Interpretable machine learning for predicting chronic kidney disease progression risk

Jin-Xin Zheng[1,#] 🆔, Xin Li[1,#], Jiang Zhu[2] 🆔, Shi-Yang Guan[3], Shun-Xian Zhang[4,5] and Wei-Ming Wang[1]

Abstract

Objective: Chronic kidney disease (CKD) poses a major global health burden. Early CKD risk prediction enables timely interventions, but conventional models have limited accuracy. Machine learning (ML) enhances prediction, but interpretability is needed to support clinical usage with both in diagnostic and decision-making.

Methods: A cohort of 491 patients with clinical data was collected for this study. The dataset was randomly split into an 80% training set and a 20% testing set. To achieve the first objective, we developed four ML algorithms (logistic regression, random forests, neural networks, and eXtreme Gradient Boosting (XGBoost)) to classify patients into two classes—those who progressed to CKD stages 3–5 during follow-up (positive class) and those who did not (negative class). For the classification task, the area under the receiver operating characteristic curve (AUC-ROC) was used to evaluate model performance in discriminating between the two classes. For survival analysis, Cox proportional hazards regression (COX) and random survival forests (RSFs) were employed to predict CKD progression, and the concordance index (C-index) and integrated Brier score were used for model evaluation. Furthermore, variable importance, partial dependence plots, and restrict cubic splines were used to interpret the models' results.

Results: XGBOOST demonstrated the best predictive performance for CKD progression in the classification task, with an AUC-ROC of 0.867 (95% confidence interval (CI): 0.728–0.100), outperforming the other ML algorithms. In survival analysis, RSF showed slightly better discrimination and calibration on the test set compared to COX, indicating better generalization to new data. Variable importance analysis identified estimated glomerular filtration rate, age, and creatinine as the most important predictors for CKD survival analysis. Further analysis revealed non-linear associations between age and CKD progression, suggesting higher risks in patients aged 52–55 and 65–66 years. The association between cholesterol levels and CKD progression was also non-linear, with lower risks observed when cholesterol levels were in the range of 5.8–6.4 mmol/L.

Conclusions: Our study demonstrated the effectiveness of interpretable ML models for predicting CKD progression. The comparison between COX and RSF highlighted the advantages of ML in survival analysis, particularly in handling non-linearity and high-dimensional data. By leveraging interpretable ML for unraveling risk factor relationships, contrasting predictive techniques, and exposing non-linear associations, this study significantly advances CKD risk prediction to enable enhanced clinical decision-making.

Keywords

Chronic kidney disease, machine learning, survival analysis, interpretable models, random survival forests

Submission date: 16 August 2023; Acceptance date: 15 December 2023

[1]Department of Nephrology, Ruijin Hospital, Institute of Nephrology, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2]Liver Transplantation Center, West China Hospital, Sichuan University, Chengdu, China
[3]Department of Statistics, Second Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China
[4]School of Global Health, Chinese Center for Tropical Diseases Research – Shanghai Jiao Tong University School of Medicine, Shanghai, China
[5]Clinical Research Center, Longhua Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China

[#]Jin-Xin Zheng and Xin Li contributed equally to this work.

Corresponding author:
Shun-Xian Zhang, School of Global Health, Chinese Center for Tropical Diseases Research – Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China; Clinical Research Center, Longhua Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai 200032, China.
Email: zhangshunxian110@163.com

Wei-Ming Wang, Department of Nephrology, Ruijin Hospital, Institute of Nephrology, Shanghai Jiao Tong University School of Medicine, Shanghai, China
Email: wweiming01@126.com

Background

Chronic kidney disease (CKD) poses a significant global burden, with an estimated annual death toll of over 5 million by the year 2040.[1] This prevalence is mainly attributed to the fact that CKD often remains asymptomatic, making it challenging to detect and diagnose in a timely manner.[2] According to the World Health Organization (WHO), in 2019, kidney diseases account for 1.7 million years of life lived with disability, equivalent to 168 YLDs per 100,000 population, and its prevalence continues to surge, particularly among elderly individuals and those afflicted by coexisting conditions such as diabetes and hypertension.[3]

The timely and accurate prediction of incident CKD risk holds profound clinical significance, as it empowers healthcare practitioners to implement effective preventive measures and ensure optimal patient outcomes.[4] Early identification of individuals at heightened risk of developing CKD enables targeted interventions, aimed at arresting disease progression, averting complications, and easing the burden on healthcare systems.[5] Moreover, proactive management of high-risk patients may lead to enhanced therapeutic efficacy and potential cost reductions associated with the management of advanced CKD cases.[6]

Traditional CKD risk prediction models have traditionally relied on clinical factors such as age, gender, blood pressure (BP), and laboratory results, which, although informative, exhibit limitations in predictive accuracy and risk assessment comprehensiveness.[7,8] These conventional models may inadvertently overlook critical risk contributors that could significantly impact disease progression and patient outcomes.[9] In response to these challenges, recent advancements in machine learning (ML) techniques have demonstrated immense promise across various medical domains, particularly in the field of disease risk prediction, including CKD and acute kidney injury.[10–12] ML models possess the unique ability to discern intricate patterns and associations hidden within vast and diverse datasets, allowing for a more comprehensive understanding of CKD risk factors, and enhancing the precision of risk assessments.[11] By leveraging ML's computational power, we can potentially improve the predictive capabilities of CKD risk models, and with high accurate prediction, offering a more holistic and personalized approach to patient care and early intervention strategies.[13,14]

However, the predominant focus on predictive accuracy has led to complex black-box models lacking in interpretability, and the growing adoption of ML models has also raised concerns regarding their interpretability and transparency, particularly in healthcare settings.[15] This poses barriers to clinical adoption, as trust necessitates some level of explainability. To address this challenge, an emerging paradigm of interpretable ML strives to balance performance with transparency.[16] A myriad of techniques facilitates model interpretation, from visualizing decision rules and feature importance to approximating complex models with simpler surrogates.[17] When applied judiciously in clinical studies, interpretable ML can elucidate predictive patterns and relationships within patient data.[18] For instance, partial dependence plots (PDPs) can reveal non-linear associations between risk factors and outcomes. Such transparency uncovers actionable insights to guide clinical decision-making while avoiding reliance on inscrutable models.[19] Interpretable ML represents a nascent but active field with immense potential to bridge the gap between predictive analytics and clinical implementation. These interpretable models offer valuable insights into the underlying factors contributing to risk assessments, thereby empowering clinicians to make informed decisions based on actionable information. Interpretable ML models for CKD risk prediction remain an open challenge needing further research. Progress in this area can lead to clinically useful tools that improve CKD management.

This study makes several novel contributions through the in-depth investigation and evaluation of interpretable ML models for predicting CKD progression risk. We compare multiple algorithms for CKD classification, demonstrating eXtreme Gradient Boosting (XGBoost)'s superior performance. For predicting CKD progression, we evaluate and contrast Cox regression and random survival forests (RSFs), showcasing ML's advantages for survival analysis. By utilizing interpretable models, we gain insights into non-linear relationships between variables like age, cholesterol, and CKD risk. External validation of predictive models on new data is vital for further improving generalizability. Our research aims to advance CKD risk stratification through accurate and interpretable models, enabling tailored interventions and enhanced CKD management. The deployment of such models holds promise for optimizing healthcare resources and improving patient outcomes.

Methods

Study design

This hospital-based cross-sectional study aims to investigate the application of interpretable ML techniques for predicting the progression of CKD to stages 3–5 in a well-defined patient population. The study design was carefully selected to assess the feasibility and use interpretable ML models to assess CKD risk and compare their performance in survival analysis.

Data source and study population

Data for this study were obtained from the electronic medical records of 491 patients collected at Tawam Hospital, located in Al-Ain city, Abu Dhabi, United Arab Emirates, the full dataset covers patients from 1 January 2008 to 31 December 2008, with up to 10 years of follow-up based on the 2018 publication.[20] The hospital's

electronic medical record system provided access to a rich and comprehensive dataset, which was made available for research purposes. The study population consisted of 491 consecutive patients who were either diagnosed with cardiovascular disease (CVD) or presented with a high CVD risk. CVD patients were included in this study due to the close association between CVD and CKD. Previous research indicates CVD is both a risk factor for and complication of CKD, underscoring the value of studying this patient population for enhanced understanding of CKD progression. Each patient's electronic medical record included a collection of 21 essential clinical variables, representing a wide range of laboratory test results, examination findings, and pertinent medical history. The dataset was de-identified to ensure patient confidentiality, and the study protocol received ethical approval from both the Tawam Hospital and the United Arab Emirates University Research and Ethics Board.[20]

Predictors and outcome

The study cohort consisted of 491 consecutive patients with available medical records, each encompassing a chart of 13 pertinent clinical variables, comprising laboratory test results, examination findings, and historical medical data (Table 1). We conducted an extensive evaluation of ML models to assess their feasibility and accuracy in identifying CKD risk, employing a dataset comprising diverse predictor variables. The predictor variables comprised essential clinical and demographic factors, including "Sex," "Age Baseline," "History Diabetes," "History CHD" (coronary heart disease), "History Vascular," "History Smoking," "History HTN" (hypertension), "History DLD" (dyslipidemia), "History Obesity," "DLD meds" (medication for dyslipidemia), "DM meds" (medication for diabetes), "HTN meds" (medication for hypertension), "ACEIARB" (angiotensin-converting enzyme inhibitors-ACEI or angiotensin II receptor blockers-ARB), "Cholesterol Baseline," "Creatinine Baseline," "eGFR Baseline" (estimated glomerular filtration rate), "sBP Baseline" (systolic blood pressure), "dBP Baseline" (diastolic blood pressure), "BMI Baseline" (body mass index), and "Time To Event " (time to the occurrence of the event).

The primary outcome of interest, "EventCKD35," pertains to the development of CKD within the follow-up period. The staging of CKD is predominantly based on the estimated glomerular filtration rate (eGFR), which quantifies the kidney's ability to effectively filter waste from the blood. The stages are classified as Stage 1 CKD, Stage 2 CKD, Stage 3 CKD, Stage 4 CKD, and Stage 5 CKD.[21] Stage 3–5 represents the severe stage of CKD, indicating kidney failure, and patients are likely to experience more severe symptoms and complication, necessitate kidney replacement therapy through dialysis or transplantation.

In our study, the primary objective is to predict the risk of progression of CKD to stages 3 to 5, and the survival risk to CKD progression based on the "EventCKD35" outcome. By leveraging a diverse set of predictor variables and employing advanced ML methodologies, our research endeavors to contribute novel insights into the accurate identification and risk stratification of CKD stages, which could profoundly impact clinical decision-making and patient management strategies in the context of CKD.

Model development

Before develop model to prediction, the dataset was checked for any duplicated or incomplete entries, which were removed. For missing values, median or mean imputation was used for continuous variables and mode imputation for categorical variables given the low missingness rate (∼1–3%). No explicit noise removal techniques were applied since the data originated from standard clinical measurements.

The objectives of this study are twofold: firstly, to discriminate the outcomes of CKD stages 3 to 5 (EventCKD35) based on clinical data, and secondly, to predict the risk of patients progressing to CKD stages 3 to 5, utilizing the available follow-up data.

To achieve first objective, we randomly allocated patients to either a development dataset or a test dataset using an 8 : 2 sample ratio, where the development dataset comprised both training and validation data. Prior to model training, all continuous variables were normalized post-splitting to prevent information leakage and ensure fair application across the training, validation, and test datasets. For the classification task, we considered four different ML algorithms: linear regression (LM), random forests (RF), neural networks (NNET), and XGBOOST. To reduce the impact of data partitioning variability, each iteration involved random sampling with different seeds from the observational data. After obtaining the final models, we thoroughly evaluated, compared, and reported their predictive performance on the test dataset. Employing 10-fold cross-validation, we computed average values from 10 different partitions to minimize sensitivity to random data splits and yield more robust performance estimates. For the ML classification models, all clinical variables, except time-related factors, were used as input, while the dependent variable was set to EventCKD35 (CKD stages 3 to 5).

Secondly, given the recorded follow-up time for patients progressing to CKD3–5, survival analysis models were also employed to predict the risk of progression. For the survival analysis task, we utilized two algorithms: Cox regression (Cox) and RSF. The Cox regression model is well-suited for survival analysis due to its ability to handle censored data efficiently and accommodate time-varying covariates.[22] Its flexible assumption of proportional hazards allows for a wide range of applications and robustness to

Table 1. Baseline information of study patients.

Variables	Clinical implication	Unit
Sex	Gender of patient (Male-1 and Female-0)	
Age Baseline	Age of the patient	year
BMI Baseline	Body mass index of the patient	kg/m^2
ACETARB	If the patient has taken ACEI or ARB (Yes-1 OR No-0)	
History CHD	Patient history of coronary heart disease (Yes-1 OR No-0)	
History Diabetes	Patient history of diabetes (Yes-1 OR No-0)	
History DLD	Patient history of dyslipidemia (Yes-1 OR No-0)	
History HTN	Patient history of hypertension (Yes-1 OR No-0)	
History Obesity	Patient history of obesity (Yes-1 OR No-0)	
History Smoking	Patient history of smoking (Yes-1 OR No-0)	
History Vascular	Patient history of vascular diseases (Yes-1 OR No-0)	
DLD meds	If the patient has taken dyslipidemia medications (Yes-1 OR No-0)	
DM meds	If the patient has taken diabetes medications (Yes-1 OR No-0)	
HTN meds	If the patient has taken hypertension medications (Yes-1 OR No-0)	
Cholesterol Baseline	Level of cholesterol	mmol/L
Creatinine Baseline	Level of creatinine in the blood	mol/L
dBP Baseline	Diastolic blood pressure	mmHg
sBP Baseline	Systolic blood pressure	mmHg
eGFR Baseline	Estimated glomerular filtration rate	ml/min/1.73 m^2
Time To Event months	Time from follow-up start to severe CKD event or last visit	month
EventCKD35	If the patient had moderate-extreme CKD stages 3–5 (Yes-1 OR No-0)	

Note: CKD: chronic kidney disease; CHD: coronary heart disease; DM: diabetes mellitus; HTN: hypertension; ACEI: angiotensin-converting enzyme inhibitors; ARB: angiotensin II receptor blockers; BMI: body mass index; eGFR: estimated glomerular filtration rate; SBP: systolic blood pressure; DBP: diastolic blood pressure.

multicollinearity, providing reliable predictions and personalized risk assessments.[23] However, the Cox regression model relies on the assumption of proportional hazards, which may not always hold true and could lead to biased results. Moreover, it does not directly estimate survival probabilities and lacks information about the shape of the survival curve, making it less suitable for predicting exact survival times. On the other hand, RSFs models offer advantages in handling survival data. As an extension of traditional RFs to address survival data, RSF excels at handling high-dimensional datasets and capturing complex interactions among predictors.[24] Unlike traditional parametric survival models, RSF does not impose strong assumptions on the underlying distribution of survival times, enhancing its flexibility and applicability to various data types. RSFs provide a powerful and versatile tool for researchers aiming to analyze survival data in a non-parametric and data-driven manner.[25]

For survival analysis, we first conducted univariate Cox regression analysis on the entire dataset to assess the

associations between each clinical variable and survival outcomes. Variables with a significant correlation ($p < 0.05$) with survival were considered potential predictive factors and were subsequently included in the multivariate Cox proportional hazards (Coxph) model. Next, to compare the predictive accuracy of the Cox regression model with RSF, we incorporated the variables significantly correlated with survival ($p < 0.05$) and randomly allocated patients to the development and test datasets using an 8:2 sample ratio. To evaluate the predictive performance of the models, we employed cross-validation techniques and assessed the model performance on the test dataset. Therefore, this study constituted a regression problem based on time-to-event (censored) data.

Model evaluation and interpretability

For the classification task, the model's performance on the testing set was assessed by calculating the area under the receiver operating characteristic curve (AUC-ROC) and its corresponding 95% confidence interval (CI).[26] The ROC curve is a graphical representation of the model's ability to discriminate between different classes by varying the classification threshold. It is constructed by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) for different threshold values. The AUC-ROC value ranges from 0 to 1, where a value of 0.5 indicates random performance, and a value of 1 represents perfect discrimination between the classes.

In the context of survival analysis models, we employed the C-index and integrated Brier score (IBS) to assess model predictive performance.[27] The C-index measures the proportion of correctly ordered pairs of predicted and actual survival times among all possible subject pairs. It ranges from 0 to 1, where 0.5 signifies random predictions, and 1 indicates perfect concordance between predicted and observed survival times. A C-index of 0.5 implies that the model performs no better than chance in predicting survival outcomes, while a value greater than 0.5 indicates some predictive ability, with a C-index approaching 1 representing strong predictive performance. On the other hand, the IBS combines calibration and discrimination aspects and quantifies the overall model calibration by comparing predicted survival probabilities with actual survival outcomes at various time points. Lower IBS values indicate better model calibration and higher prediction ability compared to models with higher IBS values.

Assessing variable importance is crucial in evaluating the significance of input features in ML models. The caret package for R utilizes distinct techniques for each algorithm to calculate variable importance.[28] For RF models, the mean decrease in accuracy method is employed, wherein each feature's contribution is measured by randomly permuting its values and observing the subsequent decrease in model accuracy.[29,30] The average decrease across all trees in the forest determines the final importance score for each feature. Higher importance scores indicate more pivotal features considered by the ML model, providing valuable insights into feature significance, and aiding in model interpretability.

To interpret the impact of individual features on predicted outcomes, we employed the PDP technique in ML model interpretation.[19,30] PDP provides a comprehensive understanding of the relationship between the target variable and a specific input feature while holding other features constant at predetermined values. By plotting the selected feature's values against the corresponding predictions, PDP illustrates the partial effect of the feature on the target variable, revealing insights into its influence and potential interactions with other features. PDPs serve as valuable tools for model transparency and explainability, enhancing model understanding and supporting decision-making in complex ML applications.

In time-to-event analysis, dealing with covariates exhibiting challenging functional forms, time-dependent effects, and interaction structures necessitates the implementation of restrict cubic splines (RCS).[31,32] RCS captures non-linear associations between continuous predictors and binary outcomes by replacing continuous predictors with piecewise cubic functions, smoothly joined at specific knots. The "restrict" attribute limits the spline's linearity beyond boundary knots, preventing extreme extrapolations outside the observed data range. The number and positions of knots are often determined by quantiles of the predictor variable's distribution. RCS models were employed to evaluate potential non-linear associations between hazard ratios (HRs) and covariates, accommodating the complex relationship patterns in survival analysis.

Statistical analysis

Continuous variables underwent the Kolmogorov–Smirnov test to assess the normality of their distribution. Normally distributed data were compared using the t-test, while non-normally distributed data were compared using the Mann–Whitney U test. For binary classification tasks, we employed the caret package in R, employing various ML algorithms, including linear regression (glm), random forests (randomForest), neural networks (nnet), and XGBoost (XGBOOST). Survival data were visualized using Kaplan–Meier (KM) curves, with median survival estimates, and further analyzed using Cox proportional hazards regression (survival). Additionally, RFs survival analysis was conducted using the randomForestSRC package. The R Statistics software version 4.2.2 (R Foundation for Statistical Computing, Vienna, Austria) was used for statistical analyses.

Results

Baseline clinical characteristics

A total of 491 patients' clinical data were collected for this study. Among them, 241 (49.1%) were female, and 250

Table 2. Baseline characteristics of participants with and without follow-up chronic kidney disease (CKD) progression.

Variables	All data (N = 491)	Event CKD3–5		*p*-value
		Yes (N = 435)	No (N = 56)	
Sex				0.047
Female	241 (49.1%)	221 (50.8%)	20 (35.7%)	
Male	250 (50.9%)	214 (49.2%)	36 (64.3%)	
Age	53.2 (13.8)	52.0 (13.9)	62.2 (9.51)	<0.001
History Diabetes				<0.001
0	276 (56.2%)	267 (61.4%)	9 (16.1%)	
1	215 (43.8%)	168 (38.6%)	47 (83.9%)	
History CHD				<0.001
0	446 (90.8%)	407 (93.6%)	39 (69.6%)	
1	45 (9.16%)	28 (6.44%)	17 (30.4%)	
History Vascular				0.036
0	462 (94.1%)	413 (94.9%)	49 (87.5%)	
1	29 (5.91%)	22 (5.06%)	7 (12.5%)	
History Smoking				0.051
0	416 (84.7%)	374 (86.0%)	42 (75.0%)	
1	75 (15.3%)	61 (14.0%)	14 (25.0%)	
History HTN				<0.001
0	156 (31.8%)	151 (34.7%)	5 (8.93%)	
1	335 (68.2%)	284 (65.3%)	51 (91.1%)	
History DLD				0.002
0	174 (35.4%)	165 (37.9%)	9 (16.1%)	
1	317 (64.6%)	270 (62.1%)	47 (83.9%)	
History Obesity				0.529
0	243 (49.5%)	218 (50.1%)	25 (44.6%)	
1	248 (50.5%)	217 (49.9%)	31 (55.4%)	

Table 2. Continued.

| Variables | All data (N = 491) | Event CKD3−5 | | *p*-value |
		Yes (N = 435)	No (N = 56)	
DLD meds				<0.001
0	220 (44.8%)	208 (47.8%)	12 (21.4%)	
1	271 (55.2%)	227 (52.2%)	44 (78.6%)	
DM meds				<0.001
0	330 (67.2%)	315 (72.4%)	15 (26.8%)	
1	161 (32.8%)	120 (27.6%)	41 (73.2%)	
HTN meds	0.62 (0.49)	0.59 (0.49)	0.84 (0.37)	<0.001
ACEIARB				<0.001
0	272 (55.4%)	259 (59.5%)	13 (23.2%)	
1	219 (44.6%)	176 (40.5%)	43 (76.8%)	
Cholesterol	4.98 (1.10)	5.03 (1.09)	4.54 (1.09)	0.002
Creatinine	67.9 (17.9)	65.9 (17.1)	82.8 (17.1)	<0.001
eGFR	98.1 (18.5)	100 (17.8)	79.6 (12.5)	<0.001
sBP	131 (15.7)	131 (15.3)	137 (17.7)	0.017
dBP	76.9 (10.7)	77.1 (10.5)	74.8 (11.8)	0.154
BMI	30.2 (6.24)	30.2 (6.27)	30.0 (6.06)	0.829
Time To Event	83.1 (26.0)	87.3 (22.1)	50.5 (31.2)	<0.001

Notes: categorical variables are coded as No for 0 and Yes for 1. In addition, categorical variables are expressed in terms of frequencies and percentages (N(proportion)), and continuous variables are expressed in terms of means and standard deviations (Mean(sd)). CKD: chronic kidney disease; CHD: coronary heart disease; DM: diabetes mellitus; HTN: hypertension; ACEI: angiotensin-converting enzyme inhibitors; ARB: angiotensin II receptor blockers; BMI: body mass index; eGFR: estimated glomerular filtration rate; SBP: systolic blood pressure; DBP: diastolic blood pressure.

(50.9%) were male, with an average age of 53.2 years. Out of these patients, 56 (11.4%) progressed to CKD stages 3 to 5. Common comorbidities included hypertension, diabetes, and coronary heart disease (CHD). Specifically, 215 (43.7%) patients had a history of CHD, and 75 (15.2%) had a smoking history. Detailed information is presented in Table 2. Notably, age showed a significant difference ($p < 0.001$), with patients who progressed to CKD stages 3 to 5 having a higher mean age (62.2 years) compared to those who did not (52.0 years). Other variables with notable differences included History Diabetes ($p < 0.001$), History CHD ($p < 0.001$), History Vascular ($p = 0.036$), History Smoking ($p = 0.051$), History HTN ($p < 0.001$), History DLD ($p = 0.002$), DLD meds ($p < 0.001$), DM meds ($p < 0.001$), ACEIARB ($p < 0.001$), Cholesterol ($p = 0.002$), Creatinine ($p < 0.001$), eGFR ($p < 0.001$), sBP ($p = 0.017$), and Time To Event months ($p < 0.001$).

Machine learning models for CKD classification

We employed six ML models to predict CKD outcomes based on patients' clinical characteristics and medical history. Notably, the XGBoost model demonstrated the best performance in predicting CKD progression, with an AUC of 0.867 (95% CI: 0.728–0.100) on the testing set. This outperformed the RF model, which achieved an AUC of 0.834 (95% CI: 0.676–0.983), as well as the LM and NNET

models with AUCs of 0.787 (95% CI: 0.621–0.952) and 0.5 (95% CI: 0.318–0.682), respectively (Figure 1). These results indicate that the XGBoost model exhibits higher accuracy and reliability in predicting CKD progression.

### Variable importance and interpretability in XGBOOST

Furthermore, by comparing the variable importance across the different models, we identified the most significant clinical features contributing to the prediction of CKD outcomes. Among these, three key clinical variables were identified as the most influential in CKD progression: eGFR, age, and creatinine (Figure 2). These findings offer valuable insights into the pivotal clinical factors influencing CKD progression and contribute to enhanced understanding and management of the disease.

Subsequently, we utilized interpretable ML models to delve deeper into the relationships between the variables,

namely eGFR, age, and creatinine, and their marginal effects on predicting CKD outcomes. The results revealed that age exhibited a non-linear relationship with CKD progression, with patients aged 52–55 and 65–66 years showing a higher risk of experiencing CKD progression among our study's population. As cholesterol levels increased, the risk of CKD progression decreased. However, when cholesterol levels fell within the range of 5.8–6.4 mmol/L, the likelihood of CKD progression was less than 5%. This implies that higher cholesterol levels may be associated with a lower risk of CKD progression. Nonetheless, it is crucial to note that when cholesterol levels are within a specific range, the risk of CKD progression might increase (Figure 3).

Machine learning models for CKD survival analysis

In our study, we conducted a univariate screening, and factors with $p < 0.05$ were included in our COX model for predicting
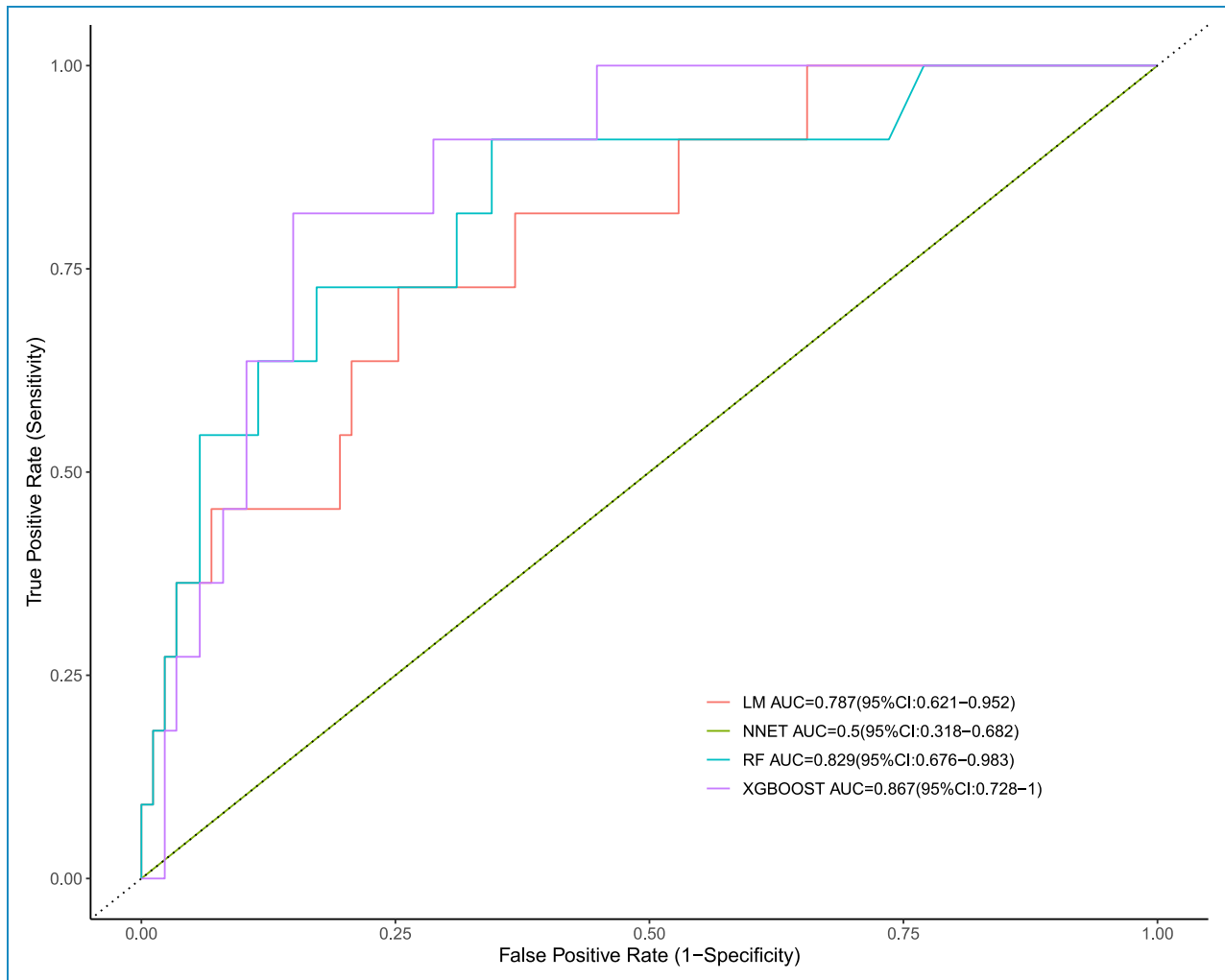
Figure 1. Area under the receiver operating characteristic curve (AUC-ROC) with 95% CI in the training and testing sets for the different trained models in classification of CKD. LM: logistic regression, RF: random forests, NNET: neural networks, and XGBOOST: eXtreme Gradient Boosting. CKD: chronic kidney disease; CI: confidence interval.
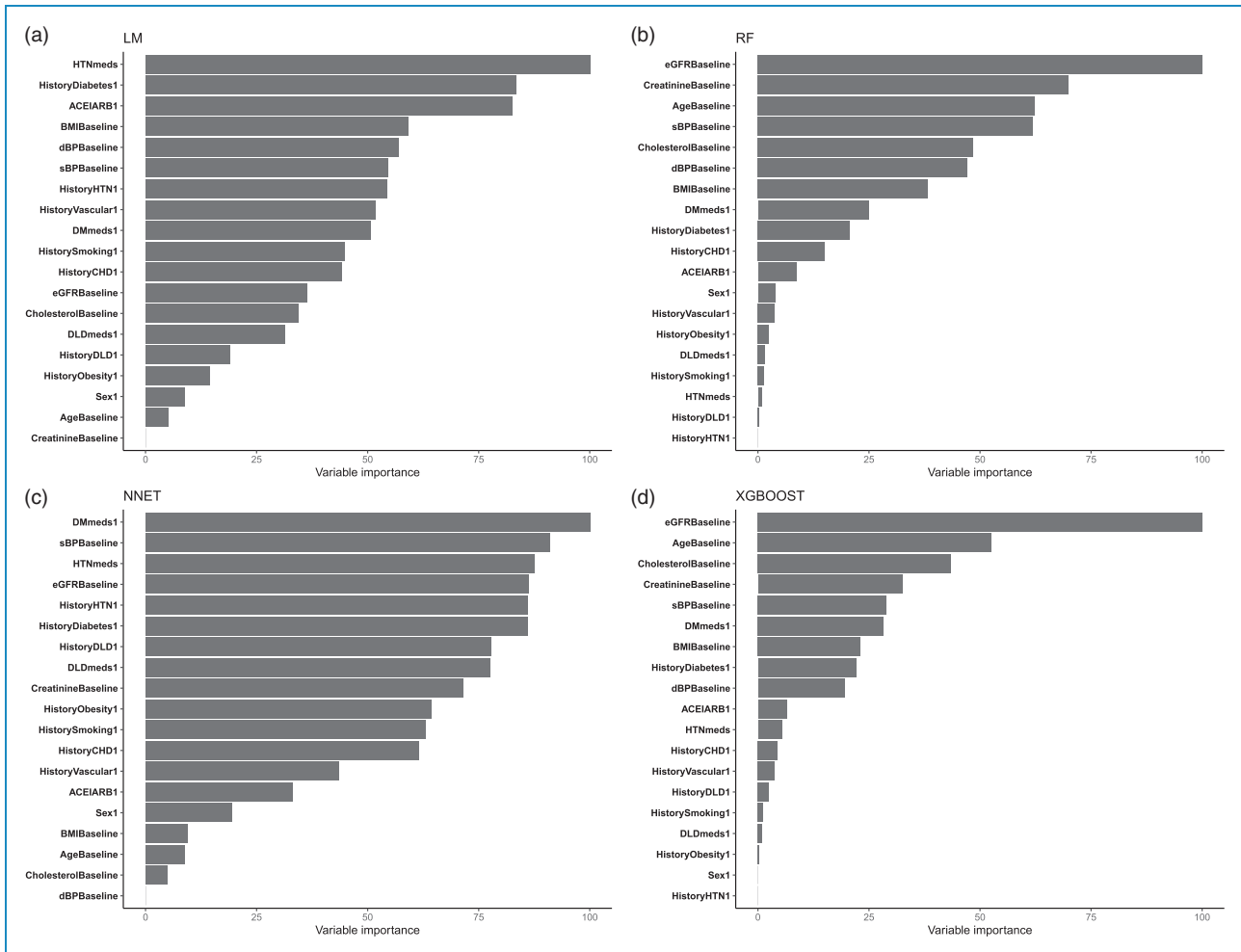
Figure 2. Relative importance ranking of each input variable for prediction of CKD in the machine learning algorithms. (a) Logistic regression. (b) Random forest. (c) Neural network. (d) Extreme gradient boosting. CHD: coronary heart disease; DM: diabetes mellitus; HTN: hypertension; ACEI: angiotensin-converting enzyme inhibitors; ARB: angiotensin II receptor blockers; BMI: body mass index; eGFR: estimated glomerular filtration rate; SBP: systolic blood pressure; DBP: diastolic blood pressure; CKD: chronic kidney disease.

CKD progression. Notably, the variables "History Obesity," "dBP Baseline," and "BMI Baseline" demonstrated no statistically significant association in the univariate COX regression analysis. Subsequently, we incorporated the remaining 16 factors into the COX model, and the Schoenfeld residual test ($p > 0.05$) confirmed that the model satisfied the proportional hazards assumption. Among the variables considered, "DLD meds" exhibited a HR of 3.06 (95% CI: 1.158–8.1), indicating that it was a significant risk factor for CKD progression. Similarly, "History Smoking" demonstrated a HR of 2.4 (95% CI: 1.125–5.1), indicating its significance as a risk factor in CKD progression (Table 3).

Model performance with COX and RSF

To facilitate a comparison with the RF model, we excluded the variables "History Obesity," "dBP Baseline," and "BMI Baseline" and split the dataset into 80% for training and 20% for testing. Upon conducting the model evaluation,

the Cox model demonstrated a C-index of 0.891 on the training set and 0.844 on the testing set. Similarly, the RSF model achieved a C-index of 0.849 on the training set and 0.854 on the testing set. Both models displayed high C-index values on the training set, with RSF marginally outperforming Cox, and this trend persisted on the testing set, with RSF showing slightly better performance (Figure 4). To further assess the models' calibration, we employed the Brier score as a metric. On the training set, the Cox model exhibited a lower Brier score, indicating better calibration compared to RSF. However, on the test set, RSF demonstrated a slightly lower Brier score, signifying better generalization in terms of calibration.

Interpretability of variables

We utilized the RSFs model to identify important predictors for CKD progression. Based on the variable importance results obtained from the RSF model, we proceeded to
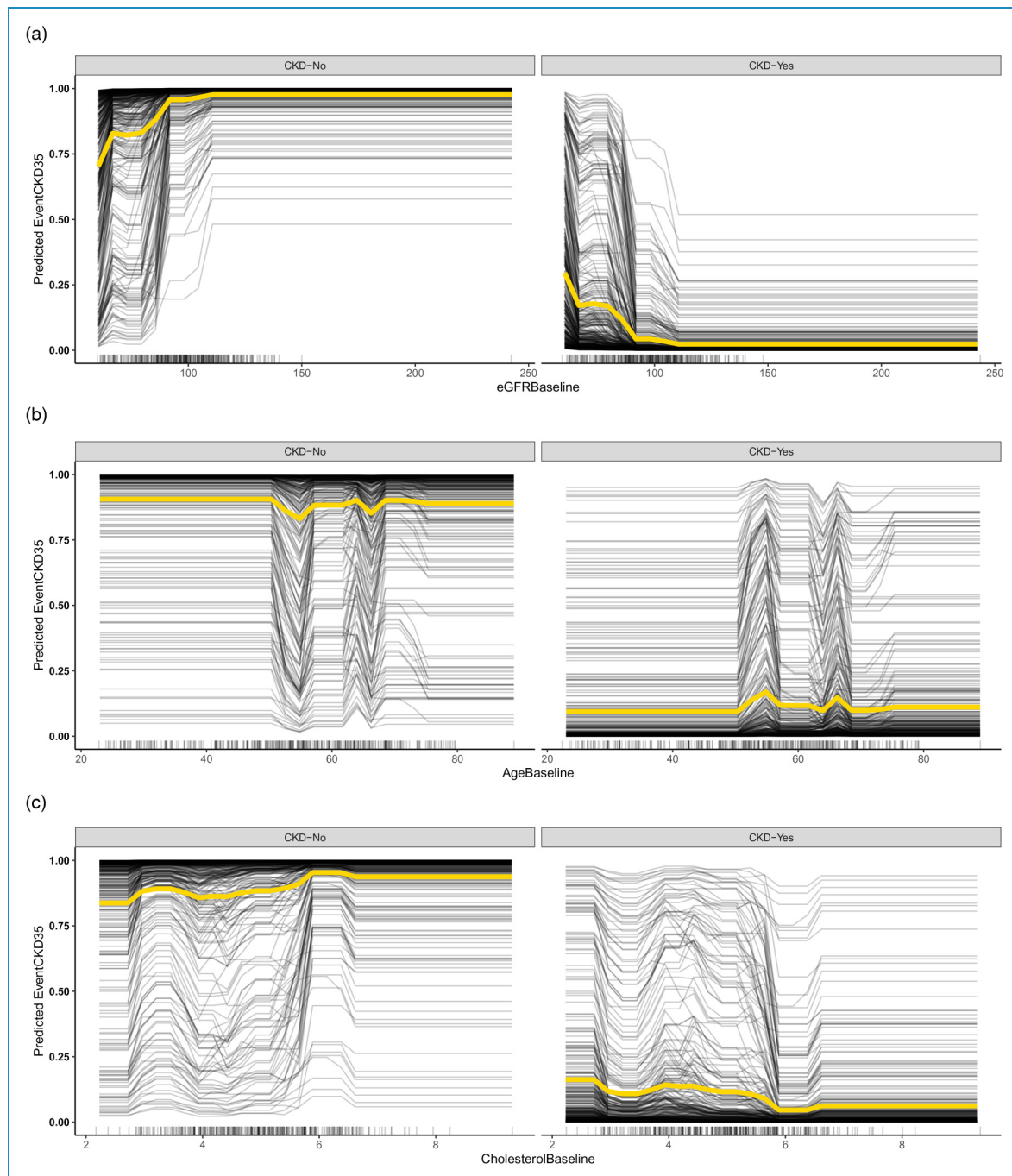
Figure 3. Dependence plots of eGFR, age, and creatinine, and their marginal effects on predicting CKD outcomes based on XGBOOST model, black line displays one line per instance that shows how the individual prediction changes with feature, and yellow line shows the main effect of the feature and ignores possible feature interactions. CKD: chronic kidney disease.

investigate the relationship between CKD progression and three key factors: eGFR Baseline, Creatinine Baseline, and Age Baseline (Figure 5a). To explore potential non-linear associations, we employed the RCS curves. The application of RCS allowed us to flexibly estimate HRs across the range of baseline covariates, thereby assessing non-linear relationships in survival data. Specifically, we

plotted the HR for CKD progression against baseline age, modeled using the RCS approach (Figure 5b, c, and d). Result shows the HR for CKD progression by baseline age, the overall *p*-value and non-linear *p*-value suggest the association is statistically significant, and there is evidence of non-linearity. Similarly, for baseline creatinine, we observed a statistically significant overall association

Table 3. Univariable and multivariable hazards ratios (HRs) and 95% confidence intervals (95% CI) of predictors for developing CKD stages 3–5 in COX model.

Variables	Univariable HR (95%CI, *p*-value)	Multivariable HR (95%CI, *p*-value)
Sex (Male)	1.90 (1.09–3.28, *p* = 0.023)	0.46 (0.03–7.85, *p* = 0.592)
Age Baseline	1.06 (1.03–1.08, *p* < 0.001)	1.02 (0.93–1.12, *p* = 0.703)
History Diabetes (Yes)	7.08 (3.47–14.47, *p* < 0.001)	1.65 (0.52–5.25, *p* = 0.395)
History CHD (Yes)	5.42 (3.06–9.61, *p* < 0.001)	1.42 (0.69–2.92, *p* = 0.338)
History Vascular (Yes)	2.53 (1.14–5.58, *p* = 0.022)	0.70 (0.30–1.62, *p* = 0.406)
History Smoking (Yes)	2.18 (1.19–4.00, *p* = 0.012)	2.40 (1.12–5.12, *p* = 0.024)
History HTN (Yes)	4.63 (1.85–11.61, *p* = 0.001)	3.22 (0.76–13.54, *p* = 0.111)
History DLD (Yes)	2.82 (1.38–5.77, *p* = 0.004)	1.67 (0.42–6.64, *p* = 0.463)
History Obesity (Yes)	1.15 (0.68–1.96, *p* = 0.601)	
DLD meds (Yes)	2.95 (1.56–5.60, *p* < 0.001)	0.60 (0.15–2.30, *p* = 0.453)
DM meds (Yes)	6.88 (3.75–12.62, *p* < 0.001)	3.06 (1.16–8.10, *p* = 0.024)
HTN meds	3.16 (1.55–6.46, *p* = 0.002)	0.23 (0.05–1.08, *p* = 0.062)
ACEIARB (Yes)	4.18 (2.25–7.79, *p* < 0.001)	1.86 (0.64–5.42, *p* = 0.253)
Cholesterol Baseline	0.68 (0.53–0.88, *p* = 0.003)	0.83 (0.63–1.09, *p* = 0.184)
Creatinine Baseline	1.05 (1.04–1.07, *p* < 0.001)	1.04 (0.89–1.20, *p* = 0.648)
eGFR Baseline	0.92 (0.90–0.94, *p* < 0.001)	0.96 (0.82–1.11, *p* = 0.576)
sBP Baseline	1.02 (1.01–1.04, *p* = 0.004)	1.01 (0.99–1.03, *p* = 0.263)
dBP Baseline	0.98 (0.96–1.01, *p* = 0.136)	
BMI Baseline	0.99 (0.95–1.03, *p* = 0.672)	

Note: CKD: chronic kidney disease; CHD: coronary heart disease; DM: diabetes mellitus; HTN: hypertension; ACEI: angiotensin-converting enzyme inhibitors; ARB: angiotensin II receptor blockers; BMI: body mass index; eGFR: estimated glomerular filtration rate; SBP: systolic blood pressure; DBP: diastolic blood pressure; CI: confidence interval.

with CKD progression. However, the evidence of non-linearity based on the non-linear *p*-value was relatively weak compared to the other factors. Regarding baseline eGFR, the overall association with CKD progression was statistically significant, yet the relatively high non-linear *p*-value suggested that the relationship was primarily linear.

Discussion

The present study aimed to explore the use of interpretable ML models for predicting CKD risk based on a diverse set of clinical characteristics and demographic data. Our results provide valuable insights into the potential of ML

techniques for enhancing the CKD classification and stratification and forecasting CKD progression. Our investigation revealed that the XGBOOST model exhibited the highest predictive accuracy in predicting CKD outcome. The superior performance of the XGBOOST model suggests its potential as a robust tool for predicting CKD progression and underscores its clinical relevance in supporting healthcare decision-making.

Accurate prediction of CKD progression can guide clinical decision-making and improve patient outcomes. ML techniques have shown promise in developing predictive models using electronic health record data. However, most ML models for CKD risk prediction are complex
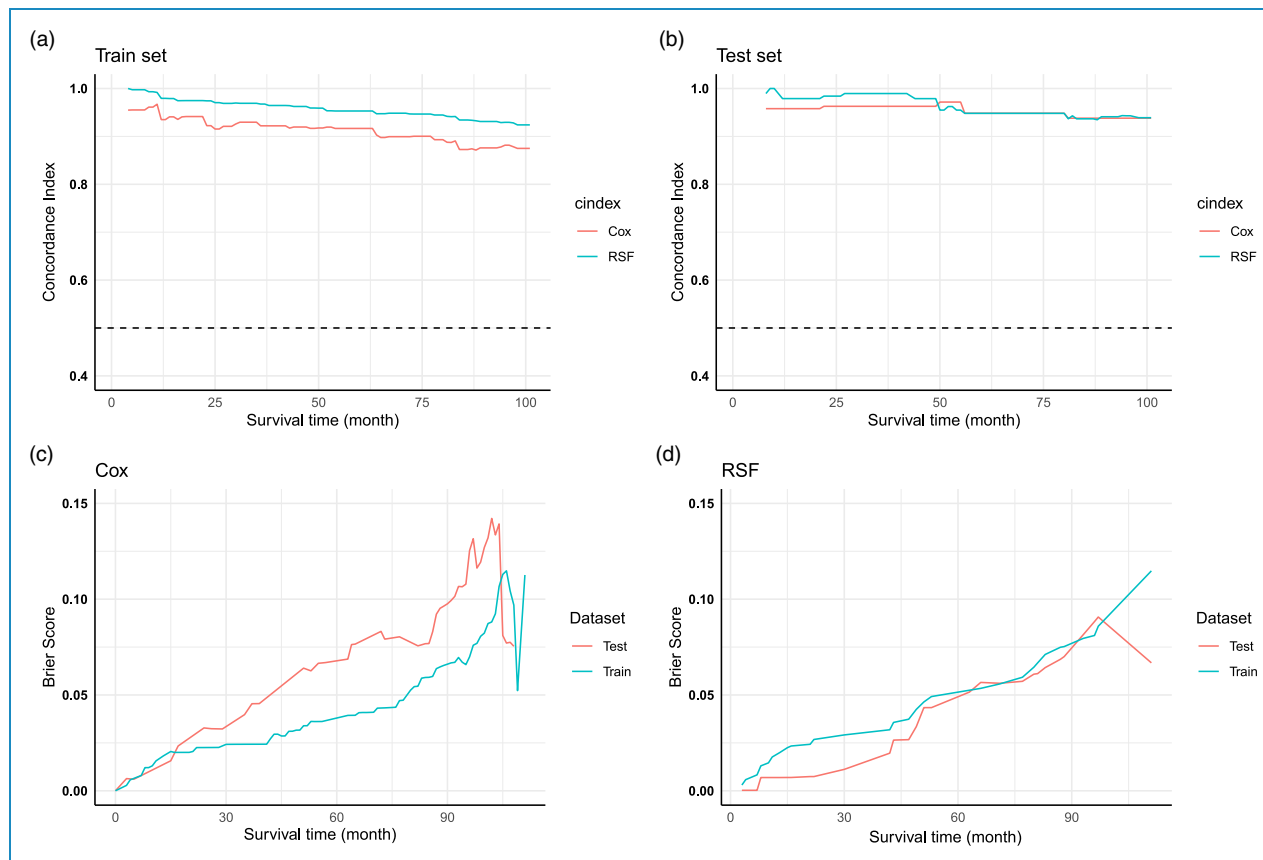
Figure 4. Overtime C-index and Brier scores for predicting disease-specific survival of CKD stages 3–5 with Cox regression and RSF in train and test dataset. (a) The overtime C-index by COX and RSF model in Trian set. (b) The overtime C-index by COX and RSF model in Test set. (c) The overtime of C-index in by COX model Trian and Test set. (d) The overtime of C-index in by RSF model Trian and Test set. CKD: chronic kidney disease; RSF: random survival forest.

black-box models.[33,34] While accuracy is important, interpretability is also critical for understanding model predictions and biases. Healthcare providers are unlikely to trust and implement black-box models. Then, one of the key strengths of our study lies in the interpretability of the ML models employed for CKD risk classification. By utilizing variable importance analysis, we identified three crucial variables that significantly contributed to the predictive performance of the models: eGFR, age, and creatinine. These findings underscore the critical role of these factors in determining CKD risk and highlight the importance of incorporating them into risk prediction models. Such insights can aid clinicians in identifying high-risk patients and implementing targeted interventions to manage CKD progression effectively. But one interest things we found in our study, that medicine taken with ACEIs and ARBs is related to CKD outcome, ACEIs and ARBs reduce intraglomerular pressure and proteinuria, which helps slow the progression of CKD. Also, the diabetes medications taken is related to CKD progress in survival analysis, tight control of diabetes with medications that lower blood glucose can significantly slow the

progression of diabetic kidney disease. These include insulin, metformin, sulfonylureas, and other anti-diabetic drugs.

eGFR, an estimate of the kidney's glomerular filtration rate, emerged as the most important predictor in our models. This finding is consistent with previous research highlighting eGFR as a fundamental marker for kidney function and a key indicator of CKD risk.[5,35] As eGFR decreases, the risk of CKD progression increases, underscoring the importance of regular eGFR monitoring in clinical practice. Age, another critical factor, demonstrated a non-linear association with CKD progression in our analysis.[36] Specifically, we observed higher risks of CKD progression in patients aged 52–55 years and 65–66 years, but the risk of CKD is still lower with 25% probability. These findings emphasize the age-related vulnerability to CKD development and highlight the need for age-specific risk assessment and preventive strategies.[37] Creatinine, a marker of kidney function, also exhibited significant predictive power in our models. Elevated levels of creatinine were associated with an increased risk of CKD progression.[21,38] This aligns with the established role of creatinine
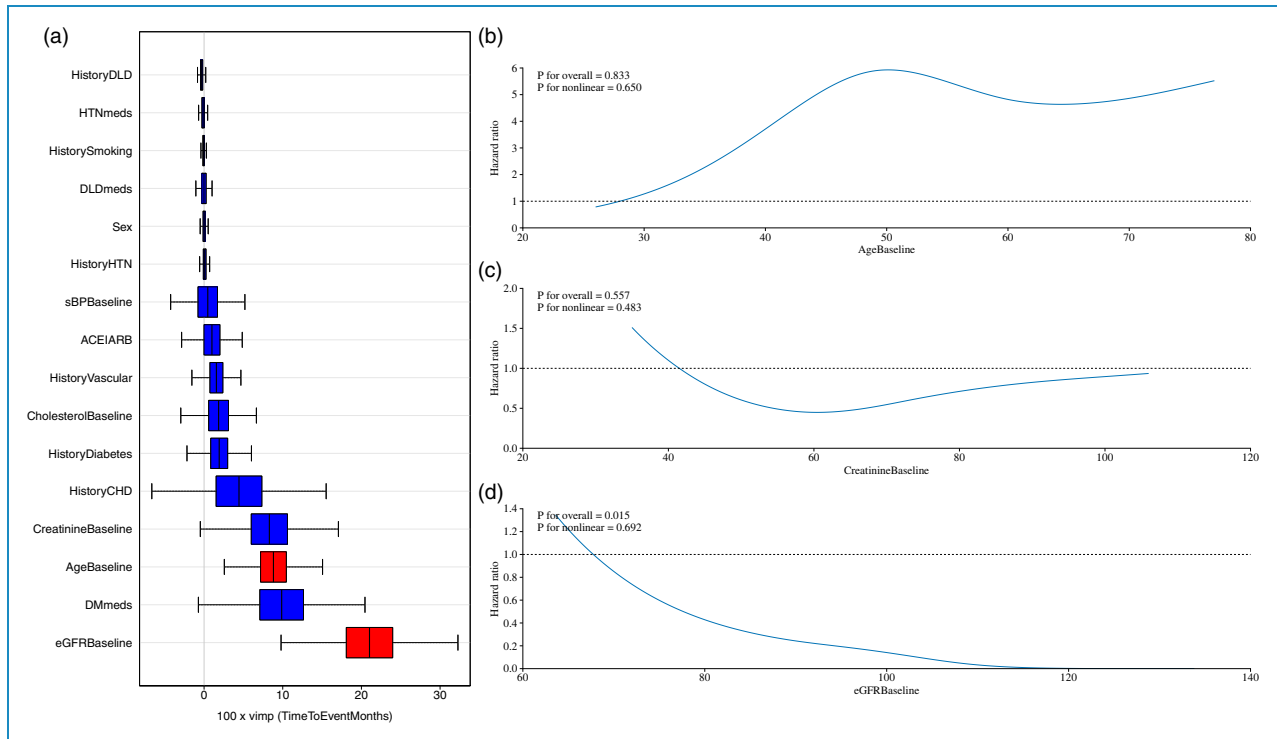
Figure 5. Variable importance of prediction CKD progression by RSF model and the relationship between eGFR baseline, creatinine baseline, and age baseline with CKD progression by restricted cubic spline (RCS) curves. (a) Variable importance of prediction CKD progression by RSF model. (b) The hazard ratio by RCS curves of Age Baseline with CKD progression. (c) The hazard ratio by RCS curves of Creatinine Baseline with CKD progression. (d) The hazard ratio by RCS curves of eGFR Baseline with CKD progression. CKD: chronic kidney disease; RSF: random survival forest.

as an essential clinical marker for kidney function and reinforces its importance in CKD risk assessment.[39] The interpretability of the variables provides a foundation for the development of risk prediction tools with improved clinical utility. Clinicians can leverage this knowledge to develop risk stratification algorithms that are easily understandable and applicable in real-world practice.[40] By incorporating eGFR, age, and creatinine into CKD risk prediction models, healthcare providers can accurately identify patients at higher risk of CKD progression, enabling early interventions and more effective disease management.

In our study, we evaluated the performance of two survival analysis models, Cox proportional hazards regression and RSFs, for predicting the progression of CKD. Both models have been widely used in survival analysis, each with its own strengths and limitations. The RSF model showed slightly higher performance compared to the COX model, with a C-index of 0.849 on the training set and 0.854 on the test set. These results suggest that RSF has a slightly better ability to generalize to new data and provides better calibration, making it a promising model for CKD progression prediction. COX is more suitable when the proportional hazards assumption holds, and the focus is on estimating HRs for individual predictors.[24] On the other hand, RSF is preferred when dealing with high-

dimensional data and exploring complex interactions between predictors.[25] RSF stands out as a powerful tool for handling non-linearity, high-dimensional data, and complex interactions among predictors. The advantages of ML, exemplified by RSF in our study, hold great potential in advancing CKD research and facilitating personalized risk prediction in clinical practice.

Furthermore, the interpretability of the ML models through variable importance analysis and RCS curves provides clinicians with valuable insights into the underlying mechanisms of CKD progression. Based on the association between BP and the risk of CKD, their study has founded the non-linear relationship with BP and risk of CKD.[32] These insights can inform the development of targeted interventions and personalized treatment plans for patients at various stages of CKD.

Despite the promising results, our study has several limitations that should be considered. Firstly, the reliance on data from a single center constrains generalizability of the findings. External validation across diverse populations and healthcare settings is imperative to rigorously assess the robustness of the models. Therefore, external validation using data from diverse healthcare settings is essential to confirm the robustness of our models. Additionally, our dataset was limited to specific clinical and demographic

variables, also the prediction of CKD 3–5 stages is based on eGFR. The CKD risk could be also defined by urinary albumin creatinine ratio, which our study is lacked with this data. Incorporating multi-modal data encompassing genetics, biomarkers, and lifestyle factors could further enhance predictive accuracy and contribute to a more comprehensive risk prediction model for CKD progression. Moreover, our study focused on predicting CKD progression based on baseline information and more advanced ML techniques such as deep neural networks warrant exploration as they may unlock substantial improvements in predictive accuracy.

This work motivates several promising directions to overcome the limitations and realize the full potential of ML for advancing clinical practice. Foremost is external validation of the models using real-world data from varied patient cohorts and care settings. Second, harnessing emerging technologies from wearables to electronic health records offers abundant opportunities to strengthen predictions through the integration of diverse longitudinal data. Evaluating model transportability across disease areas is imperative to determine wider clinical utility. The framework could also be extended to incorporate sensor data and applied to other disease contexts. Additionally, large language models like OpeanAI-GPT provide new opportunities to generate predictive features and interactions.[41] Finally, a critical next phase entails translating enhanced prognostic capabilities into personalized interventions and treatment plans tailored to individuals' risk profiles and trajectories. Such steps toward precision medicine hold immense promise for optimizing patient outcomes.

Conclusion

Our study demonstrates the potential of interpretable ML models in predicting CKD classification and CKD progression along with identifying significant risk factors. The findings highlighted the advantages of ML in handling survival data, particularly in terms of non-linearity and high-dimensional predictors. Future research efforts should focus on external validation and the integration of diverse data sources to further improve CKD risk prediction and advance personalized healthcare strategies for CKD patients.

Author's note

Shi-Yang Guan is also affiliated with Department of Epidemiology and Biostatistics, School of Public Health, Anhui Medical University, Hefei, China.

Author contributions: Conceptualization, WWM and SXZ; methodology, formal analysis, and data curation, JXZ, JZ and SYG; writing—original draft preparation, JXZ; writing—review and editing, JXZ and XL; supervision, WWM and SXZ. All authors have read and agreed to the published version of the manuscript.

ORCID iD: Jin-Xin Zheng https://orcid.org/0000-0003-1476-1903
Jiang Zhu https://orcid.org/0000-0003-4194-2355

References

1. Luyckx VA, Al-Aly Z, Bello AK, et al. Sustainable development goals relevant to kidney health: an update on progress. *Nat Rev Nephrol* 2021; 17: 15–32.
2. Webster AC, Nagler EV, Morton RL, et al. Chronic kidney disease. *Lancet* 2017; 389: 1238–1252.
3. Burden of Kidney Diseases - PAHO/WHO|Pan American Health Organization [Internet]. [accessed on 2023 Aug 7]. Available from: https://www.paho.org/en/enlace/burden-kidney-diseases
4. O'Seaghdha CM, Lyass A, Massaro JM, et al. A risk score for chronic kidney disease in the general population. *Am J Med* 2012; 125: 270–277.

5. Chen TK, Knicely DH and Grams ME. Chronic kidney disease diagnosis and management. *JAMA* 2019; 322: 1294–1304.

6. Kalantar-Zadeh K, Lockwood MB, Rhee CM, et al. Patient-centred approaches for the management of unpleasant symptoms in kidney disease. *Nat Rev Nephrol* 2022; 18: 185–198.

7. Shih CC, Lu CJ, Chen GD, et al. Risk prediction for early chronic kidney disease: results from an adult health examination program of 19,270 individuals. *Int J Environ Res Public Health* 2020; 17: 4973.

8. Lousa I, Reis F, Beirão I, et al. New potential biomarkers for chronic kidney disease management—A review of the literature. *Int J Mol Sci* 2020; 22: 43.

9. Orlando LA, Belasco EJ, Patel UD, et al. The chronic kidney disease model: a general purpose model of disease progression and treatment. *BMC Med Inform Decis Mak* 2011; 11: 41.

10. Yuan Q, Zhang H, Deng T, et al. Role of artificial intelligence in kidney disease. *Int J Med Sci* 2020; 17: 970–984.

11. Ahsan MM, Luna SA and Siddique Z. Machine-Learning-Based disease diagnosis: a comprehensive review. *Healthcare (Basel)* 2022; 10: 541.

12. Nateghi Haredasht F, Viaene L, Pottel H, et al. Predicting outcomes of acute kidney injury in critically ill patients using machine learning. *Sci Rep* 2023; 13: 9864.

13. Yao L, Zhang H, Zhang M, et al. Application of artificial intelligence in renal disease. *Clinical EHealth* 2021; 4: 54–61.

14. Islam MA, Majumder MZH and Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. *J Pathol Inform* 2023; 14: 100189.

15. Quazi S. Artificial intelligence and machine learning in precision and genomic medicine. *Med Oncol* 2022; 39: 120.

16. Petch J, Di S and Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Can J Cardiol* 2022; 38: 204–213.

17. Martin SA, Townend FJ, Barkhof F, et al. Interpretable machine learning for dementia: a systematic review. *Alzheimers Dement* 2023; 19: 2135–2149.

18. Guan C, Ma F, Chang S, et al. Interpretable machine learning models for predicting venous thromboembolism in the intensive care unit: an analysis based on data from 207 centers. *Crit Care* 2023; 27: 406.

19. Linardatos P, Papastefanopoulos V and Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy* 2021; 23: 18.

20. Al-Shamsi S, Regmi D and Govender RD. Chronic kidney disease in patients at high risk of cardiovascular disease in the United Arab Emirates: a population-based study. *PLoS One* 2018; 13: e0199920.

21. Inker LA, Astor BC, Fox CH, et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. *Am J Kidney Dis* 2014; 63: 713–735.

22. George B, Seals S and Aban I. Survival analysis and regression models. *J Nucl Cardiol* 2014; 21: 686–694.

23. Hsu JY, Roy JA, Xie D, et al. Statistical methods for cohort studies of CKD: survival analysis in the setting of competing risks. *Clin J Am Soc Nephrol* 2017; 12: 1181–1189.

24. Mogensen UB, Ishwaran H and Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw* 2012; 50: 1–23.

25. Wang H and Zhou L. Random survival forest with space extensions for censored data. *Artif Intell Med* 2017; 79: 52–61.

26. Parodi S, Verda D, Bagnasco F, et al. The clinical meaning of the area under a receiver operating characteristic curve for the evaluation of the performance of disease markers. *Epidemiol Health* 2022; 44: e2022088.

27. Choi I, Wells BJ, Yu C, et al. An empirical approach to model selection through validation for censored survival data. *J Biomed Inform* 2011; 44: 595–606.

28. Zhu J, Zheng J, Li L, et al. Application of machine learning algorithms to predict central lymph node metastasis in T1-T2, non-invasive, and clinically node negative papillary thyroid carcinoma. *Front Med (Lausanne)* 2021; 8: 635771.

29. Archer KJ and Kimes RV. Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal* 2008; 52: 2249–2260.

30. Zheng JX, Xia S, Lv S, et al. Infestation risk of the intermediate snail host of Schistosoma japonicum in the Yangtze River Basin: improved results by spatial reassessment and a random forest approach. *Infect Dis Poverty [Internet]* 2021; 10: 74. Available at: https://pubmed.ncbi.nlm.nih.gov/34011383/ (accessed 7 August 2023).

31. Chen L, Hu Y, Ma Y, et al. Non-linear association of fasting C-peptide and uric acid levels with renal dysfunction based on restricted cubic spline in patients with type 2 diabetes: A real-world study. *Front Endocrinol [Internet]* 2023; 14: 1157123.

32. Lee H, Kwon SH, Jeon JS, et al. Association between blood pressure and the risk of chronic kidney disease in treatment-naïve hypertensive patients. *Kidney Res Clin Pract* 2022; 41: 31–42.

33. Lee AM, Hu J, Xu Y, et al. Using machine learning to identify metabolomic signatures of pediatric chronic kidney disease etiology. *J Am Soc Nephrol* 2022; 33: 375–386.

34. Lei N, Zhang X, Wei M, et al. Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis. *BMC Med Inform Decis Mak* 2022; 22: 205.

35. Courtney AE, Maxwell AP and Fogarty DG. Using estimated glomerular filtration rate (eGFR) to help manage patients with chronic kidney disease. *Ulster Med J* 2007; 76: 154–156.

36. Hallan SI, Matsushita K, Sang Y, et al. Age and the association of kidney measures with mortality and end-stage renal disease. *JAMA* 2012; 308: 2349–2360.

37. Denic A, Glassock RJ and Rule AD. Structural and functional changes with the aging kidney. *Adv Chronic Kidney Dis* 2016; 23: 19–28.

38. Lopez-Giacoman S and Madero M. Biomarkers in chronic kidney disease, from kidney function to kidney damage. *World J Nephrol* 2015; 4: 57–73.

39. Gowda S, Desai PB, Kulkarni SS, et al. Markers of renal function tests. *N Am J Med Sci* 2010; 2: 170–173.

40. Echouffo-Tcheugui JB and Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS Med* 2012; 9: e1001344.

41. The Lancet Digital Health null. ChatGPT: friend or foe? *Lancet Digit Health* 2023; 5: e102.

Appendix

Notation

CKD	Chronic Kidney Disease
WHO	World Health Organization
YLD	Years Lived with Disability
ML	Machine Learning
CVD	Cardiovascular Disease
LM	Linear Regression
RF	Random Forests
NNET	Neural Networks
XGBOOST	eXtreme Gradient Boosting
RSF	Random Survival Forests
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CI	Confidence Interval
PDP	Partial Dependence Plot
RCS	Restricted Cubic Splines
HR	Hazard Ratio
BMI	Body Mass Index
eGFR	Estimated Glomerular Filtration Rate
sBP	Systolic Blood Pressure
dBP	Diastolic Blood Pressure
ACEI	Angiotensin-Converting Enzyme Inhibitors
ARB	Angiotensin II Receptor Blockers
DLD	Dyslipidemia
HTN	Hypertension
DM	Diabetes Mellitus
KM	Kaplan–Meier
IBS	Integrated Brier Score