BMC Medical Informatics
and Decision Making

RESEARCH Open Access

An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records

Hongnian Wang[1,2], Mingyang Zhang[3], Liyi Mai[4], Xin Li[5*], Abdelouahab Bellou[4,5,6,7*] and Lijuan Wu[4,8*]

Abstract

Background Identifying key variables is essential for developing clinical outcome prediction models based on high-dimensional electronic medical records (EMR). However, despite the abundance of feature selection (FS) methods available, challenges remain in choosing the most appropriate method, deciding how many top-ranked variables to include, and ensuring these selections are meaningful from a medical perspective.

Methods We developed a practical multi-step feature selection (FS) framework that integrates data-driven statistical inference with a knowledge verification strategy. This framework was validated using two distinct EMR datasets targeting different clinical outcomes. The first cohort, sourced from the Medical Information Mart for Intensive Care III (MIMIC-III), focused on predicting acute kidney injury (AKI) in ICU patients. The second cohort, drawn from the MIMIC-IV Emergency Department (MIMIC-IV-ED), aimed to estimate in-hospital mortality (IHM) for patients transferred from the ED to the ICU. We employed various machine learning (ML) methods and conducted a comparative analysis considering accuracy, stability, similarity, and interpretability. The effectiveness of our FS framework was evaluated using discrimination and calibration metrics, with SHAP applied to enhance the interpretability of model decisions.

Results Cohort 1 comprised 48,780 ICU encounters, of which 8,883 (18.21%) developed AKI. Cohort 2 included 29,197 transfers from the ED to the ICU, with 3,219 (11.03%) resulting in IHM. Among the ten ML methods evaluated, the tree-based ensemble method achieved the highest accuracy. As the number of top-ranking features increased, the models' accuracy began to stabilize, while feature subset stability (considering sample variations) and inter-method feature similarity reached optimal levels, confirming the validity of the FS framework. The integration of interpretative methods and expert knowledge in the final step further improved feature interpretability. The FS framework effectively reduced the number of features (e.g., from 380 to 35 for Cohort 1, and from 273 to 54 for Cohort 2) without significantly affecting prediction performance (Delong test, $p > 0.05$).

*Correspondence:
Xin Li
lixin@gdph.org.cn
Abdelouahab Bellou
abellou402@gmail.com
Lijuan Wu
wulj1989@163.com

Full list of author information is available at the end of the article

Conclusion The multi-step FS method developed in this study successfully reduces the dimensionality of features in EMR while preserving the accuracy of clinical outcome prediction. Furthermore, it improves the interpretability of risk factors by incorporating expert knowledge validation.

Keywords Electronic medical records, Feature selection, Machine learning, Clinical outcomes, Risk prediction, Interpretability

Background

The development of clinical outcome prediction models from electronic medical records (EMR) represents a significant advancement in healthcare, providing valuable insights into patient management and treatment strategies. This is especially crucial for conditions such as acute kidney injury (AKI) and in-hospital mortality (IHM), where timely and accurate predictions can profoundly impact patient outcomes. Leveraging large-scale EMR data for these predictions enables early detection and intervention, which are essential for enhancing patient care in acute medical settings.

AKI is a serious medical condition characterized by a rapid decline in kidney function, leading to elevated rates of morbidity and mortality [1, 2]. The Kidney Disease Improving Global Outcomes (KDIGO) guidelines define AKI by increases in serum creatinine (SCr) levels or decreases in urine output, both of which are late indicators of the condition [3]. The data-driven approach that utilizes extensive EMR datasets offers a unique analytical opportunity to facilitate the early detection of AKI [4, 5], thereby allowing for timely interventions from specialists or pharmacists aimed at improving patient outcomes [6].

IHM among patients transferred from the emergency department (ED) to the intensive care unit (ICU) is another critical outcome that stands to benefit from predictive modeling [7]. These patients often present in severe condition, and accurately predicting IHM can aid clinicians in making informed decisions regarding patient care, resource allocation, and treatment strategies [8]. Early identification of high-risk patients enables prompt interventions that may lower mortality rates and enhance overall healthcare efficiency [9].

Despite these potential benefits, the high dimensionality of EMR data presents significant challenges in identifying the most informative and relevant features for accurate prediction [10]. Including irrelevant or redundant variables can lead to overfitting, increased computational complexity, and decreased predictive performance [11]. Feature selection (FS) emerges as an effective approach to identify the most pertinent features from high-dimensional EMR datasets. FS has been widely applied in disease risk prediction models to reduce dimensionality, enhance model performance, and improve interpretability [12]. Additionally, FS can provide valuable insights into the underlying mechanisms and risk factors associated with diseases, thereby assisting clinicians in diagnosis, treatment, and prevention strategies [13–16].

However, several common issues persist in applying FS methods to high-dimensional EMR data:

a) **Which FS methods should be prioritized?** The vast array of feature selection methods often complicates the process of making an optimal choice. These methods can be broadly categorized into three types [12]: simple filter methods based on statistical measures and correlation (e.g., Chi-square test), wrapper methods that assess feature contributions by training predictive models (e.g., stepwise regression), and embedded methods that integrate feature selection with model training processes (e.g., random forest).

b) **How many top-ranked variables should be selected?** While many FS methods provide insights into variable importance, previous research has highlighted significant discrepancies in the risk factor importance derived from different FS methods due to varying criteria [13]. However, most existing studies tend to overlook the stability of selected features influenced by sample variations and the similarity of features identified by multiple methods [17, 18].

c) **Do the selected factors align with medical interpretation?** Factors identified by FS methods indicate correlations with clinical outcomes but do not necessarily imply causation [16, 19]. Some studies prioritize prediction accuracy over ensuring that the selected features align with clinical interpretation [20]. Given the impracticality of manually filtering features in high-dimensional EMR data, some research has focused on developing transparent risk-scoring models using a limited number of known risk factors, often tailored to specific situations [21].

To address these challenges, we introduce a practical multi-step FS framework that combines statistical inference with expert knowledge validation. This innovative approach is designed to prioritize the most effective FS methods, determine the optimal number of top-ranked variables, and ensure that the selected factors are medically interpretable. By focusing on accuracy, stability, similarity, and interpretability, key attributes that enhance the reliability and generalizability of clinical outcome prediction models—our primary objectives are twofold: (1) to identify cost-effective and significant predictors while maintaining high predictive performance; and (2) to select EMR features that offer reasonable medical interpretability, enabling a better understanding of the underlying factors that contribute to clinical outcomes. The effectiveness of this multi-step FS framework was validated using two de-identified EMR datasets corresponding to distinct clinical outcomes. By providing a more stable and reliable feature selection approach, this method has the potential to enhance the credibility and interpretability of predictive models in clinical applications.

Methods

Ethics

The clinical data repository utilized in this study includes the Medical Information Mart for Intensive Care III (MIMIC-III, version 1.4) [22] and the MIMIC-IV Emergency Department (MIMIC-IV-ED, version 2.0) [23]. We obtained permission to extract data from both datasets (certification number: 34034170).

Study participants

Cohort 1 comprised 46,520 ICU patients across 61,051 encounters sourced from the MIMIC-III database, covering the timeframe from 2001 to 2012. The objective was to predict AKI within 48 h of ICU admission, with AKI defined according to the KDIGO SCr criteria [24] (see Supplementary Table S1). Baseline SCr was determined using the last measurement within two days preceding admission, if available, otherwise, the first measurement post-admission was utilized. Exclusion criteria included patients younger than 18 years, those with an initial SCr ≥ 4 mg/dL, and individuals with a history of end-stage renal disease or chronic dialysis [25]. Cohort 2 consisted of patients from the MIMIC-IV-ED database, which documented 447,712 ED visits from 2011 to 2019. The focus here was on predicting IHM for patients transitioned from the ED to the ICU. Exclusions were applied to patients under 18 years of age; those lacking critical emergency records; and individuals with illogical temporal recording sequences, and patients who were not transferred to the ICU.

Data extraction and processing

Table 1 provides a detailed overview of patient characteristics for both cohorts. Cohort 1 includes demographic information (age, gender, race), details of admission (e.g.,

Table 1 Patient characteristics used in both cohorts

Feature Category	Number of Variables	Details/Examples
The 380 features from Cohort 1		
Demographics	3	Age, Gender, Race;
Admission information	3	Admission type, Admission location, First care unit;
Vital signs	9	Heart rate, Systolic blood pressure, Diastolic blood pressure, Mean arterial pressure, Respiratory rate, Temperature, SpO2, Height, weight;
Lab tests	19	Anion gap, Albumin test, Band neutrophils test, Bicarbonate test, Bilirubin test, Creatinine test, Chloride test, Glucose test, Hematocrit (red blood cells) test, Hemoglobin test, Lactate test, Platelet count blood test, Potassium blood test, Partial thromboplastin time (PTT), International normalized ratio (INR), Prothrombin time (PT), Sodium blood test, Blood urea nitrogen (BUN), White blood count (WBC);
Intervention	346	Information from Tables INPUTEVENTS and PROCEDUREVENTS, for example, medications, vasopressor, ventilation, etc.
The 273 features from Cohort 2		
Demographics	3	Age, Gender, Race;
Admission information	2	Insurance, Arrival transport;
Triage observations	7	Temperature, Heart rate, Respiratory Rate, SpO2, Systolic blood pressure, Diastolic blood pressure, Acuity Level;
Vital signs	6	Temperature, Heart rate, Respiratory Rate, SpO2, Systolic blood pressure, Diastolic blood pressure;
Lab tests	45	Anion gap, Albumin test, Bicarbonate test, Bilirubin test, Creatinine test, Chloride test, Glucose test etc.;
Medications	160	Grouped using the hierarchical ontology of the Enhanced Therapeutic Class (ETC), e.g., Analgesic Opioid Agonists, Antiemetics, Diuretics, etc.;
Diagnosis*	48	Defined by the International Statistical Classification of Diseases and Related Health Problems, 9th and 10th Revision (ICD-9 and ICD-10), e.g., Gastrointestinal Hemorrhage, Pneumonia, Shortness of Breath, etc.;
Time-Related information	2	EDLOS (the time from arrival to departure from the ED), EDBT (the time spent in the ED waiting for an inpatient bed after admission decision);

Notes: In Cohort 1, some tables that do not have accurate timestamps, such as DIAGNOSES_ICD, CPTEVENTS, DRGCODES, PROCEDURES_ICD, CAREGIVERS, and PRESCRIPTIONS, were not used in the analysis. Additionally, NOTEEVENTS containing unstructured data was also excluded

*Diagnosis codes represent only the first three digits of ICD codes, as the dataset is specific to ED patients admission type and location), vital signs, laboratory tests, and interventions derived from the "*INPUTEVENTS*" and "*PROCEDUREEVENTS*" tables. The most recent values of vital signs and laboratory tests recorded prior to ICU admission were employed. In Cohort 2, patient characteristics encompass demographics, specifics of admission (such as insurance and arrival mode), triage observations, vital signs, laboratory tests, medications organized by Enhanced Therapeutic Class (ETC), and diagnoses coded using ICD-9 and ICD-10 classifications. Time-related information, including emergency department length of stay (EDLOS) and boarding time (EDBT), were also extracted and computed. For vital signs and laboratory values, the most recent measurements before the patient's transfer to the ICU were used. During the data processing phase, essential procedures such as unit conversion, outlier handling, and aggregation of semantically similar features were conducted to minimize missing data and enhance the overall robustness of the analysis [22]. The median value was utilized to impute missing data when training traditional models (KNN, NB, LR, and DTC), whereas no imputation was applied for tree-based ensemble models (RF, XGBoost, LightGBM, and CatBoost) because tree-based methods inherently handle missing values by splitting on available data, allowing them to make use of the information from other features without the need for explicit imputation.

Multi-step feature selection framework

To identify an optimal subset of features that prioritizes accuracy, stability, and interpretability, we developed a multi-step FS framework that integrates data-driven statistical inference through expert knowledge verification (see Fig. 1). The process is detailed below:

Step 1: Univariate feature selection. In this initial step, we independently assess the correlation between each feature x_i $(1 \leq \ i \leq \ N)$ and the target variable y . Features demonstrating a significant statistical correlation with the target variable (e.g., $p < 0.05$) are selected based on specific evaluation metrics, such as t-test, Chi-square test, and Wilcoxon rank-sum test. This helps eliminate many redundant variables.

Step 2: Multivariate feature selection. This step focuses on selecting the most predictive subset of features by capturing potential interactions and dependencies among them. To ensure consistent selection results, we conducted additional analyses to evaluate the stability of selection outcomes across sample variations [26], and the similarity of feature importance rankings generated by different FS methods [27].

First, we identified suitable embedded FS methods that effectively combine feature selection with model training processes. Selecting an appropriate classifier model with strong discriminatory power for the current data and research problem is essential. Thus, we compared ten ML methods, including Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees Classifier (DTC), Naive Bayes (NB), Multi-Layer Perceptron (MLP), Random Forest (RF), XGBoost [28], Catboost [29], and LightGBM [30].

Subsequently, these ML models, along with post-hoc interpretation techniques such as SHAP [31], were utilized to derive inherent feature importance. A forward FS strategy [32] was employed to identify the top K ranked features $\left\{X_{[1:top1]}, X_{[1:top2]}, \cdots, X_{[1:topK]}\right\}$ (where $1 \leq \ K \leq \ N$). The value of K was determined by observing the stability of the predicted performance curve as the number of top-ranking features increases.

We then characterized the stability and similarity trend of the selected top K features. The Similarity Index [27] can be defined as

$$\text{Similarity Index}_{M_i,M_j} = \frac{1}{h}\sum_{n=1}^{h}\frac{\left|S_n^{(i)} \cap S_n^{(j)}\right|}{K}, \quad (1)$$

where h denotes the number of cross-validation folds (e.g., h=10 for ten-fold cross-validation), $S_n^{(i)}$ and $S_n^{(j)}$ are the corresponding subsets of top K ranked features obtained by the different methods M_i and M_j in the n -th fold, respectively. The average similarity index for the method M_i across all other methods is calculated as: $\frac{1}{M-1}\sum_{j=1,j\neq i}^{M}\text{Similarity Index}_{M_i,M_j}$. This formula averages the similarity indices between method M_i and each other method M_j (where j encompasses all methods except i).

The Stability Index is computed as follows:

$$\text{Stability Index} = \frac{2}{h(h-1)}\sum_{(i=1)}^{(h-1)}\sum_{(j=i+1)}^{h} I_C(S_i, S_j), \quad (2)$$

where $I_C\left(S_i, S_j\right) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$ stands for the consistency index between two feature subsets [26], $|\cdot|$ denotes cardinality, '$\cap$'denotes intersection, and '$\cup$'denotes union of sets.

Once the measurement criteria for accuracy (as represented by AUROC), similarity, and stability, were established, we determined the optimal K values for each criterion by applying the elbow method to their respective curves, identifying points at which performance improvements plateau. We derived K_{accuracy} from the AUROC curve, $K_{\text{similarity}}$ from the similarity index curve, and $K_{\text{stability}}$ from the stability index curve. Ultimately, $K_{\text{optimal}} = max\left(K_{\text{accuracy}}, K_{\text{similarity}}, K_{\text{stability}}\right)$. This selection ensures that all three metrics reach stable and optimal levels, facilitating a robust and reliable feature selection process.

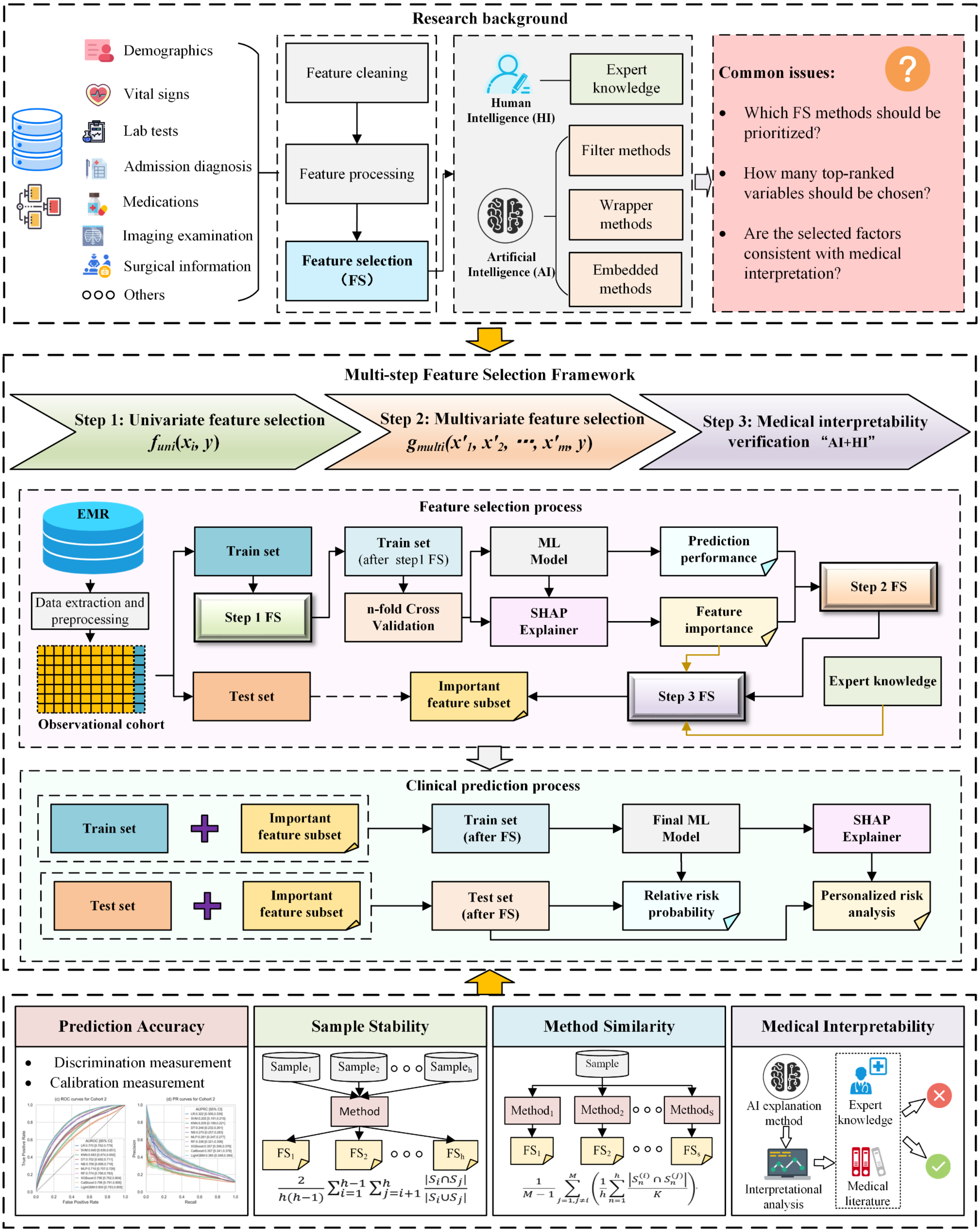

Fig. 1 Overview of the multi-step FS framework for clinical risk assessment

Finally, the data-driven method aids not only in the interpretability analysis of risk features (e.g., significance or weight), but also filters out a feasible subset of features for expert knowledge validation. This systematic selection process enables the identification of a reliable and stable predictor subset (i.e., top K features) that significantly impacts the target variable.

Table 2 Demographics and admission information of two cohorts

Characteristics	Negative	Positive	P values
In Cohort 1, *n* (%)	**39,897 (81.79)**	**8,883 (18.21)**	
Age, year, median [Q1, Q3]	65 [52, 78]	69 [58, 79]	<0.001
Gender, male, n (%)	22,090(55.37)	5314(59.82)	< 0.001
Race, n (%)			< 0.001
Unknown	4126(10.34)	1147(12.91)	
White	29,161(73.09)	6357(71.56)	
Black	3232(8.1)	707(7.96)	
Hispanic	1365(3.42)	235(2.65)	
Asian	916(2.3)	207(2.33)	
Other	1097(2.75)	230(2.59)	
Type of hospital admission, n (%)			< 0.001
Elective	5473(13.72)	1726(19.43)	
Urgent	976(2.45)	269(3.03)	
Emergency	33,448(83.84)	6888(77.54)	
Type of ICU when first admitted, n (%)			< 0.001
Coronary Care Unit	5659(14.18)	1394(15.69)	
Cardiac Surgery Recovery Unit	6032(15.12)	2889(32.52)	
Medical Intensive Care Unit	15,855(39.74)	2693(30.32)	
Surgical Intensive Care Unit	6970(17.47)	1171(13.18)	
Trauma/Surgical Intensive Care Unit	5381(13.49)	736(8.29)	
In Cohort 2, n (%)	**25,978 (88.97)**	**3,219 (11.03)**	
Age, year, median [Q1, Q3]	67 [54, 79]	74 [62, 84]	< 0.001
Gender, male, n (%)	13,950 (53.7)	1715 (53.3)	0.664
Race, n (%)			< 0.001
Unknown	1671 (6.4)	378 (11.7)	
White	17,247 (66.4)	2100 (65.2)	
Black	3835 (14.8)	378 (11.7)	
Hispanic	1122 (4.3)	99 (3.1)	
Asian	916 (3.5)	137 (4.3)	
Other	1187 (4.6)	127 (3.9)	
Type of arrival transport, n (%)			< 0.001
Unknown	3106 (11.96)	550 (17.09)	
Ambulance	16,100 (61.98)	2057 (63.90)	
Walk-in	6397 (24.62)	550 (17.09)	
Helicopter	319 (1.23)	54 (1.68)	
Other	56 (0.22)	8 (0.25)	
Insurance, Medicare or Medicaid, n (%)	13,792 (53.09)	1940 (60.27)	< 0.001

Abbreviations: Q1, first quartile (25th percentile); Q3, third quartile (75th percentile); Patient characteristics were compared between the positive and negative cases in both the AKI and IHM cohorts using the t-test for normally distributed continuous variables, the Wilcoxon rank-sum test for non-parametric continuous variables, and the Chi-square test for categorical variables.

Step 3: Medical interpretability verification. Considering the complexity of medical features, ensuring medical interpretability is essential during the final screening of the important feature subsets identified through data-driven methods. The initial two steps provide interpretability analysis (e.g., SHAP value or feature weights) of significant risk features while filtering out a limited number of features that facilitate expert knowledge validation. In this study, we primarily verified the correlation and rationality of the selected features by referring existing medical literature and consulting two experienced clinicians for evaluation.

Statistical analysis

Population characteristics were presented as median [IQR] for continuous variables and as proportions for categorical variables. The t-test was employed for normally distributed continuous variables, while the Wilcoxon rank sum test was used for non-parametric continuous variables, and the Chi-square test assessed categorical variables. For model development, the dataset was divided into 80% training set and 20% test set using stratified random sampling; at the training stage, we used 10-fold cross-validation to evaluate the model's performance. Discrimination performance was evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Individual class metrics were assessed using precision, recall, and F1 score. The optimal classification threshold was determined using the Youden index [33]. Calibration was assessed via the Brier score and calibration chart, using observed and expected event rates per deciles as specified by the Hosmer-Lemeshow C statistic. Delong's test was applied to calculate the statistical significance of differences between two or more ROC curves [34]. Augmented Dickey-Fuller (ADF) test [35] was utilized to verify when the predictive performance of the top K feature model began to plateau. Two-tailed $p < 0.05$ denoted statistical significance for all comparisons. 95% confidence interval (CI) was calculated using bootstrapping ($n = 1{,}000$). Data extraction was conducted using PostgreSQL, while data processing and analysis were performed using Python 3.7 with open-source packages (e.g., 'lightgbm' and 'shap') and scikit-learn library.

Results

Characteristics of two study cohorts

Table 2 presents the demographics and admission information for the two cohorts after applying the exclusion criteria. Cohort 1 included 48,780 patient encounters, with an AKI incidence of 18.21%. Notably, AKI was more prevalent among male patients than female patients ($p < 0.001$) and predominantly affected older individuals, with a median age of 69 years [IQR: 58–79] compared to 65 years [IQR: 52–78] for non-AKI patients ($p < 0.001$). Additionally, patients initially admitted to the Cardiac Surgery Recovery Unit (CSRU) showed a significantly higher occurrence of AKI (32.52%) relative to those without AKI (15.12%; $p < 0.001$). Furthermore, the median creatinine levels recorded in the last test before the onset of AKI were significantly elevated in the AKI group, measuring 1.20 mg/dL [IQR: 0.90–1.80], in contrast to the non-AKI group, which had a median level of 1.00 mg/dL [IQR: 0.70–1.30] ($p < 0.001$) (see Supplementary Table S2).

Cohort 2 comprised 29,197 ICU admissions from the ED, with a mortality rate of 11.03%. Deceased patients were generally older than survivors, with a median age of 74 years [IQR: 62–84] compared to 67 years [IQR: 54–79] for those who survived ($p < 0.001$). Moreover, there were distinct differences in racial distribution and modes of arrival among deceased and surviving patients ($p < 0.001$), although no significant gender disparity was observed ($p = 0.664$). Insurance coverage by Medicare or Medicaid was more prevalent among deceased patients, with rates of 60.27% compared to 53.09% for survivors ($p < 0.001$). Additionally, deceased patients exhibited lower median albumin levels (3.2 g/dL [IQR: 2.6–3.7]) in contrast to survivors, who had median level of 3.7 g/dL [IQR: 3.2–4.1] ($p < 0.001$) (see Supplementary Table S2).

Comparison of accuracy, similarity and stability

Regarding the predictive performance of the 10 ML models (see Fig. 2), the tree-based ensemble models, including LightGBM, XGBoost, and CatBoost, exhibited superior results compared to other models, with LightGBM achieving the best overall performance. Specifically, the AUROC (95% CI) achieved by LR, SVM, KNN, DTC, NB, MLP, RF, XGBoost, CatBoost, and LightGBM in Cohort 1 were 0.671 (0.665–0.677), 0.560 (0.553–0.566), 0.640 (0.633–0.646), 0.683 (0.677–0.690), 0.645 (0.639–0.652), 0.677 (0.671–0.684), 0.722 (0.715–0.727), 0.735 (0.729–0.741), 0.730 (0.724–0.736), and 0.738 (0.732–0.744), respectively. In Cohort 2, the AUROC (95% CI) achieved by LR, SVM, KNN, DTC, NB, MLP, RF, XGBoost, CatBoost, and LightGBM were 0.770 (0.762–0.779), 0.640 (0.630–0.651), 0.683 (0.674–0.692), 0.702 (0.692–0.711), 0.706 (0.696–0.716), 0.716 (0.707–0.726), 0.774 (0.766–0.783), 0.796 (0.789–0.804), 0.798 (0.791–0.806), and 0.800 (0.793–0.808), respectively.

Figure 3 illustrates the variability in feature importance for outcomes prediction across the two cohorts, influenced by discrepancies in methods and samples. The figure clearly demonstrates substantial differences in the ranking of feature importance obtained through various FS methods. Notably, in Fig. 3(a), the XGBoost model identifies the "*Hematocrit test*" as the most critical feature, which contrasts significantly with findings from other methods. This variation raises the question of which feature importance ranking method is superior—whether the model's inherent ranking or the ranking derived from the combination with SHAP method. To address this, we selected three high-precision ML models (i.e., LightGBM, XGBoost and Catboost), and compared the results of six different FS approaches (i.e., the three models' own rankings and their rankings combined with the SHAP method).

Figure 4(a-d) depict the trend of feature importance and the prediction performance in the two cohorts as the number of top-ranking features increases. The graphs reveal an exponential decline in feature importance, while the performance curve stabilizes after reaching a certain value of K, indicating that adding more features beyond this point does not significantly improve prediction accuracy. Figure 4(e-h) demonstrate noticeable fluctuations in both similarity and stability of the important features identified by different FS methods when only a limited number are selected. However, once a certain threshold is surpassed, these features exhibit reasonable levels of stability and similarity. These findings suggest that selecting an optimal number of top-ranked features can help reduce discrepancies among different FS methods. Supplementary Figures S1 and S2 present the accuracy (AUROC), similarity, and stability of the six FS methods at various top K values, indicating where each metric reaches stability. Notably, accuracy stabilizes at a lower value of K compared to similarity and stability, which do not achieve optimal levels at this stage. Consequently, we selected the maximum K value at which all three metrics are stable for the FS step 2 (see Supplementary Table S3).

Effectiveness of the multi-step FS framework

In the first step of the FS process, univariate correlation tests (i.e., $p < 0.05$) were employed to screen features, resulting in the selection of 52 features from a total of 380 in Cohort 1 and 168 features from 273 in Cohort 2. In FS step 2, the performance curves (including accuracy, similarity and stability) began to stabilize at a certain number of top K features. This analysis utilized the LightGBM model combined with SHAP, leading to the selection of 35 features for Cohort 1 and 79 features for Cohort 2. In FS step 3, final selections were made based on literature, expert opinions, and the clarity of the features, yielding 35 predictors for Cohort 1 predictors (see Supplementary Table S5) and 54 predictors for Cohort 2 (see Supplementary Table S6). This approach enhances both the clarity and clinical rationale for each feature. Figure 5(g)

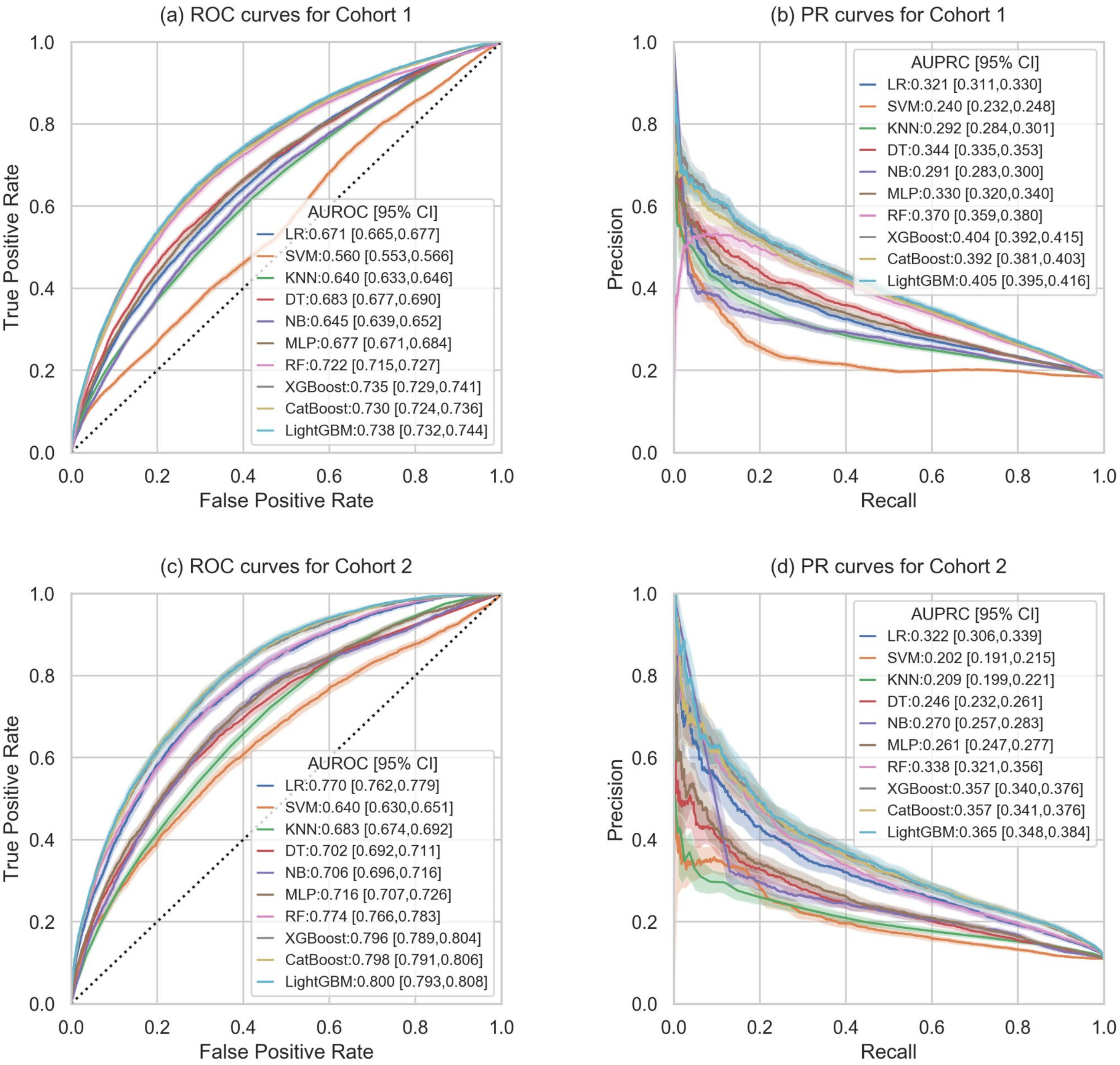

Fig. 2 The receiver operating characteristic (ROC) curves and precision-recall (PR) curves of 10 machine learning methods. (LR, Logistic Regression; SVM, Support Vector Machines; KNN, K-Nearest Neighbors; DTC, Decision Trees Classifier; NB, Naive Bayes; MLP, Multi-Layer Perceptron; RF, Random Forest; XGBoost, eXtreme Gradient Boosting; Catboost Categorical Boosting; LightGBM, Light Gradient Boosting Machine)

illustrates the variation in feature count at each stage of the selection process.

To assess the effectiveness of the multi-step FS framework, we compared the performance of four outcome prediction models developed at different FS stages: BFSM (before-FS model), AFSM-1 (after-FS-step-1 model), AFSM-2 (after-FS-step-2 model), and AFSM-3 (after-FS-step-3 model). Figure 5 presents the ROC curves, precision-recall curves, calibration plots, and decision curves for these models across both cohorts. The results indicate that the FS framework effectively reduced feature dimensionality while maintaining the predictive performance of the models. For instance, in Cohort 2, the AUROC values for BFSM and AFSM-3 were 0.800 (95% CI, 0.784–0.817) and 0.804 (95% CI, 0.789–0.821) ($p = 0.09$, Delong's test), respectively. Furthermore, the prediction performances of the four models in Cohort 1 did not show significant differences, averaging around 0.744 (95% CI, 0.731–0.756) (see Supplementary Figure S3). To evaluate the performance of individual class metrics in our predictive models, we calculated precision, recall, and F1 score, with the optimal classification threshold determined using the Youden Index (see Supplementary Table S4).

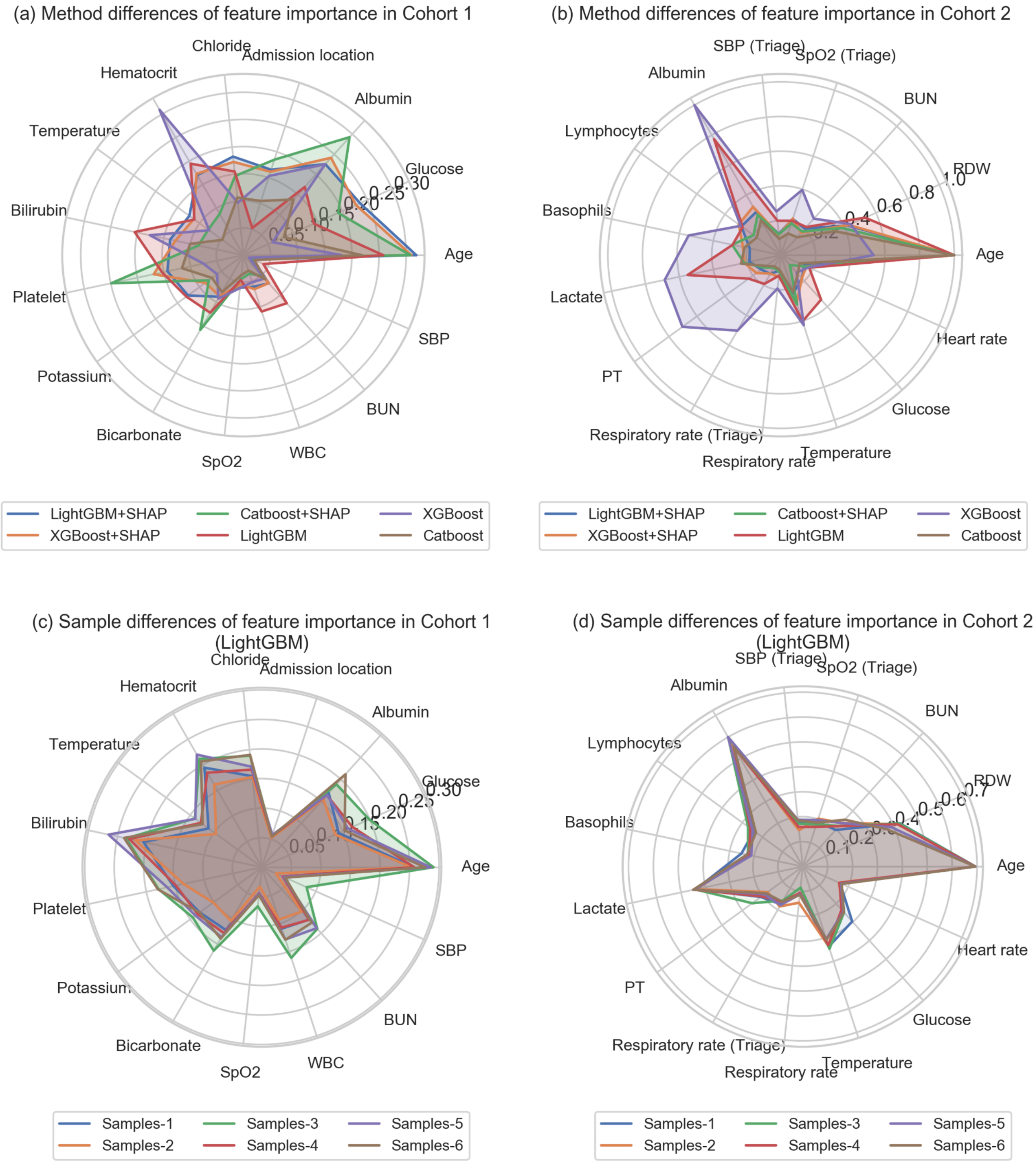

Fig. 3 Comparison of feature importance rankings across different feature selection methods (**a**-**b**) and sample variations (**c**-**d**). (The diversity of samples arises from the re-randomization of training and testing datasets, which introduces variability and helps assess the stability of feature importance rankings)

Interpretability of clinical outcome predictions

Supplementary Figure S4 highlights the top 15 predictors for both cohorts, while Fig. 6 shows the influence of the most significant predictors on the predicted outcomes. In Cohort 1, focused on predicting AKI, higher creatinine levels are associated with an increased risk of AKI (Fig. 6a). Admission to the cardiac surgery recovery unit (CSRU) is associated with the highest AKI risk (Fig. 6b), and advancing age significantly raises the AKI risk, especially for individuals over 60, underscoring the vulnerability of older adults (Fig. 6c). Interestingly, glucose levels demonstrate a protective effect within a

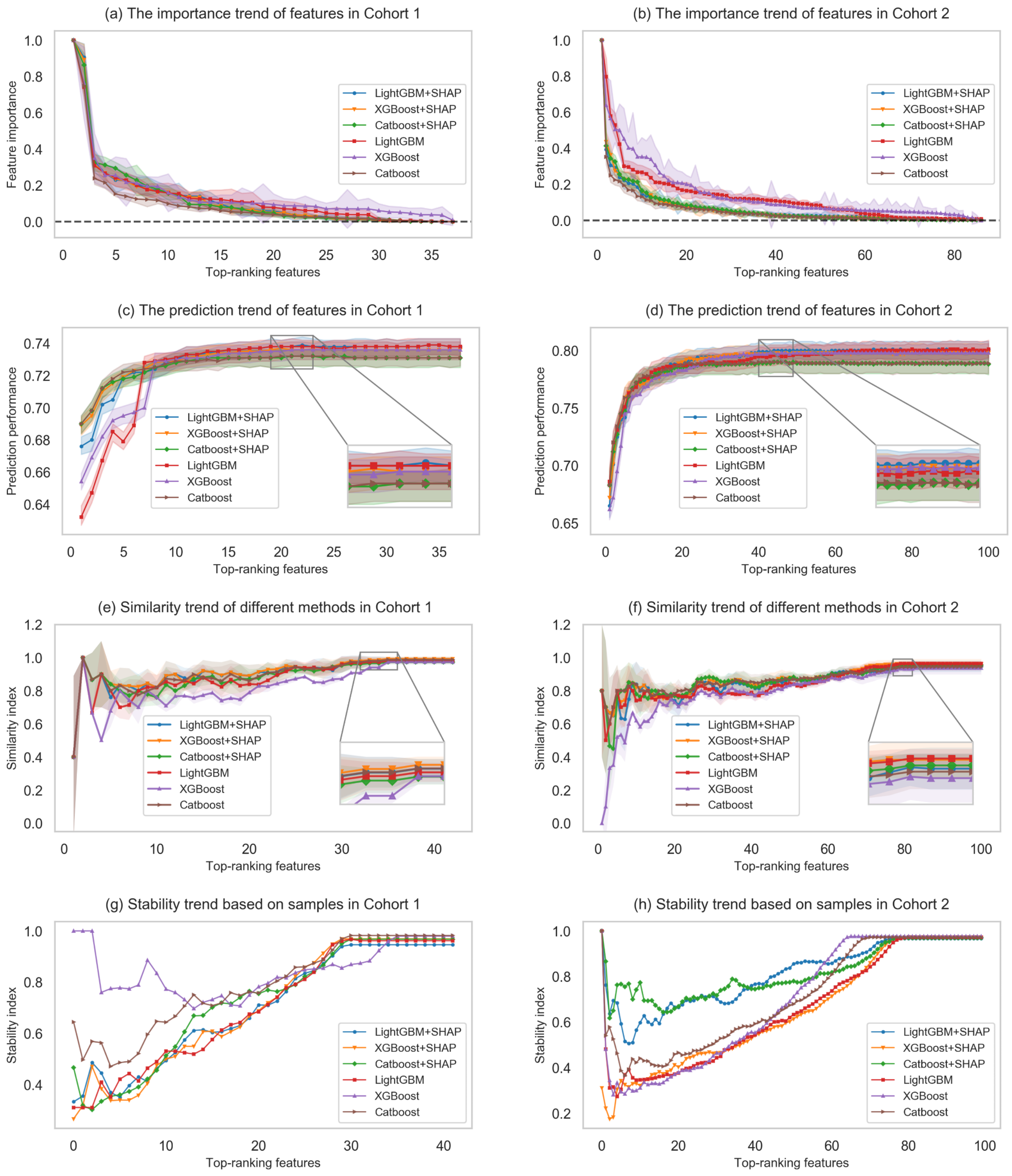

Fig. 4 Trends in feature importance (**a**-**b**), prediction performance (**c**-**d**), similarity among selected features (**e**-**f**), and stability of top-ranking features (**g**-**h**) across increasing numbers of features in the two cohorts

moderate range, but when they fall below or exceed this range, they are associated with an increased risk (Fig. 6d). In Cohort 2, which evaluates IHM, both increasing age and elevated red cell distribution width (RDW) show a strong correlation with higher mortality risks (Fig. 6e and f). Additionally, lower levels of lymphocytes and albumin are associated with increased mortality risks, highlighting their critical prognostic importance (Fig. 6g and h).

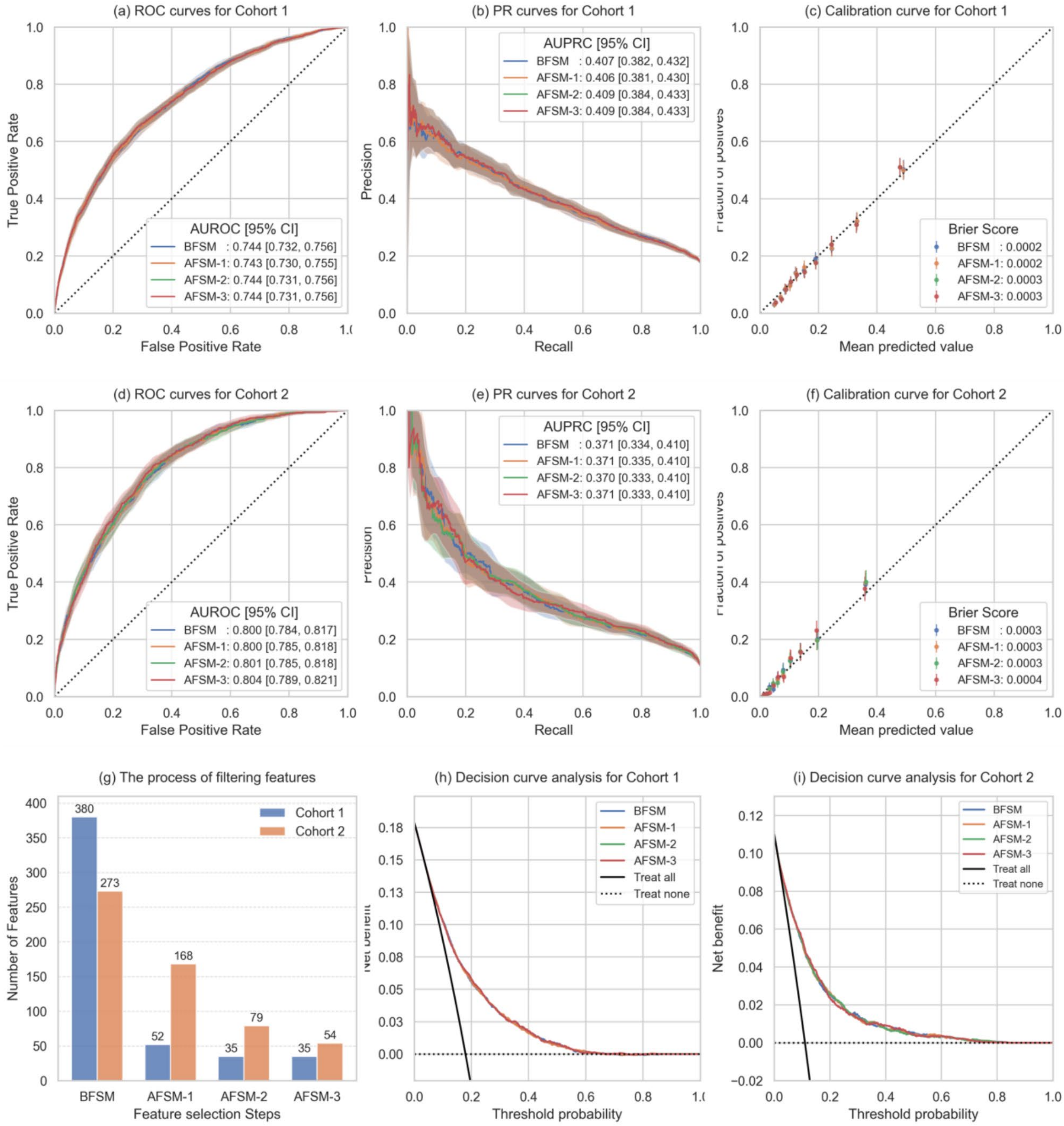

Fig. 5 Validation of the effectiveness of multi-step feature selection, including receiver operating characteristic curves (**a**, **d**), precision-recall curves (**b**, **e**), and calibration plots (**c**, **f**), feature number (**g**), and decision curves (**h**, **i**) based on different feature selection steps. (FS, feature selection; BFSM, before-FS model; AFSM-1, after-FS-step-1 model; AFSM-2, after-FS-step-2 model; AFSM-3, after-FS-step-3 model)

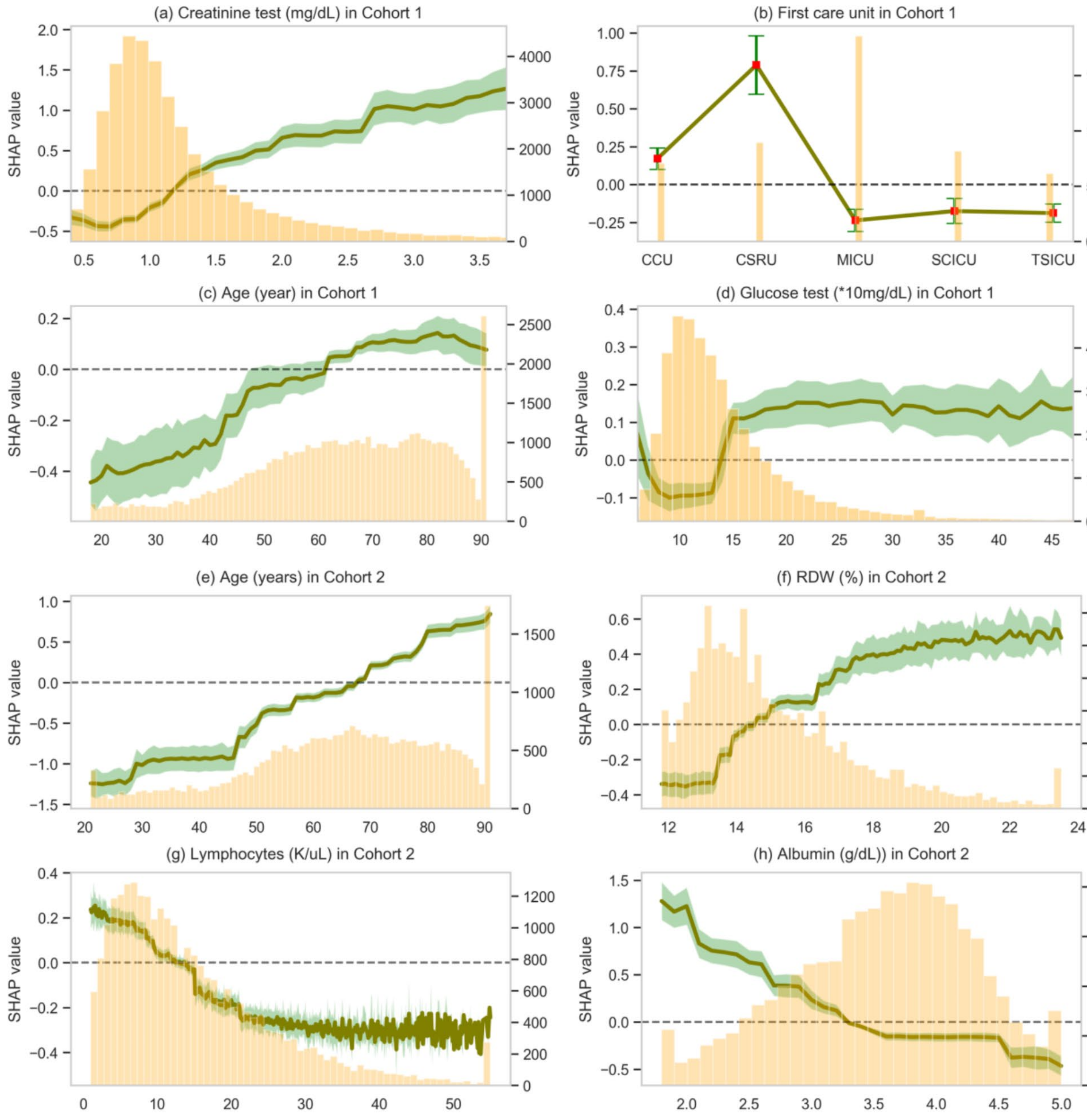

Fig. 6 Effect of varying individual feature values. (These plots show the changes in predicted outcomes across all values of a given feature. The olive line or red dot represents the average predicted outcome for all samples with a given feature value, while the light green area shows the corresponding standard deviation. The orange bars depict the distribution of the feature values. Individuals older than 89 are uniformly classified in the 91-year-old group. CCS, coronary care unit; CSRU, cardiac surgery recovery unit; MICU, medical intensive care unit; SICU, surgical intensive care unit; TSICU, trauma/surgical intensive care unit)

Discussion

Identifying important variables is a critical step in developing accurate prediction models using high-dimensional EMR data [13, 36]. Minor variations in training samples or differences in criteria for assessing feature importance can lead to divergent feature rankings and sets, complicating decision-making and reducing clinicians' trust in machine learning models [37, 38]. In this study, we developed and validated a comprehensive multi-step FS framework that not only achieves high predictive accuracy but also ensures the stability and clinical interpretability of the selected feature sets. By incorporating interpretability methods and validating findings with expert knowledge, we gained valuable insights into the underlying factors influencing risk prediction, making the selected features more understandable and trustworthy for researchers and clinicians.

The multi-step FS framework employed in this study is both reasonable and practical. In Step 1, we implemented univariate FS methods to filter out a substantial amount of redundant EMR information, a widely accepted approach [13]; in Step 2, we applied advanced multivariate FS techniques to identify the top K features based on accuracy, similarity, and stability, thereby enhancing prediction reliability. This stage further reduced dimensionality, enabling effective subsequent expert validation and screening. In Step 3, by applying artificial intelligence interpretability methods and involving clinicians, we successfully eliminated ambiguously defined features and unexplained variables that could arise from confusion or statistical artifacts. This collaborative approach enhanced the medical interpretation of the final set of selected factors, thereby increasing the likelihood of clinician trust and utilization of the predictive model.

To ensure a robust and reliable approach for selecting the most relevant and dependable feature subset, we employed efficient embedded FS methods that combine feature selection with model training processes and conducted comparative analyses focusing on accuracy, similarity, and stability in Step 2. It was imperative to select a classifier model that demonstrated excellent discrimination for our data and specific problem; we evaluated ten machine learning methods and found that tree-based ensemble models exhibited superior predictive performance (see Fig. 2). Although noticeable fluctuations existed in the similarity between the important features identified by different FS methods and in the stability of these features across varying samples, selecting an appropriate number of top-ranked features mitigated these discrepancies (see Fig. 4). In two cohorts, predictive accuracy stabilized at a lower top K value compared to similarity and stability. This suggests that including more features can enhance interpretability without sacrificing performance. We employed a conservative strategy by selecting the maximum top K value where all three metrics remained stable (see Supplementary Figure S1, S2 and Table S3).

To address this trade-off, we incorporated expert knowledge verification in Step 3. The small number of important feature subsets screened in the first two steps allows for effective final expert review. Through assessment by clinical experts and consideration of existing medical evidence, we optimized the feature set to achieve a balance between feature quantity and interpretability. Involving clinical experts enabled us to refine the feature set to include variables that were both statistically significant and relevant to clinical practice. Experts evaluated the selected features for clinical importance and removed those lacking relevance or deemed redundant. This process reduced the total number of features while maintaining or enhancing interpretability and predictive performance (see Supplementary Table S6).

Most of the risk factors identified aligned with established clinical knowledge and literature (see Supplementary Tables S5 and S6). For instance, in predicting AKI in Cohort 1, age emerged as a significant factor due to reduced kidney reserve in older individuals [39], while laboratory tests such as albumin and bilirubin levels indicated liver function issues known to interact with kidney function [40]. Cardiovascular measurements, including heart rate and blood pressure, were also critical due to their direct impact on renal health. It is important to note that factors selected through data-driven methods have strong predictive power but are not necessarily direct causal triggers [19]. For example, "*Race*" frequently appears as a strong predictor in healthcare models, reflecting underlying demographic, social, and environmental factors rather than being a direct cause of health outcomes [15]. Similarly, "*Med_2747 [Antihyperlipidemic—HMG CoA Reductase Inhibitors (statins)]*" is connected to hospitalization outcomes primarily due to pre-existing cardiovascular conditions, rather than being a direct cause hospitalization or mortality [41].

It is essential to acknowledge that, due to the complexity of medical events and the cumbersome nature of EMR data, the features selected through data-driven methods for high-risk prediction may not always be perfect. For example, some ambiguous or poorly defined features, such as "*Med_320 [Antitussives - Non-Opioid]*" may be selected by data-driven methods, but experts might remove them due to broad variability in ingredients and effects, complicating consistent impact assessment [42]. Additionally, data-driven methods may not capture all relevant risk factors accurately, and issues such as missing data and biases can undermine their reliability [43]. To overcome these limitations, it is crucial to complement data-driven approaches with expert knowledge validation [44]. Involving healthcare professionals and subject matter experts helps incorporate contextual understanding and ensures that the selected features are meaningful.

Interpretability in our framework encompasses three key aspects: (1) A stable and consistent feature set enhances clinical interpretability by providing reliable and understandable predictors [37, 38]. (2) Ensuring that selected features are consistent with existing medical evidence and manifest clear medical significance helps reduce biases inherent in purely data-driven methods, thereby enhancing the model's clinical validity and trustworthiness (see Supplementary Tables S5 and S6). (3) Utilizing SHAP values provided additional interpretability by quantifying each feature's contribution to the model's predictions (see Fig. 5 and Supplementary Figure S4). This approach aligns with high standards required in the medical field for model transparency, enabling clinicians to understand how individual features influence outcomes, which ultimately improves the credibility and utility of machine learning models in healthcare.

There are several limitations to our study. First, it was validated with only two cohorts, highlighting the need for additional testing across a wider range of clinical outcomes in future research. However, the diversity represented by these cohorts suggests the generalizability of our FS framework. Second, not all available EMR variables were utilized due to the absence of detailed timestamps for certain data categories, such as diagnosis information and surgical procedures. Including more comprehensive feature data could potentially improve prediction performance, although it is unlikely to alter the primary conclusions of this study. Third, our analysis was restricted to tabular-style data, and we observed that tree-based ensemble methods (such as LightGBM) outperformed standard deep neural networks, which aligns with previous research findings [45]. Finally, the knowledge validation of AKI risk factors in Cohort 1 was conducted by two doctors only; expanding the expert panel would enhance the reliability of our findings. To mitigate potential individual biases, we supplemented our conclusions with a comprehensive literature review. For IHM in Cohort 2, validation was based solely on literature, incorporating expert evaluations could further strengthen the robustness of the results.

Conclusions

This study presents an effective multi-step FS framework that integrates data-driven statistical inference with validation grounded in clinical domain knowledge. The effectiveness of this framework was validated using two large EMR datasets, illustrating its capability to yield consistent results across diverse scenarios. This framework enhances feature selection reliability and enables more accurate analysis in clinical outcome predictions. Our findings demonstrate its potential to advance clinical decision support systems and lay the groundwork for developing more reliable healthcare prediction models.

Abbreviations

EMR	Electronic Medical Record
FS	Feature Selection
ICU	Intensive Care Unit
ED	Emergency Department
ML	Machine Learning
AKI	Acute Kidney Injury
IHM	In-Hospital Mortality
SCr	Serum Creatinine
SHAP	SHapley Additive exPlanations
IQR	Interquartile Range
CI	Confidence Interval
AUROC	Area Under the Receiver Operating Characteristic Curve
AUPRC	Area Under the Precision-Recall Curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02922-y .

Supplementary Material 1

Acknowledgements

We are deeply grateful to Dr. Jia Zhang from the first affiliated hospital of Chongqing Medical University and Dr. Yanqin Li from the Nanfang Hospital of Southern Medical University for their effort in assessing the risk relationship between the predictors and AKI.

Author contributions

LW conceived the idea, performed the analysis, drafted and revised the manuscript. HW helped to analyze the data, frame the idea, draft and revise the manuscript. AB, XL, MZ and LM reviewed and revised the manuscript critically for scientific content. All authors reviewed the manuscript and approved the final version for publication.

Funding

This research was supported by the National Key R&D Program Intergovernmental Key Project (Grant No. 2023YFE0114300), Guangdong Natural Science Foundation General Project (Grant No. 2024A1515012112), and Guangdong Medical Research Fund Project (Grant No. A2024044). We declare that the funding body had no role in the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript. All conclusions drawn are solely those of the authors and not influenced by the funding body.

Data availability

The Medical Information Mart for Intensive Care (MIMIC) datasets are available with credentialed access from https://physionet.org. Code for reproducing the results is available at https://github.com/hongnianwang/MultiStep_EMR_FS.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]School of Management, Jinan University, Guangzhou 510632, China
[2]Key Laboratory of Digital-Intelligent Disease Surveillance and Health Governance, North Sichuan Medical College, Nanchong 637100, China
[3]School of Social Work, Henan Normal University, Xinxiang 453007, China
[4]Institute of Sciences in Emergency Medicine, Department of Emergency Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China
[5]Department of Emergency Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China
[6]Department of Emergency Medicine, Wayne State University School of Medicine, Detroit, MI 48201, USA
[7]Global Network on Emergency Medicine, Brookline, MA, USA
[8]Medical Research Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China

Received: 12 July 2024 / Accepted: 5 February 2025
Published online: 17 February 2025

References

1. Kellum JA, Romagnani P, Ashuntantang G, Ronco C, Zarbock A, Anders H-J. Acute kidney injury. Nat Rev. 2021;7:52.
2. Li X, Wang P, Zhu Y, Zhao W, Pan H, Wang D. Interpretable machine learning model for predicting acute kidney injury in critically ill patients. BMC Med Inf Decis Mak. 2024;24:148.
3. Kamel Rahimi A, Ghadimi M, van der Vegt AH, Canfell OJ, Pole JD, Sullivan C, et al. Machine learning clinical prediction models for acute kidney injury: the impact of baseline creatinine on prediction efficacy. BMC Med Inf Decis Mak. 2023;23:207.
4. Kellum JA, Bihorac A. Artificial intelligence to predict AKI: is it a breakthrough? Nat Rev Nephrol. 2019;15:663–4.
5. 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. 2022;22:17.
6. Ronco C, Bellomo R, Kellum JA. Acute kidney injury. Lancet. 2019;394:1949–64.
7. Kim HJ, Kim J, Ohn JH, Kim N-H. Impact of hospitalist care model on patient outcomes in acute medical unit: a retrospective cohort study. BMJ Open. 2023;13:e069561.
8. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. 2019;23:64.
9. Hsieh M-J, Hsu N-C, Lin Y-F, Shu C-C, Chiang W-C, Ma MH-M, et al. Developing and validating a model for predicting 7-day mortality of patients admitted from the emergency department: an initial alarm score by a prospective prediction model study. BMJ Open. 2021;11:e040837.
10. Kashani K, Herasevich V. Utilities of Electronic Medical Records to Improve Quality of Care for Acute kidney Injury: past, Present, Future. Nephron. 2015;131:92–6.
11. Omuya EO, Okeyo GO, Kimwele MW. Feature selection for classification using principal component analysis and information gain. Expert Syst Appl. 2021;174:114765.
12. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinf. 2022;2:927312.
13. Wu L, Hu Y, Liu X, Zhang X, Chen W, Yu ASL, et al. Feature ranking in Predictive models for Hospital-Acquired Acute kidney Injury. Sci Rep. 2018;8:1–11.
14. Wu L, Hu Y, Yuan B, Zhang X, Chen W, Liu K, et al. Which risk predictors are more likely to indicate severe AKI in hospitalized patients? Int J Med Informatics. 2020;143:104270.
15. Wu L, Hu Y, Liu M, Yuan B, Zhang X, Chen W, et al. Temporal dynamics of clinical risk predictors for hospital-acquired acute kidney injury under different forecast time windows. Knowl Based Syst. 2022;245:108655.
16. Wu L, Hu Y, Zhang X, Zhang J, Liu M. Development of a knowledge mining approach to uncover heterogeneous risk predictors of acute kidney injury across age groups. Int J Med Informatics. 2022;158:104661.
17. Flechet M, Falini S, Bonetti C, Güiza F, Schetz M, den Berghe G, et al. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIpredictor. Crit Care. 2019;23:1–10.
18. Sato N, Uchino E, Kojima R, Hiragi S, Yanagita M, Okuno Y. Prediction and visualization of acute kidney injury in intensive care unit using one-dimensional convolutional neural networks based on routinely collected data. Comput Methods Programs Biomed. 2021;206:106129.
19. Zhang M, Zhang X, Dai M, Wu L, Liu K, Wang H, et al. Development and validation of a multi-causal investigation and discovery framework for knowledge harmonization (MINDMerge): a case study with acute kidney injury risk factor discovery using electronic medical records. Int J Med Informatics. 2024;191:105588.
20. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572:116–9.
21. Kim WH, Lee SM, Choi JW, Kim EH, Lee JH, Jung JW, et al. Simplified clinical risk score to predict acute kidney injury after aortic surgery. J Cardiothorac Vasc Anesth. 2013;27:1158–66.
22. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. Data Descriptor: MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
23. 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.
24. Kellum JA, Lameire N, Aspelin P, Barsoum RS, Burdmann EA, Goldstein SL, et al. Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. Kidney Int Supplements. 2012;2:1–138.
25. Flechet M, Güiza F, Schetz M, Wouters P, Vanhorebeek I, Derese I, et al. AKIpredictor, an online prognostic calculator for acute kidney injury in adult critically ill patients: development, validation and comparison to serum neutrophil gelatinase-associated lipocalin. Intensive Care Med. 2017;43:764–73.
26. Kuncheva LI. A stability index for feature selection. Artificial intelligence and applications. 2007. pp. 421–7.
27. Cannas LM, Dessì N, Pes B. Assessing similarity of feature selection techniques in high-dimensional domains. Pattern Recognit Lett. 2013;34:1446–53.
28. Chen T, Guestrin C, XGBoost:. A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. pp. 785–94.
29. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018;31.
30. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–54.
31. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017. pp. 4765–74.
32. Ververidis D, Kotropoulos C. Sequential forward feature selection with low computational cost. 13th Eur Signal Process Conf. 2005;2005:1–4.
33. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32–5.
34. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett. 2014;21:1389–93.
35. Cheung Y-W, Lai KS. Lag order and critical values of the augmented Dickey–fuller test. J Bus Economic Stat. 1995;13:277–80.
36. Zhang M, Wang H, Zhao J. Use machine learning models to identify and assess risk factors for coronary artery disease. PLoS ONE. 2024;19:e0307952.
37. Spooner A, Mohammadi G, Sachdev PS, Brodaty H, Sowmya A. For the Sydney Memory and Ageing Study and the Alzheimer's Disease Neuroimaging Initiative. Ensemble feature selection with data-driven thresholding for Alzheimer's disease biomarker discovery. BMC Bioinformatics. 2023;24:9.
38. Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput Applic. 2020;32:5951–73.
39. Kane-Gill SL, Sileanu FE, Murugan R, Trietley GS, Handler SM, Kellum JA. Risk factors for acute kidney Injury in older adults with critical illness: a retrospective cohort study. Am J Kidney Dis. 2015;65:860–9.
40. Lal BB, Alam S, Sood V, Rawat D, Khanna R. Profile, risk factors and outcome of acute kidney injury in paediatric acute-on-chronic liver failure. Liver Int. 2018;38:1777–84.
41. Iihara K, Nishimura K, Kada A, Nakagawara J, Ogasawara K, Ono J, et al. Effects of Comprehensive Stroke Care capabilities on In-Hospital mortality of patients with ischemic and hemorrhagic stroke: J-ASPECT study. PLoS ONE. 2014;9:e96819.
42. Lee JH, Lim J, Han SJ, do Moon S, Moon H, Lee S-Y, et al. Clinical outcomes associated with anticholinergic burden in older hospitalized patients with advanced cancer: a single-center database study. Support Care Cancer. 2021;29:4607–14.
43. An D, Kim NH, Choi J-H. Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. Reliab Eng Syst Saf. 2015;133:223–36.
44. Sun J, Hu J, Luo D, Markatou M, Wang F, Edabollahi S et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annual Symposium Proceedings. 2012. p. 901.
45. Lundberg S, Erion GG, Chen H, Degrave AJ, Prutkin JM, Nair BG, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67.

Publisher's note