BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Explainable machine learning algorithm to predict cardiovascular event in patients undergoing peritoneal dialysis

Qiqi Yan[1,2], Guiling Liu[1,2], Ruifeng Wang[1,2], Dandan Li[1,2], Xiaoli Chen[1,2], Jingjing Cong[1,2] and Deguang Wang[1,2*]

Abstract

Objective To compare the performance of predictive models for cardiovascular event (CVE) in patients undergoing peritoneal dialysis (PD) based on machine learning algorithm and Cox proportional hazard regression.

Methods This study included patients underwent PD catheterization in our center from January 1, 2010, to July 31, 2022. The patients were randomly divided into training and validation sets in a 7:3 ratio. Cox regression, extreme gradient boosting (XGBoost), and random survival forest (RSF) models were developed using the training set and validated using the validation set. The time-dependent area under the curve (AUC) and concordance index (C-index) were used to evaluate the discriminative ability of predictive models.

Results A total of 318 patients were enrolled in this study. 110 (34.6%) patients developed CVE during the median follow-up of 31(16,56) months. The RSF model had better predictive performance, with a C-index of 0.725 and 1-, 3-, and 5-year time-dependent AUC of 0.812, 0.836, and 0.706 in the validation set, respectively. The top 5 important variables identified were platelet count, age, 4 hD/Pcr, left atrium diameter, and left ventricular diameter. Patients were classified into high-risk and low-risk groups based on the cut-off risk score calculated using the maximally selected rank statistics in the validation set. The log-rank test showed a significant difference in cumulative CVE-free survival probability between the two groups.

Conclusion The RSF model may be a useful method for evaluating CVE risk in PD patients.

Keywords Cardiovascular event, Peritoneal dialysis, Machine learning, Predictive model, Random survival forest

*Correspondence:
Deguang Wang
wangdeguang@ahmu.edu.cn
[1]Department of Nephrology, The Second Affiliated Hospital of Anhui Medical University, Hefei, China
[2]Institute of Kidney Disease, Inflammation & Immunity Mediated Diseases, The Second Affiliated Hospital of Anhui Medical University, Hefei, China

Introduction

Cardiovascular event (CVE) remains a leading cause of morbidity and mortality among patients undergoing peritoneal dialysis (PD) [1]. The burden of cardiovascular disease (CVD) in this population is significantly higher than in the general population, largely due to the presence of multiple risk factors, including inflammation, oxidative stress, abnormal calcium-phosphorus metabolism, overhydration, and high glucose load [2, 3]. Thus, early identification and management of high-risk patients are critical to improving outcomes.

Despite advances in dialysis techniques, accurately predicting the occurrence of CVE in patients undergoing PD remains challenging. Traditional risk assessment tools, such as the Framingham Risk Score [4], are not fully applicable to patients undergoing PD due to their differences in risk factors and pathophysiological characteristics. Therefore, there is an urgent need for more accurate and reliable predictive models tailored to this high-risk population.

The Cox proportional hazards regression model is a classic statistical method used to establish the relationship between survival time and specific risk factors [5]. However, it assumes proportional hazards and linear relationships between predictors and outcomes, which may not always be valid in complex clinical settings. In recent years, machine learning (ML) algorithms have emerged as powerful tools for risk prediction in various medical fields [6, 7]. These algorithms are especially effective in analyzing complex datasets, as they can handle non-linear relationships and interactions among multiple predictors, making them ideal for survival analysis in clinical settings. Among these, random survival forest (RSF) has shown promise in survival analysis by extending the random forest algorithm to handle censored data. Prediction models based on this algorithm can effectively identify prognostic risk factors and screen patients with poor prognosis [8]. Extreme gradient boosting (XGBoost), a widely used boosting algorithm, has demonstrated excellent performance in various predictive tasks, including survival analysis [9, 10]. By constructing an ensemble of decision trees, XGBoost improves model accuracy and reduces overfitting, making it a powerful tool for handling complex datasets.

In this study, we aimed to develop and validate predictive models for CVE in patients undergoing PD using both traditional Cox proportional hazard regression and ML algorithms. By comparing the performance of different models, including the Cox proportional hazard regression model, RSF model, and XGBoost model, our ultimate goal was to identify the best predictive model for enabling earlier and more accurate screening of patients at high risk of CVE. Given the superior survival data-processing capabilities of RSF, we hypothesize that the RSF model may be a useful method for evaluating CVE risk in PD patients. By using the RSF model to assess CVE risk, clinicians may be able to better stratify patients based on their risk levels, facilitating the development of personalized medical strategies and improving patient outcomes.

Materials and methods

Study population

This single-center retrospective cohort study included 422 patients who underwent PD catheterization and continuous ambulatory peritoneal dialysis or daytime ambulatory peritoneal dialysis at the Second Affiliated Hospital of Anhui Medical University from January 1, 2010, to July 31, 2022. Exclusion criteria included patients younger than 18 years old, transferred from long-term hemodialysis, with less than 3 months of PD therapy, lacking necessary data, recently experienced severe infections, used glucocorticoids or immunosuppressants within 6 months, and a history of hematological diseases or malignant tumors. Based on these criteria, 318 patients were included in our subsequent analyses. The subjects were randomly divided into a training set (70%, n = 228) to develop the model and a validation set (30%, n = 90) to validate the performance of the model. Model development and validation followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement [11]. This study received approval from the Ethics Committee of the Second Affiliated Hospital of Anhui Medical University and adhered to the Declaration of Helsinki.

Data collection

The baseline data defined at PD start were obtained from the electronic medical record system of our hospital. In total, 39 patient characteristics were collected as candidate covariates. The patient demographic data were as follows: age, sex, body mass index (BMI), causes of end-stage kidney disease (ESKD), and history of diabetes mellitus or CVD. The laboratory data were as follows: hemoglobin, white blood cell count, platelet count, C-reactive protein (CRP), high-sensitivity C-reactive protein (hs-CRP), serum creatinine, serum urea nitrogen, serum uric acid, serum corrected calcium, serum phosphorus, serum albumin, alkaline phosphatase, total cholesterol, triglycerides, fasting blood glucose, ferritin, transferrin saturation, and fibrinogen. The parameters measured in echocardiography were as follows: left ventricular (LV) end-diastolic diameter (LVEDd), left atrium (LA) diameter, right ventricular (RV) diameter, interventricular septum thickness (IVST), left ventricular posterior wall thickness (LVPWT), LV ejection fraction (EF), and pulmonary arterial pressure (PAP). The left ventricular mass (LVM) was calculated by the Devereux formula: $LVM = 1.04 \times [(IVST + LVPWT + LVE$

Dd)^3–LVEDd^3]–13.6. The LVM index (LVMI) was calculated as LVM divided by the body surface area. Left ventricular hypertrophy (LVH) was defined as LVMI > 125 g/m^2 in males and > 120 g/m^2 in females [12]. Cardiac valve calcification (CVC) was defined as bright echoes > 1 mm on one or more cusps of the aortic valve, mitral valve, or mitral annulus. In addition, data were collected on 24-hour ultrafiltration volume, 24-hour urine volume, estimated glomerular filtration rate (eGFR), total weekly Kt/V (tKt/V), total weekly creatinine clearances (tCCr) and the dialysate/plasma creatinine ratio at 4 h (4 hD/Pcr) measured using the peritoneal equilibration test after 1 month of PD treatment. Variables with more than 20% missing data (CRP, hs-CRP, PAP, eGFR, tKt/V, and tCCr) and the variable exhibiting multicollinearity (variance inflation factor > 10), ESKD, were removed. Finally, the study analyzed 32 variables. Nineteen variables (serum urea nitrogen, alkaline phosphatase, total cholesterol, triglycerides, fasting blood glucose, ferritin, transferrin saturation, fibrinogen, 4 hD/Pcr, 24-hour ultrafiltration volume, 24-hour urine volume, LV, LA, RV, IVST, LVPWT, LVEF, LVH, and CVC) had missing data, which were handled using multiple imputation by chained equations. Specifically, the predictive mean matching method was applied for imputation, with five imputations performed. No outlier values were removed for the descriptive data.

Outcome definition and follow-up

The primary outcome of this study was the incidence of CVE. CVE was defined according to the International Classification of Diseases, 10th Revision (ICD-10) codes for New York Heart Association class III–IV congestive heart failure requiring hospitalization, unstable angina, acute myocardial infarction, stroke, or cardiac arrest. The time to the first occurrence of CVE was recorded in months. The endpoint of follow-up was defined as the first occurrence of CVE, death, discontinued PD treatment, loss to follow-up, or censoring on July 31, 2023, whichever came first.

Statistical analysis

Continuous variables were presented as mean ± standard deviation or median (interquartile range), while categorical variables were presented as frequency (percentage). Data normality was tested using the Shapiro-Wilk test. The table for patient characteristics was generated using the 'tableone' package.

Cox proportional hazard regression is a simple and effective tool for identifying risk factors associated with the incidence and prognosis of diseases. As a semiparametric method, it does not assume any specific distribution for survival times, but it requires that the impact of different variables on the survival rate remains constant and that the effects of these variables are additive [13]. The Least Absolute Shrinkage and Selection Operator (LASSO) regression assumes an approximately linear relationship between predictor variables and the outcome. By adding an L1 penalty to the regression, it effectively selects variables, prevents overfitting, and mitigates the impact of multicollinearity. This method is more accurate than stepwise regression [14], especially when the sample size is not large [15]. We used LASSO regression analysis to identify potential risk factors among candidate variables. To determine the most predictive variables, we utilized 10-fold cross-validation, selecting those with the minimum cross-validated error. Subsequently, stepwise multivariate Cox regression analysis was conducted to identify the independent risk factors for CVE in patients undergoing PD. The proportional hazards assumption of the Cox model was tested using Schoenfeld residuals. Based on the minimum value of the Akaike Information Criterion (AIC), we built the final model and constructed a nomogram.

Two different ML algorithms were considered: XGBoost and RSF. XGBoost is a gradient boosting algorithm widely used for classification and regression tasks. Its excellent performance and scalability make it a powerful tool for survival prognosis analysis [16]. XGBoost models are built by gradually constructing multiple decision trees and then combining them to obtain more accurate predictions. We implemented XGBoost using the 'xgboost' package.

RSF is an adaptation of the random forest method, used for analyzing survival data through ensemble learning of decision trees [17]. In RSF, multiple decision trees are built using bootstrap samples. For each node, a subset of features is randomly selected, and the node is split based on a survival criterion that includes survival time and censoring status [18]. The final prediction is made by aggregating the results from all trees. RSF is particularly useful when dealing with a large number of predictors and complex relationships between the response and predictors [17]. Compared to the Cox proportional hazard model, the advantage of the RSF model is that it is not constrained by assumptions of proportional hazards and log-linearity. As an ensemble learning method, RSF aggregates the predictions of multiple decision trees, reducing the risk of overfitting and enhancing the reliability of predictions. Additionally, RSF provides a measure of variable importance, which helps in improving the explainability of the model and identifying the most influential predictors. The RSF model was implemented using the 'randomForestSRC' package.

All 32 variables were used to develop the XGBoost and RSF models in the training set. The hyperparameters of the models were optimized using grid search combined with 10-fold cross validation. For the RSF model,

we performed hyperparameter tuning using the concordance index (C-index) as the evaluation criterion. The C-index for each fold in the cross-validation was calculated and averaged to assess the performance of each hyperparameter set. For the XGBoost model, we used Cox-Negative Log-Likelihood as the tuning criterion, which is a standard measure of model fit in survival analysis. The goal was to minimize this value during the cross-validation process. For the XGBoost model, the optimal hyperparameters are: eta = 0.01, max_depth = 3, min_child_weight = 5, subsample = 0.3, colsample_bytree = 0.3, gamma = 5, and alpha = 1. For the RSF model, the optimal hyperparameters are: mtry = 1, nodesize = 30, ntree = 100, alpha = 0.1, and minprop = 0.1 (Additional file 1. Table S1). The discriminative ability of the predictive models was evaluated using the time-dependent area under the curve (AUC) and C-index, under the assumption that the censoring mechanism was independent of survival time. The calibration capability was assessed using calibration curves. The best predictive models in this study were selected based on their C-index and AUC at 1, 3, and 5 years to ensure optimal discrimination and predictive performance.

To better understand how the RSF model generates predictions, we utilized SHapley Additive exPlanations (SHAP) values. SHAP values, grounded in game theory, quantify the contribution of each feature to the model's prediction [6]. The maximally selected rank statistics method was employed to determine an optimal cut-off point corresponding to the strongest association with incident CVE. Kaplan-Meier analysis curves and the log-rank test were utilized to assess the distribution of incident CVE, assuming that the hazard ratio between groups remains constant over time. A web-based risk calculator was developed using the 'shiny' package. A non-probability consecutive sampling method was used in this study. The sample size was determined based on the availability of eligible patients. Data processing was carried out using R software (version 4.3.1, R Foundation for Statistical Computing, Vienna, Austria). $P < 0.05$ was considered statistically significant.

Results

Patient characteristics

A total of 318 patients undergoing PD were enrolled in this study (Fig. 1). Missing data were displayed in Additional file 2. Table S2. 110 (34.6%) patients developed

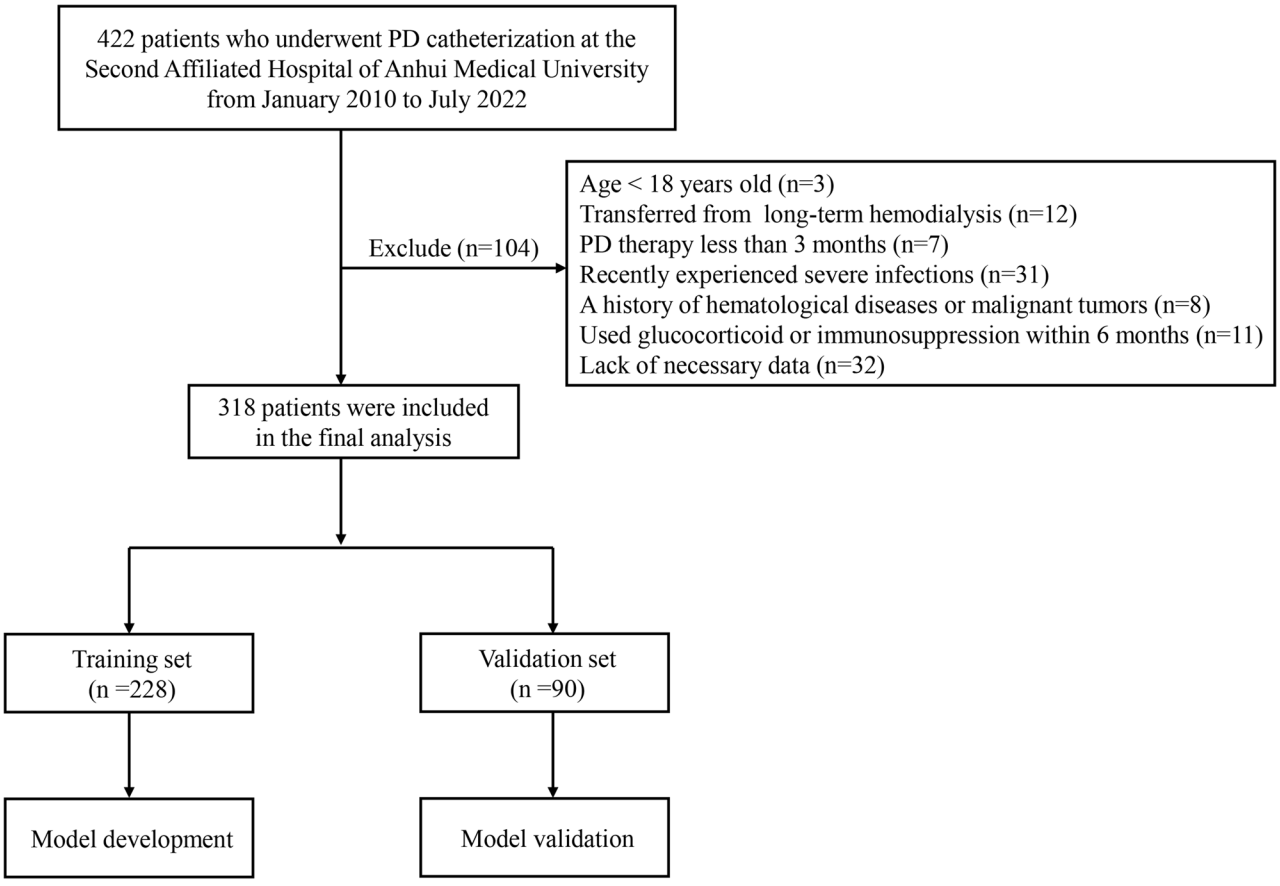

Fig. 1 The flowchart of the present study. PD: peritoneal dialysis

CVE during the median follow-up of 31(16,56) months. At 1-, 3-, and 5-years, the cumulative incidence of CVE was 9.4%, 21.7%, and 29.9%, respectively. Compared to patients without CVE, those who developed CVE were older and had significantly higher platelet count and shorter follow-up time. Additionally, they exhibited lower levels of serum urea nitrogen and serum albumin (Table 1). During the modeling process, the study population was divided into two cohorts, the training set (n = 228) and the validation set (n = 90). There was no statistical difference in any variables between the training set and the validation set, indicating that the split was balanced (Additional file 3. Table S3).

LASSO-Cox regression analysis and nomogram

For LASSO regression, we used 10-fold cross-validation and selected the parameter λ based on the one standard error of the minimum criteria. The optimal tuning parameter λ was 0.0395. At this optimal λ value, 11 variables with non-zero coefficients were selected for multivariate Cox regression analysis, including age, BMI, diabetes mellitus, fibrinogen, platelet count, serum creatinine, triglyceride, fasting blood glucose, LV diameter, IVST, and 4hD/Pcr (Additional file 4. Figure S1). By using the stepwise regression method and based on the minimum value of the AIC, the final Cox model included age, BMI, platelet count, triglyceride, fasting blood glucose, LV diameter, and 4hD/Pcr (Additional file 5. Table S4). The C-index of the Cox model in the training set and validation set was 0.688 and 0.685, respectively (Table 2). In the nomogram constructed using these variables (Fig. 2), each variable corresponds to a specific point by drawing a straight line upward to the points axis. After summing the points for all variables, locate the total points on the

Table 1 Characteristics of the study populations

Variables	Total (n = 318)	No CVE group (n = 208)	CVE group (n = 110)	*P* Value
Male n (%)	152 (47.8)	98 (47.1)	54 (49.1)	0.83
Age (years)	52.00 (41.25, 61.00)	50.00 (39.00, 60.00)	56.00 (47.00, 63.00)	**< 0.01**
Body mass index (kg/m^2)	21.74 (19.73, 23.95)	21.76 (19.70, 23.88)	21.66 (19.80, 24.37)	0.86
Diabetes mellitus n (%)	65 (20.4)	37 (17.8)	28 (25.5)	0.14
Cardiovascular disease n (%)	44 (13.8)	24 (11.5)	20 (18.2)	0.14
Fibrinogen (g/L)	3.58 (2.89, 4.48)	3.42 (2.88, 4.26)	3.71 (2.89, 4.67)	0.06
Hemoglobin (g/L)	76.00 (63.00, 90.00)	75.00 (62.75, 88.00)	77.50 (63.00, 90.75)	0.87
White blood count (10^9/L)	5.46 (4.48, 7.08)	5.37 (4.40, 6.91)	5.70 (4.67, 7.29)	0.16
Platelet count (10^9/L)	140.50 (104.00, 180.75)	132.00 (99.75, 179.25)	149.50 (116.25, 182.50)	**0.02**
Serum creatinine (μmol/L)	810.00 (665.25, 1010.00)	830.00 (692.25, 1041.25)	768.00 (625.50, 982.25)	0.06
Serum urea nitrogen (mmol/L)	29.93 (22.70, 39.48)	31.28 (23.77, 40.60)	26.66 (20.23, 38.28)	**0.01**
Serum uric acid (μmol/L)	490.00 (413.50, 591.00)	490.50 (424.00, 591.50)	490.00 (405.25, 570.25)	0.50
Serum corrected calcium (mmol/L)	2.08 (1.84, 2.23)	2.08 (1.81, 2.23)	2.09 (1.92, 2.24)	0.33
Serum phosphorus (mmol/L)	1.90 (1.56, 2.31)	1.94 (1.59, 2.33)	1.85 (1.53, 2.27)	0.54
Serum albumin (g/L)	33.22 ± 5.69	33.88 ± 5.54	31.97 ± 5.79	**< 0.01**
Alkaline phosphatase (U/L)	82.00 (64.00, 106.00)	83.00 (64.00, 105.50)	79.50 (64.00, 106.00)	0.88
Total cholesterol (mmol/L)	3.95 (3.31, 4.76)	3.84 (3.22, 4.67)	4.04 (3.40, 4.95)	0.07
Triglyceride (mmol/L)	1.15 (0.84, 1.59)	1.18 (0.83, 1.60)	1.09 (0.86, 1.55)	0.61
Fasting blood glucose (mmol/L)	4.78 (4.35, 5.38)	4.72 (4.35, 5.28)	4.87 (4.37, 5.49)	0.25
Ferritin (μg/L)	195.00 (81.60, 321.25)	183.50 (75.70, 333.50)	201.50 (86.52, 305.75)	0.87
Transferrin saturation (%)	25.00 (16.55, 34.00)	25.70 (17.00, 35.00)	24.00 (15.25, 31.90)	0.45
Left ventricular end-diastolic diameter (mm)	48.00 (44.00, 52.00)	47.00 (43.00, 51.00)	48.00 (45.00, 53.00)	0.07
Left atrium diameter (mm)	34.00 (30.00, 39.00)	34.00 (30.00, 39.00)	35.00 (30.00, 39.75)	0.51
Right ventricular diameter (mm)	21.00 (19.00, 23.00)	21.00 (19.00, 23.00)	21.00 (19.00, 23.00)	0.99
Interventricular septum thickness (mm)	10.00 (10.00, 12.00)	10.00 (9.00, 12.00)	11.00 (10.00, 12.00)	0.13
Left ventricular posterior wall thickness (mm)	10.00 (9.00, 11.00)	10.00 (9.00, 11.00)	10.00 (9.00, 11.00)	0.44
Left ventricular ejection fraction (%)	62.00 (60.00, 65.00)	62.00 (60.00, 65.00)	61.00 (58.00, 64.75)	0.14
Left ventricular hypertrophy n (%)	178 (56.0)	114 (54.8)	64 (58.2)	0.65
Cardiac valve calcification n (%)	45 (14.2)	26 (12.5)	19 (17.3)	0.32
24-hour ultrafiltration volume (mL)	334.12 ± 439.66	335.15 ± 421.80	332.17 ± 473.59	0.95
24-hour urine volume (mL)	800.00 (500.00, 1300.00)	900.00 (500.00, 1300.00)	800.00 (500.00, 1237.50)	0.27
Dialysate/plasma creatinine ratio at 4 h	0.58 (0.49, 0.69)	0.57 (0.48, 0.68)	0.60 (0.51, 0.69)	0.10
Follow-up time (months)	31.00 (16.25, 56.00)	35.00 (19.00, 59.75)	25.00 (11.00, 48.00)	**< 0.01**

CVE: cardiovascular event. Continuous variables are presented as mean ± standard deviation (for normally distributed data) or median (interquartile range) (for non-normally distributed data). Categorical variables are presented as frequency (percentage). Bold font means $P < 0.05$

Table 2 Predictive performance comparison of different methods in the training and validation sets

Indexes	Training set			Validation set		
	Nomogram	XGBoost	RSF	Nomogram	XGBoost	RSF
C-index	0.688	0.771	0.810	0.685	0.703	0.725
AUC at 1 year	0.656	0.766	0.804	0.806	0.744	0.812
AUC at 3 years	0.758	0.827	0.880	0.761	0.780	0.836
AUC at 5 years	0.720	0.776	0.825	0.610	0.719	0.706

C-index: concordance index; AUC: area under the curve; RSF: random survival forest

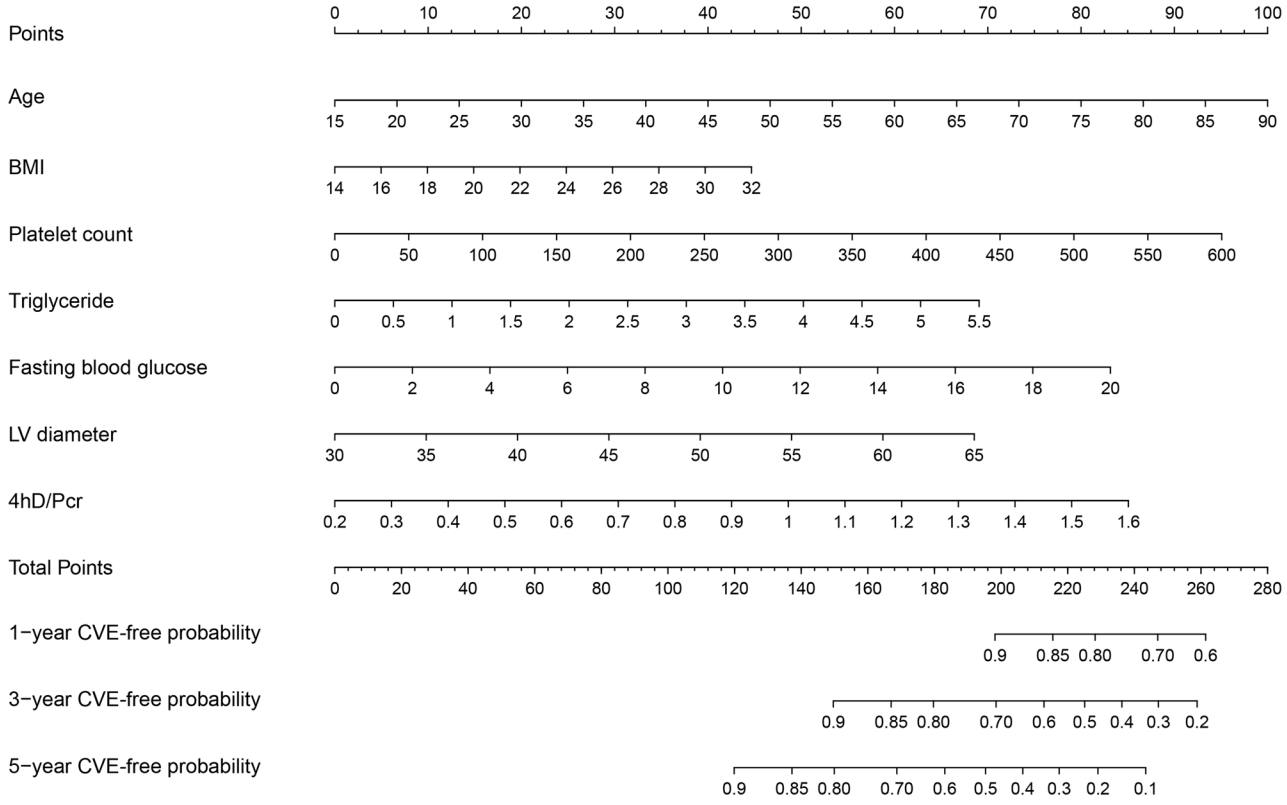

Fig. 2 Nomogram for predicting the risk of CVE in patients undergoing PD. The nomogram combines multiple clinical variables to estimate the individualized risk of CVE. BMI: body mass index; LV: left ventricular; 4hD/Pcr: dialysate/plasma creatinine ratio at 4h; CVE: cardiovascular event; PD: peritoneal dialysis

bottom scales. These total points correspond to the 1-, 3-, and 5-years CVE-free probability. To evaluate the performance of the nomogram model, we used the time-dependent receiver operating characteristic (ROC) curve. As shown in Fig. 3A, the AUC values of the nomogram model for predicting CVE in patients undergoing PD in the training set were 0.656 at 1 year, 0.758 at 3 years, and 0.720 at 5 years. In the validation set, the AUC values were 0.806 at 1 year, 0.761 at 3 years, and 0.610 at 5 years (Fig. 3D). We further plotted the calibration curves to evaluate the calibration capability of the nomogram model in the training and validation sets, and the results were shown in Additional file 6. Figure S2.

XGBoost

The C-index of the XGBoost model was 0.771 in the training set and 0.703 in the validation set (Table 2). The time-dependent ROC curves showed the AUC in the training set was 0.766 at 1 year, 0.827 at 3 years, and 0.776 at 5 years (Fig. 3B), and in the validation set was 0.744 at 1 year, 0.780 at 3 years, and 0.719 at 5 years (Fig. 3E). Additional file 7. Figure S3 depicts the calibration curves of the XGBoost model in the training and validation sets.

Random survival forest

The C-index of the RSF prediction model in the training set and validation set was 0.810 and 0.725, respectively (Table 2). The time-dependent ROC curves of the RSF model for predicting CVE in patients undergoing PD showed the AUC of 0.804, 0.880, and 0.825 for 1-, 3-,

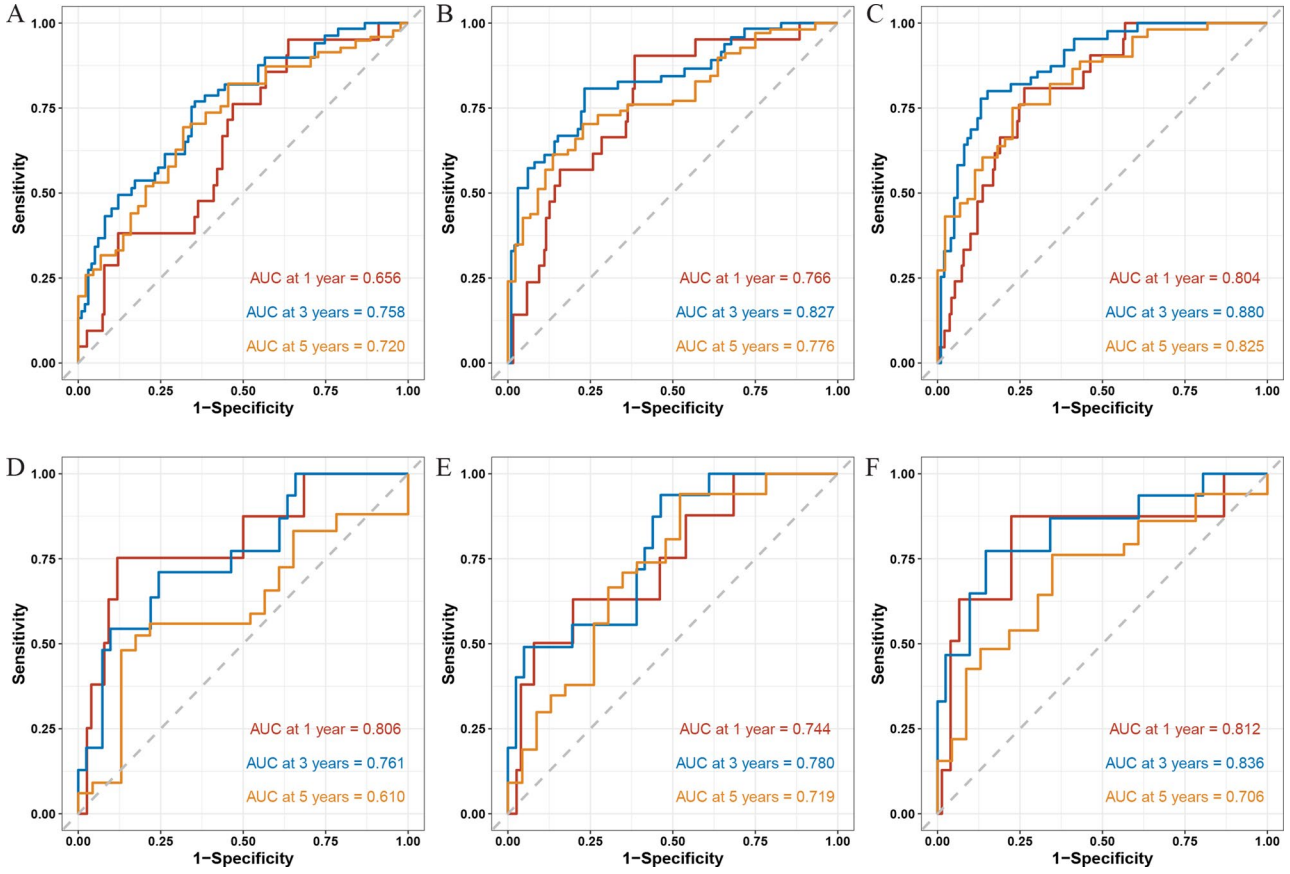

Fig. 3 Time-dependent receiver operating characteristic curves for predicting cardiovascular event in the training and validation sets. (**A**) Nomogram in the training set; (**B**) XGBoost model in the training set; (**C**) Random survival forest model in the training set; (**D**) Nomogram in the validation set; (**E**) XGBoost model in the validation set; (**F**) Random survival forest model in the validation set. The area under the curve was calculated at 1, 3, and 5 years to evaluate the models' discriminative performance

and 5-years in the training set (Fig. 3C), and 0.812, 0.836, and 0.706 in the validation set (Fig. 3F), indicating a more reliable prediction of CVE incidence. Figure 4A shows the feature importance ranking of the RSF model using SHAP summary plots, indicating that platelet count, age, 4hD/Pcr, LA diameter, and LV diameter were the top 5 contributors to the model. Figure 4B shows the relationship between individual input variable and the RSF model's predictions. The calibration curves of the RSF model in the training and validation sets were depicted in Additional file 8. Figure S4. Using maximally selected rank statistics, we calculated a cut-off risk score of 18.41 in the validation set (Fig. 5). Patients were classified into high-risk and low-risk groups based on this cut-off risk score, and the Kaplan-Meier curve was plotted (Fig. 6). The log-rank test showed a significant difference in cumulative CVE-free survival probability between the two groups, indicating that the RSF model risk score could effectively stratify patients undergoing PD by CVE risk. The 'shiny' package was used to develop a visual and operational web-based risk calculator (https://yanxzsw.shinyapps.io/RSFmodel/) for the RSF prediction model. Users can directly obtain risk score by entering the values of variables into the calculator.

Discussion

In this study, three models for predicting CVE in patients undergoing PD were developed and compared, and the results showed that the RSF model had better predictive performance, with a C-index of 0.725 and 1-, 3-, and 5-year time-dependent AUC of 0.812, 0.836, and 0.706 in the validation set, respectively. To our knowledge, this is the first study using the RSF algorithm to screen patients undergoing PD at high CVE risk.

The Cox proportional hazard regression model is a widely used statistical method for analyzing survival data. Numerous studies have employed Cox regression to explore independent risk factors for CVE or mortality and have constructed nomograms based on these findings [19, 20]. In this study, we developed a nomogram using the predictive factors screened by LASSO-Cox regression analysis, including age, BMI, platelet count, triglycerides, fasting blood glucose, LV diameter, and 4hD/Pcr. Previous research has consistently identified

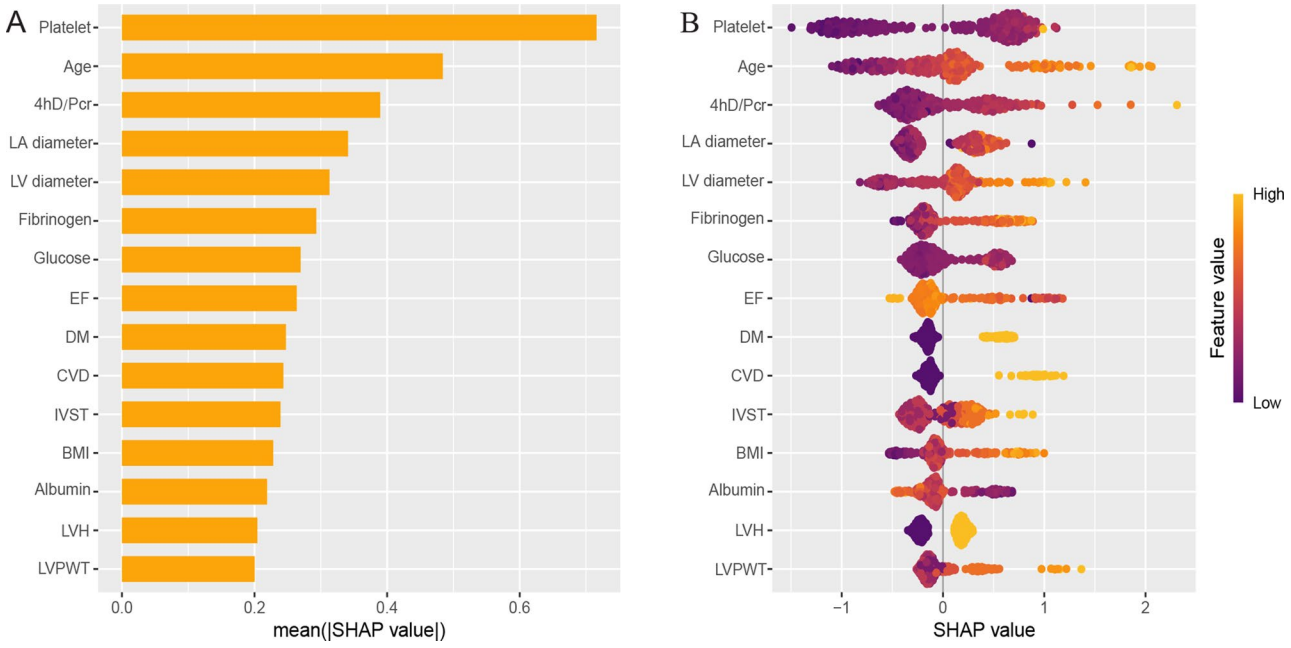

Fig. 4 SHAP summary plot for the random survival forest model. (**A**) The top 15 important features ranked by mean SHAP values, which represent the average contribution of each feature to the model's prediction. (**B**) Each patient was represented by a dot, with the x-axis position indicating the SHAP value for the corresponding feature. 4hD/Pcr: dialysate/plasma creatinine ratio at 4h; LV: left ventricular; LA: left atrium; EF: ejection fraction; DM: diabetes mellitus; CVD: cardiovascular disease; IVST: interventricular septum thickness; BMI: body mass index; LVH: left ventricular hypertrophy; LVPWT: left ventricular posterior wall thickness; SHAP: SHapley Additive exPlanations

age as a significant risk factor for CVE in patients undergoing PD [21]. However, the relationship between BMI and cardiovascular outcomes is complex. While a higher BMI predicts CVE and mortality in the general population, patients with chronic kidney disease may exhibit an "obesity paradox" where a higher BMI is associated with improved survival [22]. In our study, stepwise multivariate Cox regression analysis showed that BMI was not an independent risk factor for CVE. However, when BMI was included in the final model, the AIC was the smallest. Therefore, the final model included BMI. Platelet count were independently associated with CVE [23, 24], and patients undergoing PD with higher platelet count has an increased risk of cardiovascular mortality [25]. Multivariate Cox regression also identified that triglycerides and fasting blood glucose were independent risk factors for CVE, consistent with previous research findings. A prospective epidemiological study involving 22 countries showed that the triglyceride-glucose (TyG) index, representing insulin resistance and calculated as ln (fasting triglycerides [mg/dl] × fasting blood glucose [mg/dl]/2), was significantly associated with myocardial infarction, stroke, and cardiovascular mortality [26]. In the general population [27] and kidney transplant recipients [28], the TyG index was also associated with CVE. A larger LV diameter was associated with increased left atrial systolic force, which was related to LVH and can predict CVE incidence [29]. As for the 4hD/Pcr, it is an indicator reflecting peritoneal transport characteristics. Patients with a high 4hD/Pcr have an increased risk of atherosclerosis [30]. Overall, the predictors used to construct this nomogram are common, easily accessible, and potentially associated with the incidence of CVE. In addition, the nomogram is simple, intuitive, and easy to understand. However, it's important to note that Cox proportional hazard regression assumes proportional hazards and a linear relationship between the log hazard and predictors, which may not always hold in real-world data.

ML algorithms, a branch of artificial intelligence, have advanced rapidly in recent years and effectively complement traditional statistical methods. Previous studies have shown that the XGBoost and RSF models can predict patient prognosis and the risk of CVE [6, 31, 32]. In this study, we constructed prediction models using these two ML algorithms, and the results showed that the C-index and time-dependent ROC of the RSF model were mostly higher than those of the Cox and XGBoost models, with only the AUC value at 5 years in the validation set slightly lower than that of the XGBoost model, indicating its superior predictive performance for CVE in patients undergoing PD. The top 5 important variables in the RSF model were platelet count, age, 4hD/Pcr, LA diameter, and LV diameter. The relationship between each of the 4 indicators (platelet count, age, 4hD/Pcr, and LV diameter) and CVE have been discussed in detail above. LA function was associated with a poor prognosis

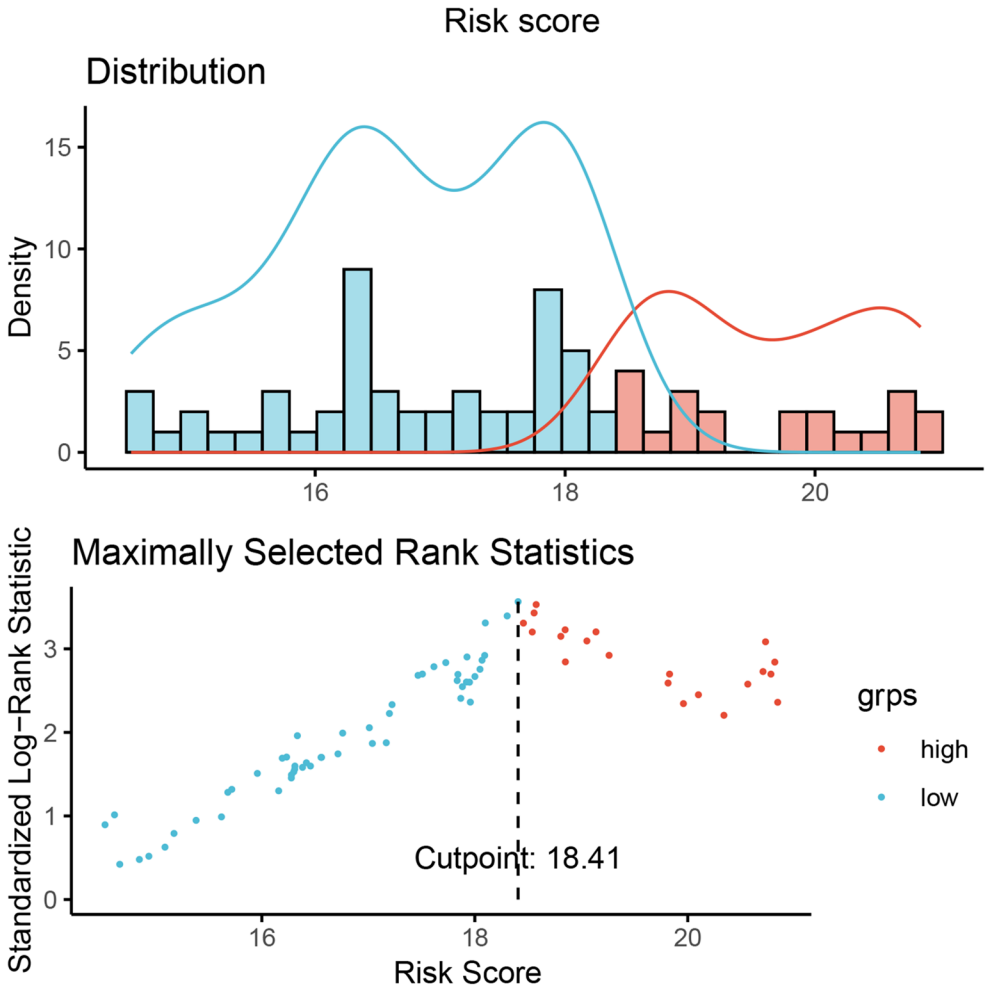

Fig. 5 Cut-off risk score calculated using the maximally selected rank statistics in the validation set. The optimal cut-off point corresponds to the strongest association with incident cardiovascular event. This method was used to divide patients into high-risk and low-risk groups based on their individual predicted risk scores

in patients with heart failure with preserved or mildly reduced ejection fraction, and atrial failure had significant predictive value for CVE [33]. Studies have also shown that LA enlargement was closely related to atrial fibrillation, which is a known risk factor for CVE [34, 35].

Due to the fact that traditional risk assessment tools are not fully applicable to patients undergoing PD, developing a predictive model for this population is both crucial and meaningful. Some studies focusing on patients undergoing PD have already used ML algorithms to construct predictive models. Xu et al. used three ML algorithms, including XGBoost, RF, and adaptive boosting (AdaBoost), to predict the risk of heart failure and all-cause mortality in patients undergoing PD. The results showed that the ML algorithms were superior to the Cox model; however, this study used an RF that was not designed for survival analysis, which may have limited its ability to fully utilize survival time [36]. Yang et al. used ML algorithms to predict adverse prognoses in patients undergoing PD, and their results showed that the categorical boosting (CatBoost) model had the best predictive performance [37]. Another study proposed an attention-based deep learning model to predict major adverse CVE in patients undergoing PD, but the model could not calculate and analyze certain medical indicators for risk prediction. Additionally, its scalability and generalizability need further improvement [38]. In this study, we fully considered the survival time in survival data when constructing ML models to enhance the accuracy of risk prediction, and ultimately found that the RSF model has better predictive accuracy and robustness. We also used SHAP values to visualize the overall feature importance of the RSF model, improving its interpretability. Moreover, the clinical indicators included in the RSF model are easily measurable and collectible, making the model practical and accessible for routine use

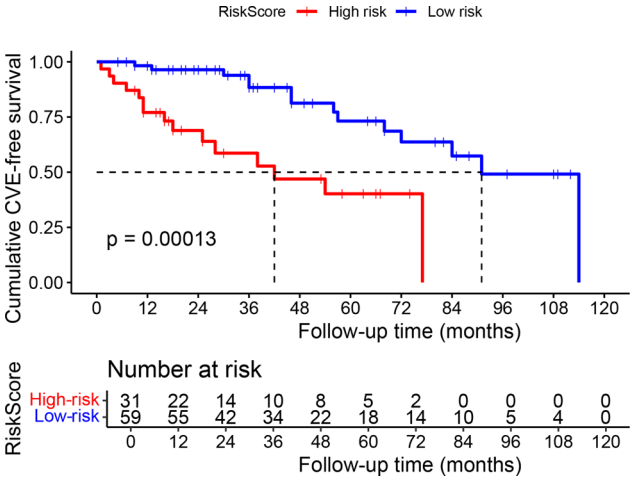

Fig. 6 Kaplan-Meier curve showing the cumulative CVE -free survival probability for high-risk and low-risk groups divided based on the cut-off risk score from the random survival forest model. The log-rank test was used to compare survival between the two groups. CVE: cardiovascular event

without the need for additional, costly tests. We have also developed a web-based risk calculator that allows users to easily calculate risk score by entering the values of variables. By using the RSF model to calculate the risk score for CVE, clinicians can effectively stratify patients based on their risk level. This enables better identification of high-risk individuals and the development of personalized medical strategies. Such a personalized approach allows for targeted interventions, proactive management of high-risk populations, and the timely initiation of treatment strategies, such as enhanced monitoring or preventive therapies.

There are several limitations in this study. Firstly, the study was conducted at a single center, and the sample size was limited, which affected the statistical power and generalizability of the findings. The calibration curve for the RSF model showed suboptimal calibration, even in the training set, which may indicate the RSF model was undertrained. This study also did not enforce consistent feature sets across models, which may affect the fairness of performance comparisons. Secondly, variables with more than 20% missing data were excluded, and multiple interpolation methods were used to handle the remaining missing values. However, there is still potential for interpolation bias, which may affect model accuracy. Additionally, some important indicators, such as CRP and hs-CRP, were excluded during model construction due to missing data, which may limit the comprehensiveness of the analysis. Thirdly, common risk factors for CVE in ESKD patients, such as hypertension and dyslipidemia, were not included in the analysis due to data limitations. Fourth, the median follow-up period of 31 months may not be long enough to comprehensively assess long-term cardiovascular risk, particularly for 5-year outcomes. This could lead to data insufficiency, estimation uncertainty, and potential censoring bias. Finally, the findings of this study require external validation to confirm their generalizability and robustness.

Conclusions

The RSF model may be a useful method for evaluating CVE risk in PD patients.

Abbreviations

CVE	Cardiovascular event
PD	Peritoneal dialysis
XGBoost	Extreme gradient boosting
RSF	Random survival forest
AUC	Area under the curve
C-index	Concordance index
CVD	Cardiovascular disease
ML	Machine learning
BMI	Body mass index
ESKD	End-stage kidney disease
CRP	C-reactive protein
hs-CRP	High-sensitivity C-reactive protein
LV	Left ventricular
LVEDd	Left ventricular end-diastolic diameter
LA	Left atrium
RV	Right ventricular
IVST	Interventricular septum thickness
LVPWT	Left ventricular posterior wall thickness
EF	Ejection fraction
PAP	Pulmonary arterial pressure
LVM	Left ventricular mass
LVMI	Left ventricular mass index
LVH	Left ventricular hypertrophy
CVC	Cardiac valve calcification
eGFR	Estimated glomerular filtration rate
tKt/V	Total weekly Kt/V
tCCr	Total weekly creatinine clearances
4hD/Pcr	Dialysate/plasma creatinine ratio at 4h
LASSO	Least Absolute Shrinkage and Selection Operator
AIC	Akaike Information Criterion
SHAP	SHapley Additive exPlanations
ROC	Receiver operating characteristic
TyG	Triglyceride-glucose
AdaBoost	Adaptive boosting
CatBoost	Categorical boosting

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03003-w.

Supplementary Material 1: Additional file 1. Table S1. Hyperparameters of the models.

Supplementary Material 2: Additional file 2. Table S2. Missing number (%) for variables.

Supplementary Material 3: Additional file 3. Table S3. Characteristics of the training and validation sets.

Supplementary Material 4: Additional file 4. Figure S1. LASSO regression analysis for CVE in patients undergoing PD.

Supplementary Material 5: Additional file 5. Table S4. Multivariate Cox regression analysis of CVE in patients undergoing PD.

Supplementary Material 6: Additional file 6. Figure S2. Calibration curves of the nomogram model predicted CVE-free probability of 1-, 3-, and 5-years in the training set and validation set.

Supplementary Material 7: Additional file 7. Figure S3. Calibration curves of the XGBoost model predicted CVE-free probability of 1-, 3-, and 5-years in the training set and validation set.

Supplementary Material 8: Additional file 8. Figure S4. Calibration curves of the random survival forest model predicted CVE-free probability of 1-, 3-, and 5-years in the training set and validation set.

Acknowledgments
We express our gratitude to all the researchers, patients, and their families for participating in this study.

Author contributions
Q.Y. designed the study, performed analysis, and drafted the manuscript. D.W. designed and directed the study. G.L., R.W., D.L., X.C., and J.C. collected data. All authors reviewed the manuscript.

Funding
This work was supported by the Research Funds of the Center for Big Data and Population Health of IHM (JKS2023005, JKS2022001), the National Natural Science Foundation of China (82370748), and the Natural Science Foundation of Anhui Province (2208085MH207).

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committee of the Second Affiliated Hospital of Anhui Medical University. All participants provided signed written informed consent.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 September 2024 / Accepted: 11 April 2025
Published online: 22 April 2025

References

1. Shah S, Weinhandl E, Gupta N, Leonard AC, Christianson AL, Thakar CV.Cardiovascular Outcomes in Patients on Home Hemodialysis and Peritoneal Dialysis. Kidney360. 2024;5(2):205–15.
2. Krediet RT, Balafa O. Cardiovascular risk in the peritoneal dialysis patient. Nat Rev Nephrol. 2010;6(8):451–60.
3. Sarnak MJ, Amann K, Bangalore S, Cavalcante JL, Charytan DM, Craig JC, et al. Chronic Kidney Disease and Coronary Artery Disease: JACC State-of-the-Art Review. J Am Coll Cardiol 2019;74:1823–38.
4. D'Agostino Sr RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008;117:743–53.
5. Su PF, Lin CK, Hung JY, Lee JS. The Proper Use and Reporting of Survival Analysis and Cox Regression. World Neurosurg. 2022;161:303–09.
6. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G.Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Sci Rep. 2021;11(1):6968.
7. Tseng PY, Chen YT, Wang CH, Chiu KM, Peng YS, Hsu SP, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Critical Care 2020;24:478.
8. Chen Y, Li G, Jiang W, Nie RC, Deng H, Chen Y, et al. Prognostic risk factor of major salivary gland carcinomas and survival prediction model based on random survival forests. Cancer Med 2023;12:10899–907.
9. Zhong X, Lin Y, Zhang W, Bi Q.Predicting diagnosis and survival of bone metastasis in breast cancer using machine learning. Sci Rep. 2023;13(1):18301.
10. Hou N, Li M, He L, Xie B, Wang L, Zhang R, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. J Transl Med 2020;18:462.
11. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg. 2015;102(3):148–58.
12. Liu G, Zhang SL, Liu PM, Yin GS, Tang JY, Ma DJ, et al. Biomarkers of endothelial dysfunction and risk of early organ damage: a comparison between patients with primary aldosteronism and essential hypertension. Zhonghua Xin Xue Xing Bing Xue Za Zhi 2012;40:640–44.
13. Crichton N.Cox proportional hazards model. J Clin Nurs. 2002;11(6):723.
14. Tibshirani R.The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385–95.
15. Tong L, Erdmann C, Daldalian M, Li J, Esposito T. Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. BMC Med Res Method. 2016;16:26.
16. Lim J, Jeon HG, Seo Y, Kim M, Moon JU, Cho SH. Survival Prediction Model for Patients with Hepatocellular Carcinoma and Extrahepatic Metastasis Based on XGBoost Algorithm. J Hepatocell Carcinoma. 2023;10:2251–63.
17. Taylor JM. Random Survival Forests. J Thorac Oncol. 2011;6(12):1974–75.
18. Xiao J, Mo M, Wang Z, Zhou C, Shen J, Yuan J, et al. The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: retrospective Cohort Study. JMIR Medical Informatics 2022;10:e33440.
19. Wu J, Li X, Zhang H, Lin L, Li M, Chen G, et al. Development and validation of a prediction model for all-cause mortality in maintenance dialysis patients: a multicenter retrospective cohort study. Renal Failure 2024;46:2322039.
20. Huang DD, Li YY, Qi XM, Wu YG.A nomogram predicts cardiovascular events in patients with peritoneal dialysis-associated peritonitis. Renal Failure. 2022;44(1):1558–67.
21. Hepburn KS, Lambert K, Mullan J, McAlister B, Lonergan M, Cheikh Hassan HI.Peritoneal dialysis-related peritonitis as a risk factor for cardiovascular events. Internal Med J. 2021;51(3):404–10.
22. Zimmermann S, Mathew A, Schöppe R, Mangova G, Biemann R, Surov A, et al. Fat tissue quantity, waist circumference or waist-to-hip ratio in patients with chronic kidney disease: a systematic review and meta-analysis. Obes Res Clin Pract 2024;18:81–87.
23. He S, Lei W, Li J, Yu K, Yu Y, Zhou L, et al. Relation of Platelet Parameters With Incident Cardiovascular Disease (The Dongfeng-Tongji Cohort Study). Am J Cardiol 2019;123:239–48.
24. Yu Z, Xiong J, Yang K, Huang Y, He T, Yu Y, et al. The association between platelet indices and cardiovascular events in chronic kidney disease patients without dialysis. Int Urol Nephrol 2021;53:961–71.
25. Peng F, Li Z, Yi C, Guo Q, Yang R, Long H, et al. Platelet index levels and cardiovascular mortality in incident peritoneal dialysis patients: a cohort study. Platelets 2017;28:576–84.
26. Lopez-Jaramillo P, Gomez-Arbelaez D, Martinez-Bello D, Abat MEM, Alhabib KF, Avezum Á, et al. Association of the triglyceride glucose index as a measure of insulin resistance with mortality and cardiovascular disease in populations from five continents (PURE study): a prospective cohort study. The Lancet Healthy Longevity 2023;4:e23–e33.
27. Liu X, Tan Z, Huang Y, Zhao H, Liu M, Yu P, et al. Relationship between the triglyceride-glucose index and risk of cardiovascular diseases and mortality in the general population: a systematic review and meta-analysis. Cardiovasc Diabetol 2022;21:124.
28. Colladant M, Chabannes M, Crepin T, Bamoulid J, Courivaud C, Ducloux D.Triglyceride-Glucose Index and Cardiovascular Events in Kidney Transplant Recipients. Kidney Int Rep. 2023;8(11):2307–14.
29. Chinali M, de Simone G, Wachtell K, Gerdts E, Gardin JM, Boman K, et al. Left atrial systolic force in hypertensive patients with left ventricular hypertrophy: the LIFE study. J Hypertens 2008;26:1472–76.
30. Sezer S, Tutal E, Arat Z, Akçay A, Celik H, Ozdemir FN, et al. Peritoneal transport status influence on atherosclerosis/inflammation in CAPD patients. J Ren. Nutr: The Official J Council Renal Nutr National Kidney Foundation 2005;15:427–34.
31. Castel-Feced S, Malo S, Aguilar-Palacio I, Feja-Solana C, Casasnovas JA, Maldonado L, et al. Influence of cardiovascular risk factors and treatment exposure on cardiovascular event incidence: assessment using machine learning algorithms. PLoS One 2023;18:e0293759.

32. Tapak L, Sheikh V, Jenabi E, Khazaei S.Predictors of mortality among hemodialysis patients in Hamadan province using random survival forests. J Prev Med Hyg. 2020;61(3):E482–e8.
33. Backhaus SJ, Schulz A, Lange T, Schmidt-Schweda LS, Evertz R, Kowallick J, et al. Real-time cardiovascular magnetic resonance imaging for non-invasive characterisation of heart failure with preserved ejection fraction: final outcomes of the HFpEF stress trial. Clin Res Cardiol 2024;113:496–508.
34. Habibi M, Zareian M, Ambale Venkatesh B, Samiei S, Imai M, Wu C, et al. Left Atrial Mechanical Function and Incident Ischemic Cerebrovascular Events Independent of AF: insights From the MESA Study. JACC Cardiovasc Imaging 2019;12:2417–27.
35. Alonso A, Kraus J, Ebert A, Nikolayenko V, Kruska M, Sandikci V, et al. Left atrial area index provides the best prediction of atrial fibrillation in ischemic stroke patients: results from the LAETITIA observational study. Front Neurol. 2023;14:1237550.
36. Xu L, Cao F, Wang L, Liu W, Gao M, Zhang L, et al. Machine learning model and nomogram to predict the risk of heart failure hospitalization in peritoneal dialysis patients. Renal Failure 2024;46:2324071.
37. Yang J, Wan J, Feng L, Hou S, Yv K, Xu L, et al. Machine learning algorithms for the prediction of adverse prognosis in patients undergoing peritoneal dialysis. BMC Med Inf Decis Making 2024;24:8.
38. Xu Z, Xu X, Zhu X, Niu K, Dong J, He Z. Attention-Based Deep Learning Model for Prediction of Major Adverse Cardiovascular Events in Peritoneal Dialysis Patients. IEEE J Biomed Health Inform. 2024;28(2):1101–09.

Publisher's Note