Multiple candidate algorithms were implemented using scikit-learn and compared, including Gradient Boosting, L2-penalized logistic regression, and extremely randomized trees[16, 17]. Hyperparameters were tuned using randomized search with fivefold cross-validation and average precision as the optimization metric. Gradient Boosting was selected as the primary model, with L2-penalized logistic regression and extremely randomized trees retained as comparator models.

Model evaluation and interpretation

Discrimination was evaluated using AUROC and AUPRC. AUPRC was emphasized because objective RRT-NEED was infrequent and precision-recall metrics are informative for imbalanced binary outcomes[18]. Calibration was assessed using Brier score, calibration plots, calibration intercept, calibration slope, mean predicted risk, and observed-to-expected ratio, because calibration is a core property of clinical prediction models when predicted probabilities may inform risk stratification[19].

Threshold-specific sensitivity, specificity, positive predictive value, and negative predictive value were reported at clinically plausible risk thresholds. Decision-curve analysis was performed for the primary Gradient Boosting model to evaluate threshold-dependent net benefit across risk thresholds from 0.01 to 0.30, with comparison against treat-all and treat-none reference strategies[20]. DCA was interpreted descriptively as an assessment of potential clinical usefulness rather than as evidence supporting RRT initiation. Uncertainty for model performance was estimated using patient-level bootstrap resampling, with performance estimates recalculated in each bootstrap sample and summarized as percentile-based 95% confidence intervals. Model interpretation was performed using SHAP values[21].

Statistical analysis

Continuous variables were summarized as median (interquartile range), and categorical variables were summarized as number (percentage). Baseline differences by outcome status were described using absolute standardized mean differences rather than formal significance testing. Overall characteristics

of the MIMIC-IV and eICU-CRD analytic cohorts were also summarized descriptively to assess case-mix differences between the model-development and external-validation populations. Formal P values were not used for baseline comparisons, predictor screening, model selection, or interpretation of model findings because this was a prediction-model study rather than a hypothesis-testing or causal-inference study. Candidate predictors were selected based on clinical availability, missingness criteria, and model-based feature selection within the training data rather than univariable statistical significance. The MIMIC-IV cohort was divided into training and internal test sets in a 7:3 ratio with stratification by objective RRT-NEED, and the eICU-CRD cohort was used for external validation. Model performance measures were reported for both internal and external validation cohorts. All statistical analyses and model development were performed using Python 3.12 with pandas, NumPy, scikit-learn, matplotlib, and SHAP packages.

Ethics approval and consent to participate

The MIMIC-IV database was approved for data sharing by the Institutional Review Board of Beth Israel Deaconess Medical Center, with a waiver of informed consent for the de-identified research resource. The eICU-CRD is a publicly available, de-identified multicenter critical care database released for research use after removal of protected health information. All data used in this study were de-identified, and individual informed consent was not required. All methods were performed in accordance with relevant guidelines and regulations.

Results

Study population and endpoint frequency

The final MIMIC-IV development cohort included 2,126 patients, of whom 185 (8.7%) developed objective RRT-NEED during 24–72 h after the stage 3 AKI index time. The eICU-CRD external validation cohort included 1,916 patients, of whom 121 (6.3%) developed objective RRT-NEED. These event rates reflect new objective RRT indications emerging after the first 24 h following stage 3 AKI onset, rather than overall RRT initiation after severe AKI. Detailed cohort flow and safety checks are provided in Supplementary Table S1.

The composite endpoint comprised severe azotemia, severe hyperkalemia, severe metabolic acidosis, and fluid overload with hypoxemia. In mutually exclusive hierarchical descriptive counts, severe azotemia was the first recorded component in 64 patients in MIMIC-IV and 54 patients in eICU-CRD; severe hyperkalemia was the first recorded component in 56 and 41 patients, respectively. Component frequencies are reported separately to clarify the composition of the endpoint. Actual RRT after objective RRT-NEED and death without recorded objective RRT-NEED were reported descriptively rather than counted as outcome components (Supplementary Table S3).

Selected cohort characteristics by outcome status are shown in Table 1, and full baseline characteristics by outcome status are provided in Supplementary Table S2. In both cohorts, patients who developed objective RRT-NEED generally had more severe renal dysfunction, lower urine output, higher blood urea

nitrogen, higher serum creatinine, and greater overall illness severity during the predictor window than patients without objective RRT-NEED. Overall characteristics of the MIMIC-IV and eICU-CRD analytic cohorts were also compared to describe case-mix differences between the model-development and external-validation populations (Supplementary Table S14).

Model development and selected predictors

Recursive feature elimination within the imputed MIMIC-IV training set selected 10 predictors for the final model: maximum and minimum blood urea nitrogen, maximum and minimum serum creatinine, urine output during the predictor window, maximum platelet count, minimum platelet count, maximum heart rate, minimum serum potassium, and maximum white blood cell count. Feature-selection stability is reported in Supplementary Table S8.

Model performance and calibration

Gradient Boosting was selected as the primary model because it provided balanced overall performance across discrimination and calibration in internal validation, while L2-penalized logistic regression and extremely randomized trees were retained as comparator models. ROC curves, precision-recall curves, and calibration plots are shown in Fig. 2. Performance of the selected models is shown in Table 2.

In the MIMIC-IV internal test set, the Gradient Boosting model achieved an AUROC of 0.848 (95% CI 0.790-0.896), an AUPRC of 0.459 (95% CI 0.349-0.599), and a Brier score of 0.062 (95% CI 0.050-0.078). In eICU-CRD external validation, the corresponding values were 0.764 (95% CI 0.717-0.804), 0.181 (95% CI 0.141-0.258), and 0.065 (95% CI 0.057-0.073). External validation preserved risk ranking but showed calibration drift: the mean predicted risk in eICU-CRD was approximately 12.5%, whereas the observed event rate was 6.3%. Calibration deciles are reported in Supplementary Table S9.

Threshold performance, decision-curve analysis, and interpretation

Threshold-specific performance of the primary model is shown in Table 3. In eICU-CRD, a 10% risk threshold yielded sensitivity of 0.777, specificity of 0.644, PPV of 0.128, and NPV of 0.977. At a 15% threshold, sensitivity was 0.595, specificity was 0.795, PPV was 0.164, and NPV was 0.967. Decision-curve analysis showed limited and threshold-dependent net benefit for the primary Gradient Boosting model (Fig. 3). In the MIMIC-IV internal validation cohort, net benefit remained positive across the evaluated thresholds, decreasing from 0.058 at a 0.05 threshold to 0.018 at a 0.30 threshold. In the eICU-CRD external validation cohort, net benefit was smaller and was mainly confined to low thresholds, decreasing from 0.025 at a 0.05 threshold to 0.004 at a 0.15 threshold and becoming negative at thresholds of 0.20 and 0.30. These findings do not support use of the model as a stand-alone decision rule for RRT initiation.

SHAP analysis showed broadly consistent predictor patterns across the internal and external validation cohorts. Higher maximum blood urea nitrogen and higher maximum serum creatinine increased predicted risk, reflecting azotemia and impaired kidney clearance (Fig. 4). Lower urine output during the

predictor window also increased predicted risk, consistent with persistent oliguria. Higher maximum heart rate and higher maximum white blood cell count contributed to higher predicted risk, likely reflecting systemic stress or inflammation, whereas platelet-count features may capture hematologic dysfunction and overall severity of critical illness. SHAP values were interpreted as model-level contributions rather than causal effects.

Discussion

In this multicohort prediction-model study, we developed and externally validated machine-learning models to predict objective indications for RRT within 24–72 h after creatinine- or urine-output-defined stage 3 AKI. The primary Gradient Boosting model showed good internal performance and moderate external performance, although external validation demonstrated overprediction. These findings support the model as a short-term risk-stratification and structured-reassessment tool rather than as a stand-alone rule for RRT initiation.

The definition of objective RRT indications is central to the study and was deliberately anchored in the rescue or safety criteria used in major randomized trials of RRT timing in severe AKI[4–7]. In those trials, delayed or standard strategies were not passive non-treatment strategies; continued observation was allowed only until prespecified biochemical, acid-base, fluid overload with respiratory compromise, or azotemic criteria were met. RRT was generally initiated or strongly recommended when patients developed complications such as severe hyperkalemia, severe metabolic acidosis, pulmonary edema or fluid overload with respiratory compromise, or severe azotemia. These criteria define a clinically meaningful safety boundary at which the rationale for further RRT deferral becomes weaker. Our endpoint was designed to predict future crossing of this trial-informed boundary, rather than whether clinicians actually initiated RRT.

This distinction is clinically important. Actual RRT initiation is both a clinical event and a treatment decision. It may reflect physiological deterioration, but it is also influenced by local practice thresholds, staffing and machine availability, vascular access feasibility, hemodynamic tolerance, goals of care, and perceived reversibility of kidney injury. A model trained to predict actual RRT use may therefore learn practice behavior as well as biological deterioration. By contrast, objective RRT-NEED captures prespecified biochemical and physiological states that commonly prompt RRT consideration across RRT-timing trials. The model should therefore be interpreted as estimating short-term risk of developing objective RRT indications, not as estimating the causal effect of RRT or recommending automatic RRT initiation.

The composite nature of objective RRT-NEED requires careful interpretation. Its components differ in immediacy and clinical severity: severe hyperkalemia, severe acidosis, and fluid overload with hypoxemia are typically more acute warning states, whereas severe azotemia may reflect persistent biochemical deterioration rather than an immediately life-threatening event. We therefore reported component frequencies separately in addition to the composite endpoint. The composite was used because the

clinical prediction question was whether a patient would develop any objective indication that commonly prompts RRT consideration during the subsequent 24–72 h, rather than which single complication would occur. These criteria should be interpreted as objective warning states, not as automatic indications for immediate RRT in every patient[4–7].

The stage 3 AKI index time is another important design feature. Many previous ICU prediction models use ICU admission as time zero. That approach is useful for broad early warning, but it mixes patients who never develop severe AKI with those already in a high-risk kidney phenotype. In practice, the RRT question becomes most urgent after severe AKI is recognized. By anchoring prediction to creatinine- or urine-output-defined stage 3 AKI and using the first 24 h after that index time as the predictor window, the model targets a narrower and more clinically actionable decision interval. This design also avoids using RRT itself to define stage 3 AKI, which would create circularity for an RRT-related prediction task.

Our findings should be interpreted in the context of prior RRT, CRRT, and AKI-progression prediction studies. Previous models in rhabdomyolysis, sepsis, and general ICU cohorts have shown that routinely collected clinical variables can identify patients who are more likely to receive kidney support[8–11]. However, most of these studies used actual RRT or CRRT initiation as the prediction target. Actual treatment use is clinically important, but it is also affected by clinician judgment, center-specific thresholds, resource availability, vascular-access feasibility, hemodynamic tolerance, and goals of care. Prediction models trained on actual RRT initiation may therefore learn local treatment behavior as well as biological deterioration.

Related AKI-progression models address a different clinical question. These models often predict progression to stage 3 AKI or severe AKI from an earlier ICU time point[22]. By contrast, our study starts after creatinine- or urine-output-defined stage 3 AKI has already occurred, which is the point at which RRT-related reassessment becomes clinically urgent. Biomarker-based studies and severe-AKI prediction models further suggest that RRT requirement is not driven by one isolated marker but reflects multiple domains, including kidney clearance, oliguria, metabolic derangement, inflammation, and overall critical illness severity[23,24]. The selected predictors and SHAP patterns in our model are consistent with this multidomain pathophysiology.

The present study therefore extends prior work in two ways. First, it separates the prediction target from actual treatment initiation by using trial-informed objective RRT indications as the endpoint. Second, it focuses on the short interval after stage 3 AKI onset and uses only information available during the first 24 h after that index time. This design is intended to support structured reassessment after severe AKI rather than broad ICU admission-based risk prediction or automatic RRT initiation.

The external validation results provide information beyond internal discrimination alone. In eICU-CRD, discrimination was retained, but absolute risk was overestimated. Differences in case mix between MIMIC-IV and eICU-CRD, including differences in body weight distribution, AKI phenotype, illness severity, urine output, vasopressor use, and laboratory profiles, may have contributed to this calibration drift. The comparison of overall cohort characteristics supports interpreting external validation as a test of

transportability across heterogeneous ICU populations rather than as a direct replication in an identical population. The model therefore appears to capture a partially transportable risk ordering, but its predicted probabilities require local recalibration before implementation. It should not be transported to a new ICU as an uncalibrated probability calculator.

The low event rate also requires careful interpretation. The outcome should not be interpreted as the overall frequency of RRT initiation among patients with stage 3 AKI. Patients who had already received RRT before or during the first 24 h after stage 3 AKI onset, those who already met objective RRT-NEED during the predictor window, and those who received actual RRT before any recorded objective criterion during the outcome window were excluded. The endpoint therefore represents new objective indications arising between 24 and 72 h among patients who remained under observation after the first post-index day. This design makes the prediction task more specific, but it also lowers outcome prevalence and directly affects PPV, AUPRC, and decision-curve results. Accordingly, the modest external AUPRC and low PPV limit the model's ability to rule in an immediate need for RRT, whereas the high NPV at low-to-moderate thresholds supports a potential role in short-term risk stratification and structured bedside reassessment.

Decision-curve analysis further constrained the clinical interpretation of the model. Although internal validation suggested positive net benefit across the evaluated threshold range, external validation showed only small net benefit at low thresholds and no net benefit at higher thresholds. This pattern is consistent with the low event rate, modest external AUPRC, low PPV, and overestimation of absolute risk in the external cohort. Because decision-curve analysis is sensitive to outcome prevalence and calibration, these findings argue against using raw predicted probabilities as direct treatment thresholds. The model's more plausible role is therefore to support standardized monitoring, structured reassessment, and early multidisciplinary review after severe AKI, rather than to determine RRT initiation. The final decision to initiate RRT must remain individualized and incorporate clinical trajectory, reversibility, hemodynamic tolerance, vascular-access feasibility, and patient goals of care.

The selected predictors and SHAP patterns were clinically plausible. Blood urea nitrogen and serum creatinine capture azotemia and impaired excretory function; urine output reflects oliguria and persistent kidney failure; potassium directly relates to an RRT-triggering complication; and heart rate, white blood cell count, and platelet count likely capture systemic stress, inflammation, shock physiology, and hematologic dysfunction. These variables are not novel biomarkers, but they are routinely available and align with the multidomain nature of deterioration after severe AKI. SHAP values are model-level explanations rather than causal estimates and should not be interpreted to mean that modifying a predictor would necessarily change the probability of objective RRT-NEED.

This study has several strengths. It used two public ICU databases, separated model development from external validation, avoided RRT-defined stage 3 AKI, and handled actual RRT as an intercurrent event when it preceded any recorded objective criterion. Model evaluation included discrimination, AUPRC, calibration, Brier score, decision-curve analysis, threshold-specific metrics, bootstrap uncertainty, and

SHAP interpretation, which is more informative than reporting AUROC alone for an infrequent and clinically consequential endpoint.

Several limitations should be considered. First, the study was retrospective and based on structured EHR data; unstructured bedside assessments, clinician intent, nuanced volume-status evaluation, vascular-access considerations, and goals-of-care discussions were unavailable. Second, objective RRT-NEED was an operational endpoint rather than an adjudicated clinical diagnosis, and misclassification is possible when laboratory, blood gas, fluid-balance, or oxygenation data were incompletely captured. Third, patients who received actual RRT during the outcome window before any recorded objective criterion were excluded because RRT can modify subsequent potassium, acid-base status, blood urea nitrogen, fluid balance, and oxygenation. However, this may remove a clinically important subgroup and may contribute to the low event rate. Fourth, death without recorded objective RRT-NEED was not counted as an endpoint component, preserving outcome specificity but potentially underestimating severe clinical deterioration. Fifth, missingness and informative measurement patterns, particularly for blood gas variables in eICU-CRD, may have contributed to residual bias and external calibration drift. Sixth, some selected predictors were correlated, and collinearity may distribute importance across related variables when interpreting SHAP values. Finally, external validation showed overestimation of absolute risk, indicating that local recalibration, prospective validation, and evaluation of clinical workflow impact are required before the model can be used for bedside decision support.

In conclusion, we developed and externally validated a Gradient Boosting model using routinely collected ICU variables to predict objective indications for RRT within 24–72 h after stage 3 AKI. The model showed good internal performance and moderate external performance, with evidence of external overprediction. It may support short-term risk stratification and structured reassessment after severe AKI, but it requires local recalibration and prospective validation and should not be used as a stand-alone trigger for RRT initiation.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.L. Supplementary Information accompanies this manuscript.

Funding

No funding was received for this study.

Author Contribution

C.C. and W.L. conceived the study. C.C. performed data extraction and statistical analysis. C.C., P.T., F.G., and X.C. contributed to study design, data processing, model development, interpretation of results, and manuscript preparation. W.L. supervised the study and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

None.

Data Availability

The data used in this study were obtained from two publicly available critical care databases: MIMIC-IV and the eICU Collaborative Research Database. Both databases are available to credentialed users through PhysioNet after completion of the required training and data-use agreements. The authors do not redistribute raw patient-level data. Derived aggregate tables supporting the findings of this study, including cohort flow, endpoint components, model performance, calibration, decision-curve analysis, and cohort-level case-mix comparisons, are provided in the Supplementary Information where permitted by the database-use agreements. Derived patient-level analytic files remain subject to the same data-use restrictions as the source databases.

Code availability

The SQL extraction scripts and Python model-development code can be made available from the corresponding author upon reasonable request, subject to the data-use agreements governing MIMIC-IV and eICU-CRD and institutional approval where applicable.

References

1. Bellomo, R., Kellum, J. A. & Ronco, C. Acute kidney injury. *Lancet* **380**, 756–766 (2012).
2. Hoste, E. A. J. et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. *Intensive Care Med.* **41**, 1411–1423 (2015).
3. Kellum, J. A., Lameire, N. & for the KDIGO AKI Guideline Work Group. Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary (Part 1). *Crit. Care*. **17**, 204 (2013).

4. Gaudry, S. et al. Initiation Strategies for Renal-Replacement Therapy in the Intensive Care Unit. *N Engl. J. Med.* **375**, 122–133 (2016).
5. Gaudry, S. et al. Comparison of two delayed strategies for renal replacement therapy initiation for severe acute kidney injury (AKIKI 2): a multicentre, open-label, randomised, controlled trial. *Lancet* **397**, 1293–1300 (2021).
6. Barbar, S. D. et al. Timing of Renal-Replacement Therapy in Patients with Acute Kidney Injury and Sepsis. *N Engl. J. Med.* **379**, 1431–1442 (2018).
7. The STARRT-AKI Investigators. Timing of Initiation of Renal-Replacement Therapy in Acute Kidney Injury. *N Engl. J. Med.* **383**, 240–251 (2020).
8. Liu, C. et al. Development and validation of a model for the early prediction of the RRT requirement in patients with rhabdomyolysis. *Am. J. Emerg. Med.* **46**, 38–44 (2021).
9. Song, B. et al. Development and validation of a predictive model for continuous renal replacement therapy in sepsis patients using the MIMIC-IV database. *Sci. Rep.* **15**, 21559 (2025).
10. Mahmoud, M., Bader, M., McNicholas, J., Esmeli, R. & Optimizing, I. C. U. Care: Machine Learning and PCA for Early Prediction of Renal Replacement Therapy Requirement. *Stud. Health Technol. Inf.* **316**, 879–883 (2024).
11. Nierhaus, A. et al. Predicting the requirement for renal replacement therapy in intensive care patients with sepsis. *Crit. Care*. **22**, 201 (2018).
12. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*. **10**, 1 (2023).
13. Pollard, T. J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data*. **5**, 180178 (2018).
14. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
15. Moons, K. G. M. et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025).
16. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
17. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, 1189–1232 (2001).
18. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. **10**, e0118432 (2015).
19. Van Calster, B. et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).
20. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26**, 565–574 (2006).

21. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777Curran Associates Inc., Red Hook, NY, USA, (2017).
22. Wei, C., Zhang, L., Feng, Y., Ma, A. & Kang, Y. Machine learning model for predicting acute kidney injury progression in critically ill patients. *BMC Med. Inf. Decis. Mak.* **22**, 17 (2022).
23. Bhatraju, P. K. et al. A Prediction Model for Severe AKI in Critically Ill Adults That Incorporates Clinical and Biomarker Data. *Clin. J. Am. Soc. Nephrol.* **14**, 506–514 (2019).
24. Klein, S. J. et al. Biomarkers for prediction of renal replacement therapy in acute kidney injury: a systematic review and meta-analysis. *Intensive Care Med.* **44**, 323–336 (2018).

Tables

Table 1.

Baseline characteristics of patients with and without objective RRT-NEED.

Variables	MIMIC-IV cohort		SMD	eICU-CRD cohort		SMD
	No objective RRT-NEED (n = 1,941)	Objective RRT-NEED (n = 185)		No objective RRT-NEED (n = 1,795)	Objective RRT-NEED (n = 121)	
Demographics						
Age, years	65 (53, 76)	64 (53, 71)	0.153	64 (53, 74)	66 (55, 77)	0.153
Male sex, n (%)	1005 (51.8%)	117 (63.2%)	0.234	1048 (58.4%)	81 (66.9%)	0.178
Documented body weight, kg	104 (79.60, 158.40)	95 (76.10, 113)	0.369	88 (71.70, 108.40)	87 (67.80, 104.33)	0.054
AKI phenotype						
Creatinine-defined stage 3 AKI, n (%)	819 (42.2%)	147 (79.5%)	0.826	1130 (63.0%)	106 (87.6%)	0.596
Urine-output-defined stage 3 AKI, n (%)	1466 (75.5%)	117 (63.2%)	0.269	858 (47.8%)	39 (32.2%)	0.322
Time from ICU admission to stage 3 AKI, h	28.50 (19.22, 57.59)	25.63 (5.97, 59.77)	0.047	26 (15.87, 64.63)	34 (15.47, 83.40)	0.049
Severity and organ support						
SOFA score	8 (5, 11)	11 (7.75, 13)	0.717	5 (4, 8)	7 (5, 10)	0.562
Mechanical ventilation, n (%)	1275 (65.7%)	121 (65.4%)	0.006	1098 (61.2%)	75 (62.0%)	0.017
Vasoactive drug use, n (%)	801 (41.3%)	99 (53.5%)	0.247	306 (17.0%)	21 (17.4%)	0.008
Suspected sepsis during predictor window, n (%)	1449 (74.7%)	151 (81.6%)	0.169	843 (47.0%)	63 (52.1%)	0.102
Vital signs						
Maximum heart rate, beats/min	104 (90, 119)	107 (92, 123.50)	0.209	111 (96, 125)	108 (93.50, 122)	0.084
Renal and fluid variables						
Urine output during predictor window, mL	961 (535, 1605)	455 (178, 1135)	0.347	615 (72.50, 1282.50)	195 (0, 788)	0.367
Cumulative fluid balance during	1028.74 (0, 2440.08)	993.75 (0, 2700.90)	0.082	-70 (-930, 1053.75)	0 (-350, 1644)	0.266

Variables	MIMIC-IV cohort		SMD	eICU-CRD cohort		SMD
	No objective RRT-NEED (n = 1,941)	Objective RRT-NEED (n = 185)		No objective RRT-NEED (n = 1,795)	Objective RRT-NEED (n = 121)	
predictor window, mL						
Maximum serum creatinine, mg/dL	1.40 (0.80, 2.60)	3.50 (2.30, 4.60)	0.985	3.02 (1.50, 4.60)	4.23 (3.31, 4.95)	0.482
Maximum blood urea nitrogen, mg/dL	31 (19, 49)	69.50 (39.75, 96.25)	1.141	47 (29, 67)	80 (55.50, 94.50)	0.967
Electrolyte and acid-base variables						
Maximum serum potassium, mmol/L	4.30 (3.90, 4.70)	4.75 (4.28, 5.20)	0.708	4.30 (3.90, 4.80)	4.60 (4, 5.12)	0.347
Minimum serum bicarbonate, mmol/L	22 (19, 25)	18 (15, 21)	0.730	20 (17, 24)	19 (16, 23)	0.193
Minimum arterial pH	7.35 (7.29, 7.40)	7.31 (7.24, 7.36)	0.467	7.34 (7.27, 7.40)	7.31 (7.25, 7.37)	0.105
Hematology						
Minimum platelet count	158 (99, 232.25)	114 (65.50, 195)	0.267	154 (99, 223)	115 (74, 174)	0.346
Maximum white blood cell count	12.40 (8.90, 16.70)	14.20 (9.22, 20.30)	0.268	13.40 (9.60, 18.90)	14.99 (10.03, 20.27)	0.174

Values are median (interquartile range) or number (percentage). Absolute SMD values are reported to describe between-group differences. SMD, standardized mean difference; AKI, acute kidney injury; eICU, eICU Collaborative Research Database; ICU, intensive care unit; RRT, renal replacement therapy; SOFA, Sequential Organ Failure Assessment.

Table 2. Performance of selected models.

Model	Dataset	AUROC (95% CI)	AUPRC (95% CI)	Brier (95% CI)	Mean predicted risk	Observed-to-expected ratio
Gradient Boosting	MIMIC-IV internal test	0.848 (0.790-0.896)	0.459 (0.349-0.599)	0.062 (0.050-0.078)	0.087	1.01
Gradient Boosting	eICU-CRD external validation	0.764 (0.717-0.804)	0.181 (0.141-0.258)	0.065 (0.057-0.073)	0.125	0.51
L2-penalized logistic regression	MIMIC-IV internal test	0.812 (0.740-0.873)	0.442 (0.295-0.596)	0.064 (0.052-0.079)	0.087	1.0
L2-penalized logistic regression	eICU-CRD external validation	0.770 (0.729-0.815)	0.197 (0.153-0.257)	0.060 (0.054-0.067)	0.124	0.51
Extremely randomized trees	MIMIC-IV internal test	0.834 (0.779-0.900)	0.395 (0.270-0.549)	0.065 (0.051-0.079)	0.088	0.99
Extremely randomized trees	eICU-CRD external validation	0.775 (0.731-0.814)	0.228 (0.166-0.316)	0.062 (0.056-0.068)	0.131	0.48

AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; CI, confidence interval.

Table 3. Threshold-specific performance of the primary Gradient Boosting model.

Dataset	Threshold	Sensitivity	Specificity	PPV	NPV	Patients above threshold	Patients below threshold
MIMIC-IV internal test	5%	0.893	0.615	0.182	0.984	274	364
MIMIC-IV internal test	10%	0.75	0.814	0.28	0.971	150	488
MIMIC-IV internal test	15%	0.536	0.912	0.37	0.953	81	557
MIMIC-IV internal test	20%	0.446	0.947	0.446	0.947	56	582
MIMIC-IV internal test	30%	0.339	0.969	0.514	0.938	37	601
eICU-CRD external validation	5%	0.909	0.342	0.085	0.982	1292	624
eICU-CRD external validation	10%	0.777	0.644	0.128	0.977	733	1183
eICU-CRD external validation	15%	0.595	0.795	0.164	0.967	440	1476
eICU-CRD external validation	20%	0.446	0.872	0.19	0.959	284	1632
eICU-CRD external validation	30%	0.331	0.93	0.242	0.954	165	1751

PPV, positive predictive value; NPV, negative predictive value.

Figures

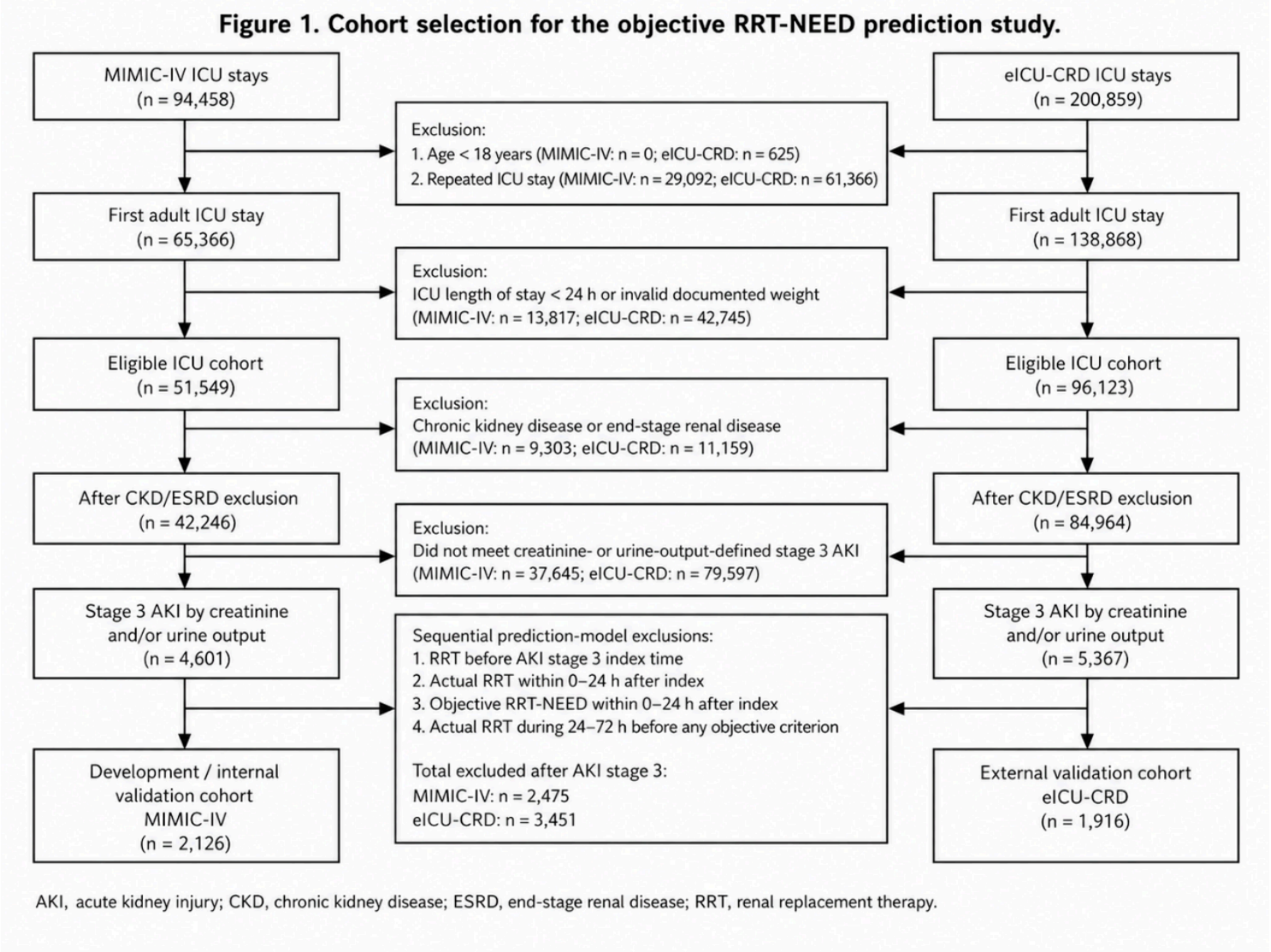


Figure 1

Cohort selection for the objective RRT-NEED prediction study. Adult ICU stays from MIMIC-IV and eICU-CRD were screened according to prespecified eligibility and exclusion criteria. The final MIMIC-IV cohort was used for model development and internal validation, and the final eICU-CRD cohort was used for external validation. AKI, acute kidney injury; CKD, chronic kidney disease; ESRD, end-stage renal disease; RRT, renal replacement therapy.

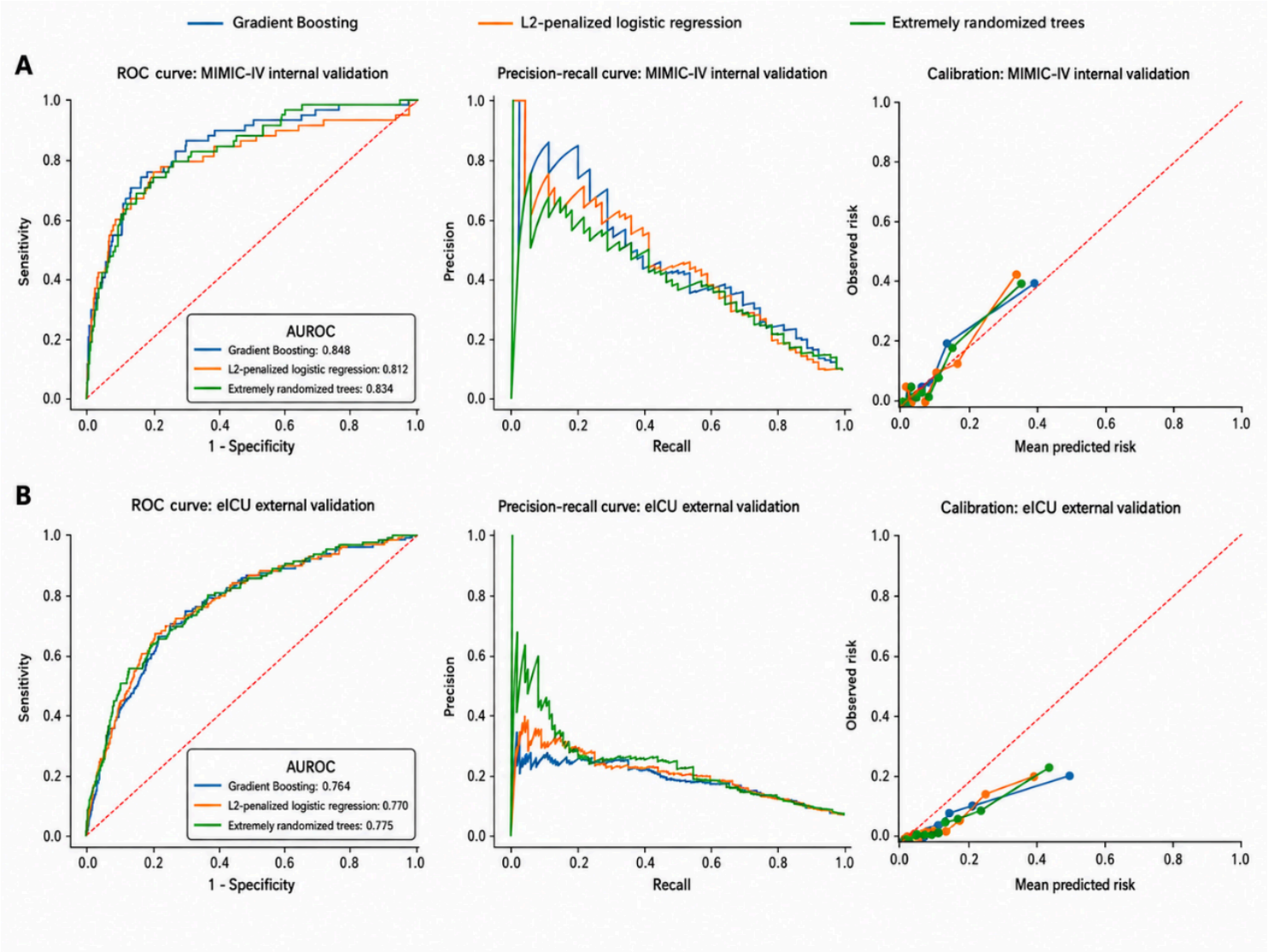


Figure 2

Model discrimination and calibration. (A) Receiver operating characteristic curves, precision-recall curves, and calibration plots in the MIMIC-IV internal validation cohort. (B) Corresponding plots in the eICU-CRD external validation cohort. The dashed diagonal lines indicate no-discrimination performance in ROC plots and perfect calibration in calibration plots.

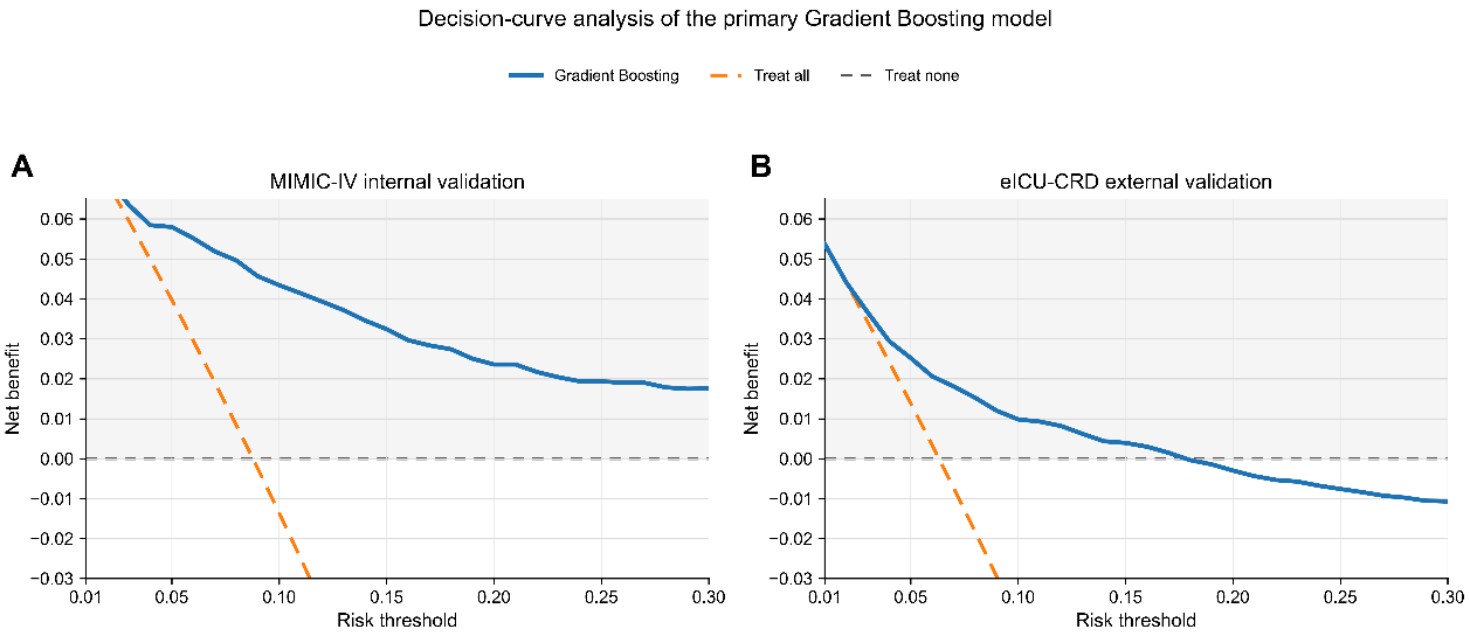


Figure 3

Decision-curve analysis of the primary Gradient Boosting model. Net benefit was evaluated across risk thresholds from 0.01 to 0.30 in (A) the MIMIC-IV internal validation cohort and (B) the eICU-CRD external validation cohort. Treat-all and treat-none strategies are shown as reference curves.

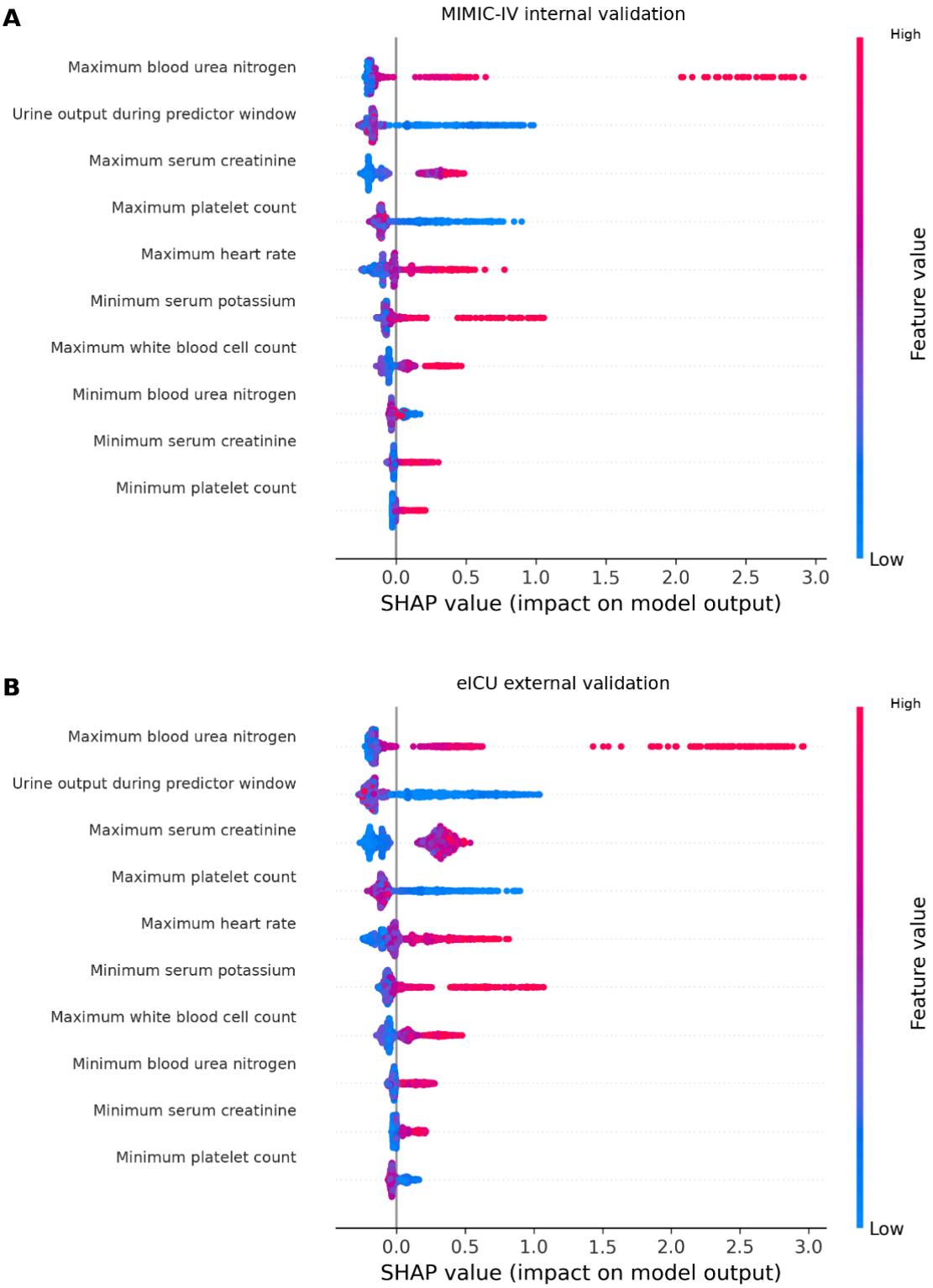


Figure 4

SHAP-based interpretation of the primary Gradient Boosting model. (A) SHAP summary plot in the MIMIC-IV internal validation cohort. (B) SHAP summary plot in the eICU-CRD external validation cohort. Predictors are ranked by their overall contribution to model output. SHAP, Shapley additive explanations.

Supplementary Files

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

- SupplementaryInformation.docx
- supplementarytables.zip