scientific reports

Check for updates

OPEN

Machine learning-based prediction models for renal impairment in Chinese adults with hyperuricaemia: risk factor analysis

Tianchen Wu[1,4], Hui Yang[2,4], Jinbin Chen[1] & Wenwen Kong[3]✉

In hyperuricaemic populations, multiple factors may contribute to impaired renal function. This study aimed to establish a machine learning-based model to identify characteristic factors related to renal impairment in hyperuricaemic patients, determine dose–response relationships, and facilitate early intervention strategies. Data were collected through the big data platform of Nanjing Hospital of Traditional Chinese Medicine, encompassing 2,705 patients with hyperuricaemia (1,577 with renal impairment, 828 without) from June 2019 to June 2022. After multiple imputations for missing values, the dataset was randomly split into training (70%) and validation (30%) sets. We employed three machine learning algorithms for feature selection: random forest (with 100 decision trees and an OOB error rate of 23.34%), LASSO regression (optimal lambda of -3.58), and XGBoost (learning rate of 0.3, maximum tree depth of 1, and 50 rounds of boosting). The intersection of features identified by these algorithms through Venn diagram analysis yielded four key predictors. A logistic regression model was subsequently constructed and evaluated for discrimination (AUC), calibration (Brier score), and clinical utility (DCA). Restricted cubic spline (RCS) curves were utilized to analyse the dose–response relationships. The model, which incorporates age, cystatin C (Cys-C), uric acid (UA), and sex, demonstrated robust performance, with an AUC of 0.818 [95% CI (0.796–0.817)] in the training set and an AUC of 0.82 [95% CI (0.787–0.853)] in the validation set. Calibration tests yielded Brier scores of 0.160 and 0.158, respectively. Clinical decision curves revealed optimal prediction probability intervals of 6-99.02% and 7-93.14%. In the hyperuricaemic population, each 0.5 mg/L increase in Cys-C, 10-year increase in age, and 100 μmol/L increase in UA corresponded to increased risks of 13%, 81%, and 73%, respectively. RCS analysis revealed nonlinear relationships for Age and Cys-C and a linear relationship for UA, with sex-specific distribution patterns. The machine learning-based model incorporating these four indicators demonstrated excellent predictive performance for renal impairment in hyperuricaemic patients. These findings suggest that monitoring Cys-C and UA levels while considering age and sex differences is crucial for risk assessment and prevention strategies.

Keywords Hyperuricaemia, Renal impairment, Machine learning, Random forest, LASSO, XGBoost, RCS curve

Renal impairment, which is chronic progressive renal parenchymal damage caused by various factors, prevents the maintenance of essential functions[1]. The causes of renal impairment are complex, and hyperuricaemia is one of the critical causes of its occurrence and development. Uric acid (UA) is the end product of purine metabolism in the body, and at normal levels, it can play a role in maintaining blood pressure, improving cognitive function, and resisting oxidative stress[2,3]. However, excessive accumulation leads to hyperuricaemia, which not only causes diseases such as gout and UA stones but also may lead to renal damage, including acute kidney injury, chronic kidney disease, and renal failure[4].

According to the data from the 2015–2019 China Health and Nutrition Survey, the prevalence of hyperuricaemia has been increasing annually in China, while there are also significant differences in male and female ratios, age distributions, and regional distributions[5]. However, we have found that people with

[1]Department of Neurology, Nanjing Hospital of Chinese Medicine Affiliated to Nanjing University of Chinese Medicine, Nanjing, China. [2]School of Nursing, Nanjing University of Chinese Medicine, Nanjing, China. [3]Department of Endocrinology, Nanjing Hospital of Chinese Medicine Affiliated to Nanjing University of Chinese Medicine, Nanjing, China. [4]Tianchen Wu and Hui Yang contributed equally to this work. ✉email: kw412@126.com

hyperuricaemia do not necessarily have renal impairment, and the dual attributes of UA make this result more challenging to interpret. Considering that many factors contribute to renal impairment, this study aimed to explore the risk factors for hyperuricaemia and its potential internal linkages in a specific population with renal impairment through machine learning methods to provide relevant guidance and reference for the clinic.

Methods

Research design and participants

Data were collected through the big data platform of Nanjing Traditional Chinese Medicine Hospital, and 3410 patients diagnosed with hyperuricaemia in the hospital within three years (June 2019 to June 2022) based on ICD-10 codes. The original data for this study can be found in the Supplementary Files. After applying the inclusion and exclusion criteria, this study included 2705 patients, 1577 with renal function impairment and 828 without renal function impairment. This study is retrospective.

The inclusion criteria are as follows: (1) Hyperuricaemia is defined as an adult UA level greater than 7 mg/dL (420 μmol/L) for men and 6 mg/dL (360 μmol/L) for women under a regular purine diet according to the latest relevant guidelines of the United States and Europe[6]. (2) According to the simplified MDRD and the modified MDRD Eq[7] suitable for the Chinese population, we defined renal function injury as an estimated glomerular filtration rate (eGFR) of less than 90 ml/min/1.73 m2.

The exclusion criteria are as follows: (1) More than 70% of the laboratory data were missing (n = 105/705, 14.9%). (2) Severe renal diseases, renal surgery within 1 year, acute renal failure (ARF), and a history of kidney transplantation or renal tumours (n = 257/705, 36.4%). (3) Severe dysfunction or decompensation of the heart, liver, brain, or other organs (n = 322/705, 45.7%). (4) Pregnant or lactating women (n = 21/705, 3.0%) (see Flowchart Fig. 1). This observational study was approved by the Ethics Committee of Nanjing Hospital of Chinese Medicine (approval number: KY2024040). All patient data were anonymized to protect personal privacy.

Statistical methods

This study used R language (R-4.2.1 version) for data analysis. Count data are expressed as frequencies (%); normally distributed data are expressed as the mean ± SD, and skewed data are expressed as the median M (P25, P75). Group differences were compared via t tests, Mann–Whitney U tests, and chi-square tests. We designated $P < 0.05$ (two-sided) as statistically significant.

The Mice package was used for missing value analysis and multiple interpolations with ten interpolations, selection of the postinterpolation dataset according to the BIC criterion[8], and sensitivity analysis of the interpolated dataset to the original dataset. The interpolated dataset, according to a ratio of 7:3, was randomly divided into training set machine learning (random forest, LASSO regression, XGBoost) and variable screening, and a Wayne diagram was plotted to take the intersection, the common characterization factors used to establish a logistic regression model and model testing. The fitting relationship between the characteristic variables and renal impairment caused by hyperuricaemia was determined via restricted cubic spline (RCS), and the related model construction requirements and results are reported in the TRIPOD statement list[9].

Results

Baseline characteristics of the participants

Following multiple imputations for missing values, no significant differences were observed between the imputed and original datasets. In both the training and validation sets, fasting blood glucose (FBG), aspartate aminotransferase (AST) and retinol binding protein (RBP) levels were not significantly different between patients with and without renal impairment ($P > 0.05$). However, all other characteristic variables demonstrated significant between-group differences ($P < 0.05$) (Table 1).

Feature selection through three machine learning models and venn diagram analysis

Random forest-based feature selection

After a tenfold crossover, the random forest was built in the training set in the validation set according to the differentiation of the ROC curve, as shown in the figure (Fig. 2A), and the AUC area under the curve was 0.805. We used OOB. The score used to evaluate the random forest model[10] (Fig. 2B) showed that with an increase in the number of decision trees to approximately 100, the observed error rate of approximately 23.34% appears to have stabilized. We ranked the variables in the model on the basis of their importance (Fig. 2C) according to the MDA[11].

Feature selection based on LASSO regression analysis

LASSO regression analysis was performed to identify key features associated with [outcome variable]. As shown in Fig. 3, the LASSO coefficient paths (Fig. 3A) demonstrate the shrinkage of variable coefficients with increasing regularization parameters (lambda). The cross-validation plot (Fig. 3B) reveals the optimal lambda value (marked by the dotted line), where the misclassification error was minimized while maintaining model parsimony. Using the one-standard-error rule (lambda.1 se), the optimal regularization parameter was identified at $\log(\lambda) = -3.58$, as shown by the vertical dashed line in Fig. 3B. At the optimal lambda value, 5 variables were retained in the model, including Age, Cystatin C (Cys-C), UA, Male sex, and Female sex. The number of nonzero coefficients decreased from 17 to 0 as lambda increased, indicating effective feature selection through LASSO regularization.

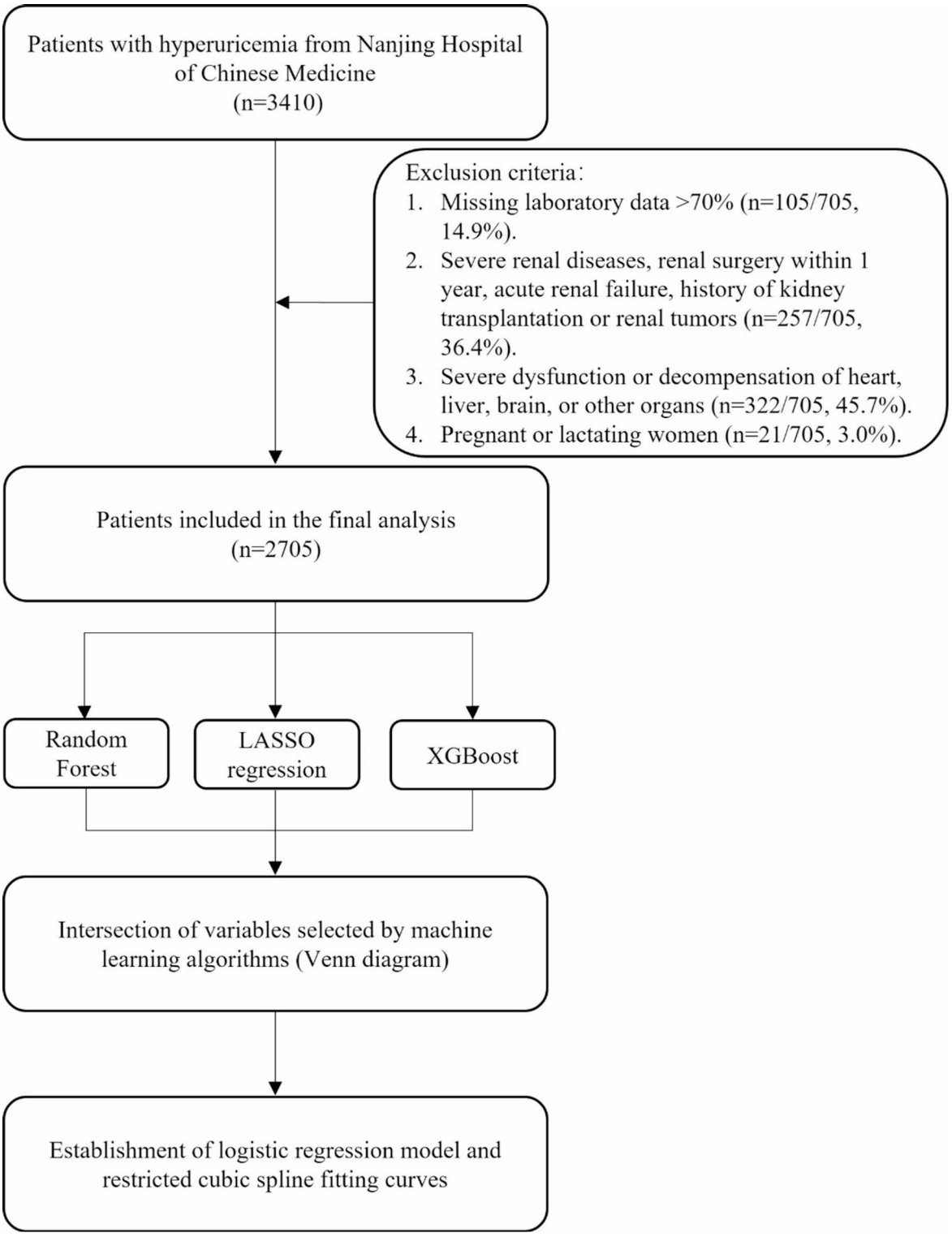

Fig. 1. Flowchart of the study selection process.

XGBoost-based feature selection and Venn diagram analysis

We trained an XGBoost model on the training set with a seed value of 1091. The optimal hyperparameters were determined through a grid search, including a binary logistic objective function, 50 rounds of boosting, a maximum tree depth of 1, a learning rate of 0.3, zero gamma, and a column sampling ratio of 0.8. The model's performance was validated on the test set, yielding an AUC of 0.828 [95% CI: 0.797–0.860], as illustrated by the ROC curve (Fig. 4A). Feature importance was assessed via gain values[12], with the results displayed in Fig. 4B.

Characteristics name	Training set (n = 1684)			Validation set (n = 721)		
	No(n = 580)	Yes (n = 1104)	p	No (n = 248)	Yes (n = 473)	p
Gender (n(%))						
Male	519 (89.5%)	772 (69.9%)	< 0.001	233 (94%)	337 (71.2%)	< 0.001
Female	61 (10.5%)	332 (30.1%)		15 (6%)	136 (28.8%)	
Age (M(P_{25}, P_{75}), year)	52.00 (37.00–65.00)	73.00 (60.00–81.00)	< 0.001	51.00 (39.00–65.00)	72.00 (60.00–81.00)	< 0.001
UA (M(P_{25}, P_{75}), μmol/L)	484.00 (449.50–535.00)	503.00 (460.00-563.50)	< 0.001	478.00 (445.00-514.5)	500.00 (457.00-563.00)	< 0.001
FBG (M(P_{25}, P_{75}), mmol/L)	5.60 (4.94–6.74)	5.66 (4.97–7.00)	0.083	5.58 (4.96–6.50)	5.79 (5.06–7.00)	0.103
HbA1c	5.90 (5.50–6.70)	6.10 (5.66–7.00)	< 0.001	5.83 (5.45–6.70)	6.10 (5.62–7.00)	< 0.001
TG (M(P_{25}, P_{75}), mmol/L)	2.00 (1.31–3.10)	1.61 (1.07–2.35)	< 0.001	1.93 (1.31–3.26)	1.56 (1.08–2.27)	< 0.001
TC (M(P_{25}, P_{75}), mmol/L)	4.92 (4.04–5.83)	4.45 (3.63–5.38)	< 0.001	4.90 (4.12–5.74)	4.31 (3.62–5.14)	< 0.001
LDL-C (M(P_{25}, P_{75}), mmol/L)	2.79 (2.13–3.37)	2.43 (1.81–3.08)	< 0.001	2.73 (2.20–3.41)	2.42 (1.85–3.12)	< 0.001
HDL-C (M(P_{25}, P_{75}), mmol/L)	1.04 (0.88–1.23)	1.06 (0.91–1.27)	0.027	1.03 (0.88–1.23)	1.05 (0.90–1.27)	0.280
AST (M(P_{25}, P_{75}), μ/L)	20.00 (15.00–27.00)	20.00 (15.00–26.00)	0.745	20.00 (15.00–26.00)	19.00 (15.00–26.00)	0.714
ALT (M(P_{25}, P_{75}), μ/L)	25.00 (16.00–42.00)	18.00 (12.00–29.00)	< 0.001	23.00 (15.00–38.00)	18.00 (13.00–30.00)	< 0.001
RBP (M(P_{25}, P_{75}), mg/L)	47.10 (38.25–58.05)	47.65 (37.95 -60.00)	0.361	50.20 (39.95–59.20)	49.80 (38.30–61.40)	0.866
Cys-C (M(P_{25}, P_{75}), mg/L)	0.94 (0.78–1.27)	1.53 (1.11–2.37)	< 0.001	0.94 (0.75–1.33)	1.54 (1.14–2.23)	< 0.001
AIP	3.64 (2.74–4.73)	3.18 (2.40–4.00)	< 0.001	3.67 (2.76–4.75)	3.13 (2.40–3.98)	< 0.001
TyG	7.52 (7.02–8.00)	7.30 (6.76–7.82)	< 0.001	7.47 (7.01–8.07)	7.32 (6.89–7.79)	0.004
SII	419.12 (291.72637.03)	482.69 (315.21-793.02)	< 0.001	421.53 (292.64-637.23)	487.15 (337.11–833.90)	< 0.001

Table 1. Baseline characteristics of the study population stratified by training and validation sets after data interpolation. Data are presented as n (%) for categorical variables and median (interquartile range) for continuous variables. UA: uric acid; FBG: Fasting Blood Glucose; HbA1c: glycated hemoglobin; TG: triglycerides; TC: total cholesterol; LDL-C: low-density lipoprotein cholesterol; HDL-C: high-density lipoprotein cholesterol; AST: aspartate aminotransferase; ALT: alanine aminotransferase; RBP: retinol binding protein; Cys-C: cystatin C; AIP: atherogenic index of plasma; TyG: triglyceride glucose index; SII: systemic immune-inflammation index. P-values were calculated using Mann-Whitney U test for continuous variables and Chi-square test for categorical variables. $P < 0.05$ was considered statistically significant. The AIP is calculated as the logarithm (base 10) of the ratio of triglycerides to high-density lipoprotein cholesterol (log[TG/HDL-C]).

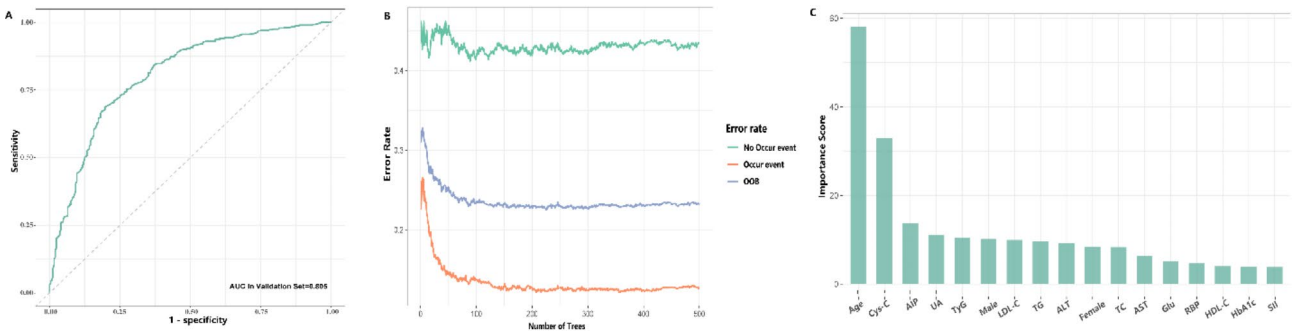

Fig. 2. (**A**) ROC diagram of RF model. (**B**) The relationship between tree. (**C**) Variable Importance Plot of RF model (Mean Decrease Accuracy)

Furthermore, we conducted a comprehensive feature evaluation by integrating SHAP values with gain values and we present the variable importance ranking based on SHAP analysis in Fig. 4C.

Through Venn diagram analysis (Fig. 4D), we examined the overlap of important features identified by the three distinct machine learning algorithms. The analysis revealed four consistent predictive variables (age, Cys-C, UA, and sex) across all methods, demonstrating their significant predictive value in identifying renal dysfunction among individuals with hyperuricaemia.

Model construction and performance evaluation

Model construction and discrimination analysis

A multicollinearity assessment of the four characteristic variables in the training set revealed that all variance inflation factors (VIFs) were less than 5, with individual factors remaining under the threshold contour

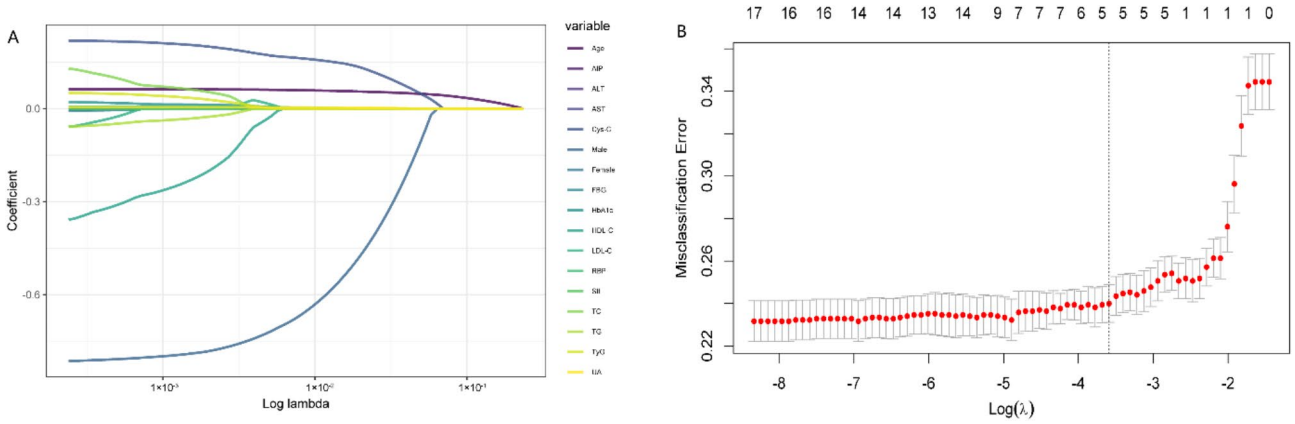

Fig. 3. (**A**) Lasso Path map. (**B**) Lasso cross-validation diagram.

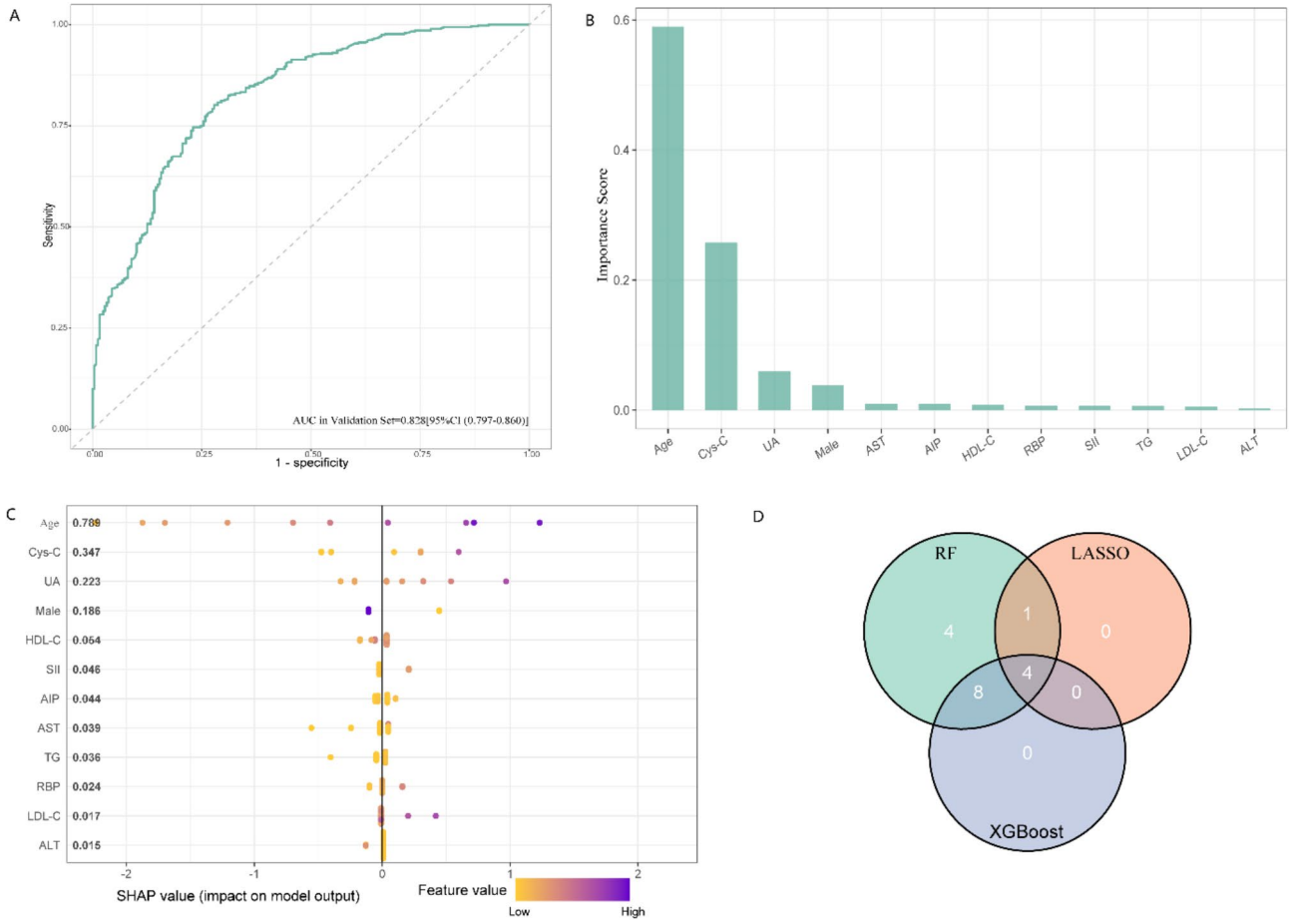

Fig. 4. (**A**) ROC diagram of XGBoost model. (**B**) Variable Importance Plot of XGBoost model (Gain). (**C**) Variable Importance Plot of XGBoost model (Gain). (**D**) Venn diagram of joint selection of characteristic variables.

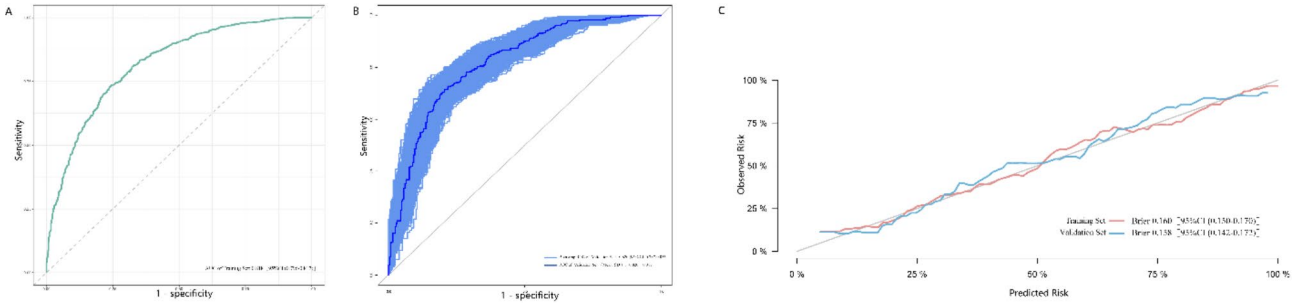

Fig. 5. (**A**) ROC Curve of a prediction model in the training set. (**B**) ROC Curve of prediction model in Validation Set(bootstrap=1000). (**C**) Training set and Verification set calibration curve.

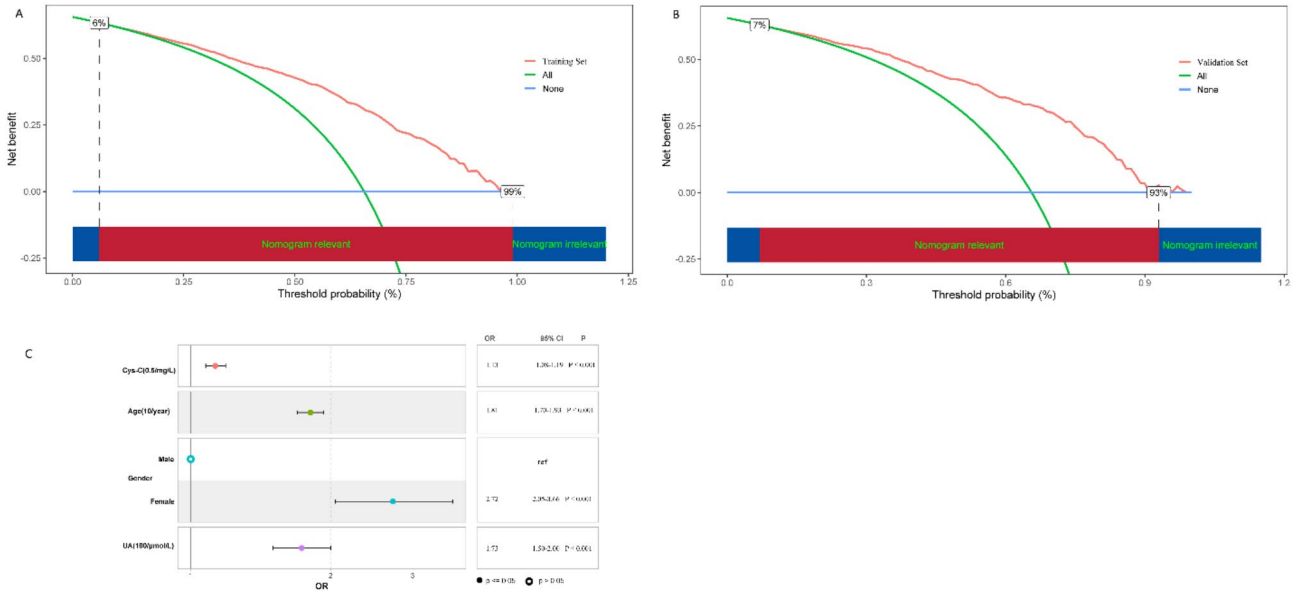

Fig. 6. (**A**) Training set Decision Curve Analysis. (**B**) Verification set Decision Curve Analysis. (**C**) Model forest map after adjusting the value of characteristic variables.

(Supplementary Fig. S1)[13]. The model had no evidence of significant multicollinearity or anomalous characteristic influencing factors. The formula for model building is as follows:

$$L = -6.271 + \mathrm{Cys-C(mg/L)} * 0.220 + \mathrm{Age(Year)} * 0.622 + \mathrm{Gender(Female)} * 0.810 + \mathrm{UA(umol/L)} * 0.001$$

The ROC discrimination test in the training set yielded an AUC = 0.818 [95% CI (0.796–0.817)] (Fig. 5A). Additionally, we performed 1000 bootstrap validations in the validation set and obtained an area under the curve (AUC) of 0.820. [95% CI (0.787–0.853)] (Fig. 5B).

Model calibration analysis

We utilized repeated sampling with the bootstrap method to calculate the confidence interval of the Brier score[14], and we set the parameters to sample 50 data points at a time during the process. We repeated the sampling procedure 1,000 times to ensure the robustness of our results (B = 50, M = 1,000), and the Brier score of the calibration curve for the training set and validation set (Fig. 5C) was 0.160 [95% CI (0.150–0.170)] and 0.158 [95% CI (0.142–0.172)], respectively. Both the training set and validation set passed the calibration degree calibration, and the consistency test of the model also passed.

Clinical utility assessment and risk factor analysis

The clinical utility of our predictive model was assessed through decision curve analysis (DCA). We applied DCA to evaluate the net benefit across different threshold probabilities in both the training and validation sets. The analysis yielded promising results, with threshold probabilities ranging from 6 to 99.02% in the training set and 7–93.14% in the validation set, indicating robust clinical decision-making capabilities (Fig. 6A, B).

Model	Tjur's R^2	RMSE	Log_loss	Score_spherical	PCP
Model 1	0.290	0.400	0.488	<0.001	0.679
Model 2	0.270	0.399	0.488	0.002	0.680
Model 3	0.290	0.399	0.487	<0.001	0.680

Table 2. Comparison of model performance metrics across three datasets. Model 1 represents the analysis using the original dataset; Model 2 was developed using the dataset with non-missing values only; Model 3 was constructed using the dataset with missing values imputed. RMSE: root mean square error; PCP: percentage of correct predictions.

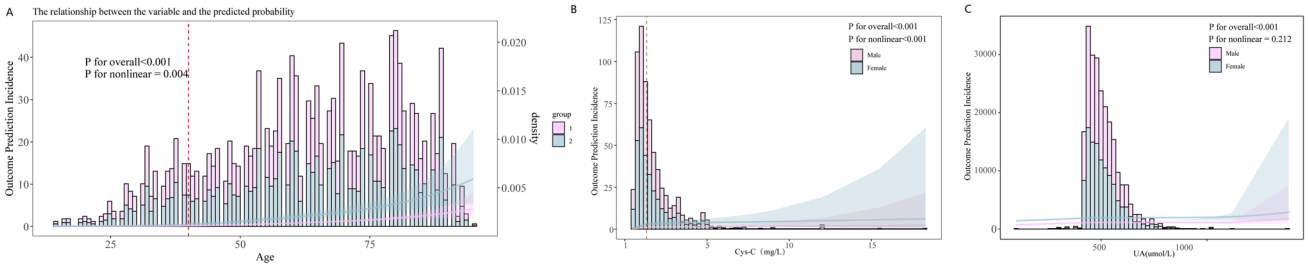

Fig. 7. (**A**) The relationship between the Age and the predicted probability. (**B**) The relationship between the Cyc-c and the predicted probability. (**C**) The relationship between the UA and the predicted probability.

To facilitate practical clinical interpretation, we recalibrated the continuous variables into clinically relevant intervals (Fig. 6C). This transformation revealed several significant associations with renal impairment risk in hyperuricaemic patients. Specifically, we observed that a cystatin C elevation of 0.5 mg/L corresponded to a 13% increased risk (OR: 1.13, 95% CI: 1.081–1.192). Age had a substantial impact, with each decade increase associated with an 81% higher risk (OR: 1.81, 95% CI: 1.704–1.932). Similarly, a UA increase of 100 µmol/L was correlated with a 73% increased risk (OR: 1.73, 95% CI: 1.501–2.001).

Sensitivity analysis

To ensure the robustness of the results, we conducted a sensitivity analysis, and we divided the data between cases with and without missing values in the original data, where the dataset cases contained missing values (n = 1593) and cases without missing values (n = 812).

The original dataset and the interpolated dataset were compared between groups, suggesting that there were some differences between the interpolated dataset and the nonmissing dataset in terms of WBC, TyG, and TBIL ($p < 0.001$), and there were no significant differences between groups in the rest of the characteristic variables. Moreover, we established models for the above four feature variables in both the original dataset and the nonmissing dataset, and the interpolated dataset, suggesting that there was no significant difference among the three in the establishment of the model and the differentiation by the likelihood ratio test ($p > 0.05$), and the relevant parameters of the test are given in Table 2.

Dose-response analysis of risk factors for renal impairment

To investigate the dose-response relationships between risk factors and renal impairment in hyperuricaemic patients, we performed restricted cubic spline (RCS) regression analyses with four knots at the 5th, 35th, 65th, and 95th percentiles. The RCS analysis revealed nonlinear associations for both age and Cys-C with renal impairment risk. For age, a significant nonlinear relationship was observed (P for nonlinearity = 0.004), with an inflection point at 40 years. The risk was more pronounced in patients aged ≤ 40 years (OR: 1.170 [95% CI: 1.116–1.235]) than in those aged > 40 years (OR: 1.054 [95% CI: 1.045–1.062]) (Fig. 7A). Similarly, Cys-C demonstrated a significant nonlinear association (P for nonlinearity < 0.001), with a threshold effect at 1.33 mg/L. The risk increase was substantially greater below this threshold (OR: 5.210 [95% CI: 3.078–8.797]) than above 1.33 mg/L (OR: 1.048 [95% CI: 0.949–1.166]) (Fig. 7B). In contrast, UA showed a linear relationship with renal impairment risk (P for nonlinearity = 0.212), suggesting a consistent dose-dependent effect across all observed concentrations (Fig. 7C).

Discussion

We conducted an observational study on the occurrence of renal function injury in a hyperuricaemic population at Nanjing Hospital of Traditional Chinese Medicine. Through data screening and inclusion and multiple interpolation methods to address missing data values, we utilized machine learning for important variable screening and identified four factors indicating renal impairment in the hyperuricaemia population. We tested the established model for discriminatory calibration and drew clinical decision curves; additionally, we conducted sensitivity analyses in various datasets, examining linear dose-response relationships through RCS curves.

The present study is a retrospective cohort study of a hyperuricaemic population in Nanjing, China, with the aim of investigating the correlation between cumulative mean hyperuricaemia and renal impairment events. Our findings may provide novel insights into critical indicators that may help advance the development of early strategies to prevent the occurrence of renal impairment in hyperuricaemic populations.

Elevated UA levels are commonly observed in patients with metabolic syndrome and are widely acknowledged as risk factors for hypertension, gout, nonalcoholic fatty liver disease, chronic kidney disease (CKD), and cardiovascular disease[15]. Hyperuricaemia is considered an independent risk factor for kidney disease, and early renal function protection can benefit from lowering UA. Zhou[16] et al. reported that higher UA levels ultimately lead to the development of renal disease and a rapid decline in the eGFR in a 6-year cohort study of 10,677 people with a normal initial eGFR and no proteinuria. Analysis of data from 6,642 patients with renal impairment included in the NHANES database from 1999 to 2018 revealed a J-shaped nonlinear relationship between UA concentration and death due to renal impairment and CVD mortality after adjustment for relevant confounders, with inflection points of 311.65 and 392.34 μmol/L, respectively; for every 50 μmol/L increase in UA, renal impairment and CVD mortality increased by 11.7% and 17.0%, respectively[17]. The above two investigative studies have several limitations because they included a population with standard UA and did not restrict the study population to the specific population with hyperuricaemia. As mentioned earlier, whether asymptomatic hyperuricaemia necessarily leads to the progression of CRF in the hyperuricaemic population has been controversial for many years[18]. For example, Jordan DM[19] used a Mendelian randomized extrapolation method, which does not support a causal relationship between high and low UA levels and eGFR levels or CRF risk.

Lowering UA levels does not help reduce the risk of CRF. In contrast, a cross-racial Mendelian randomization study by Wu, S[20] confirmed the existence of a bidirectional causal effect between UA concentration and renal function. These findings emphasize the importance of ethnic differences, and that in Asian populations, elevated levels of UA contribute to the risk of developing CRF. An experimental study in animals by Sellmayr M[21] indicated that this UA level did not reduce the risk of developing CRF unless UA crystallizes in the kidney. Crystallization in the kidney leads to the development of crystalline nephropathy, which results in kidney damage; asymptomatic hyperuricaemia does not affect the progression of CRF. Allopurinol UA-lowering therapy did not slow the decline in the eGFR in a population with established renal damage[22]. A Korean chronic kidney disease cohort study revealed that hyperuricaemia was an independent risk factor for renal outcomes, but two drugs, allopurinol and febux, were not shown to have significant renoprotective effects[23]. In the present study, in a specific population with hyperuricaemia, all three machine learning methods placed UA, a characteristic indicator, at the forefront, and after adjusting the values of the distinct variables, the likelihood of renal damage increased by 73% for every 100 μmol/L increase in UA, controlling for confounding factors such as sex and age. The increase in the RCS curve showed a linear relationship, with the likelihood of kidney damage in the hyperuricaemic population gradually increasing as the UA level increased.

The current study revealed that age and sex are also factors influencing the development of renal impairment in hyperuricaemic populations, and the question of whether the increased burden of CRF is caused by the interaction of hyperuricaemia and cardiovascular disease (CVD) risk factors or accelerated by ageing is more debated. A study from the Chinese Health and Nutrition Survey (CHNS)[24] revealed a significant cumulative effect of hyperuricaemia and age on CRF, as calculated by a multiplicative interaction model, with a synergy index of 2.26. In populations in which advanced age and hyperuricaemia coexist, the likelihood of CRF appears to be superimposed. The cumulative risk of kidney injury increases with age in hyperuricaemic populations. As age increases to approximately 40 years, there may be a step-like downwards trend in renal function, leading to decreased excretion of UA and increased serum UA levels; at the same time, the elevation of UA results in a corresponding impairment of renal function and circulatory injury.

A retrospective cohort study in Zhejiang Province, China, by Luo, Y[25] revealed that CRF due to hyperuricaemia exhibited a pattern of sex differences. After controlling for confounders, we divided the UA levels into quartiles, used the lowest quartile as a reference, and performed subgroup analyses according to sex. The highest quartile presented a 2.16-fold increased risk compared with the lowest quartile in the male subgroup, and the risk in the female subgroup presented a 2.76-fold increased risk. The spline curves showed a U-shaped pattern, which indicated a potential threshold effect of UA on the risk of CRF. At the same UA level, the risk of kidney damage may be greater in women than in men. In this study, age and the occurrence of CRF did not show a linear relationship according to the RCS curve. Before the age of 40 years, the risk of CRF in the hyperuricaemic population increased by 17% per year of age, 1.17% [95% CI (1.116–1.235)]; after the age of 40 years, this risk showed a gradual downwards trend, with an increase in the risk of incidence of 5.4% per year, 1.054% [95% CI (1.045–1.062)], and the overall trend was faster and then slower. Moreover, at the same UA level, the risk of kidney damage in women exceeded that in men.

eGFR has always been the primary indicator for determining renal impairment in clinical practice. However, it is subject to external interference, and the accuracy of this indicator decreases significantly when short-term renal function changes occur[26]. Current research suggests that this index can potentially overestimate renal function in critically ill patients[27]. Cys-C is a popular alternative indicator of the eGFR, and as an endogenous biomarker, Cys-C has many of the properties required for a good marker of renal function. Compared with that of creatinine, the plasma concentration of Cys-C is less affected by external factors[28]. Clinically, Cys-C is more accurate in assessing actual renal function because muscle loss does not appear to affect it[27,29]. A Cys-C level above normal for age indicates underlying renal problems. Additionally, Cys-C levels have significant diagnostic value in assessing renal injury in hyperuricaemic populations[30]. Guo Y's study suggested that Cys-C levels were significantly higher in the hyperuricaemic population than in healthy controls. At the same hyperuricaemic level, there was a corresponding increase in the risk of renal damage with increasing Cys-C concentrations[31]. This study revealed that, in the uricemihyperuricaemic population, the risk of developing renal damage

increased 4.21-fold for every 1 mg increase in the Cys-C concentration to 1.33 mg/L. In comparison, the risk of developing renal damage was not statistically significant after exceeding 1.33 mg/L, and the risk of developing the disease showed a smooth trend. Previous studies have shown that in hyperuricaemic individuals, an increase in the Cys-C level is indeed correlated with chronic renal failure (CRF) but not with an increase in the Cys-C concentration; thus, the risk must be significantly increased. However, there is a threshold point, and the trend tends to stabilize after exceeding the threshold.

Strengths and limitations of the study

Our research has several distinct advantages over previous studies in this field. First, we focused specifically on hyperuricaemic populations and employed a comprehensive machine learning approach, utilizing three distinct algorithms (random forest, LASSO regression, and XGBoost) to ensure robust feature selection. Second, our model showed good discriminative ability, with an AUC of 0.82, and underwent thorough validation through calibration tests and decision curve analyses. Third, we performed detailed dose–response relationship analyses using RCS curves, which revealed important threshold effects for key predictors.

However, several limitations warrant careful consideration. While our model achieved reasonable performance, with an AUC of 0.82, there remains substantial room for improvement. Future studies could enhance predictive accuracy through multiple approaches: (a) incorporating additional relevant biomarkers and clinical parameters, such as inflammatory markers and detailed medication history; (b) implementing more sophisticated deep learning architectures; (c) considering interaction terms between variables; and (d) incorporating longitudinal data to develop dynamic prediction tools. Furthermore, as our demographic data were collected from a single centre in Eastern China, the generalizability of the model across different ethnic groups and geographical regions may be limited. Multicentre validation studies would be necessary to assess the model's performance in diverse health care settings. This study relied primarily on conventional laboratory indicators, while potential novel biomarkers and environmental factors that might influence renal damage in hyperuricaemic populations were not included. Additionally, the model's applicability might be limited to specific subgroups, such as patients with acute kidney injury or those undergoing dialysis, as these populations were excluded from our study. Finally, environmental factors, dietary habits, and lifestyle differences across populations could influence the model's performance, suggesting the need for external validation in diverse settings.

Conclusion

Our study identified age, Cys-C, UA, and sex as key predictors of renal impairment in hyperuricaemic patients. Clinical management should focus on controlling Cys-C and UA levels to reduce the risk of renal impairment. The observed sex differences highlight the need for sex-specific treatment approaches, while the threshold effects of Cys-C and age on renal function provide important guidance for clinical decision-making.

Data availability

The original data for this study can be found in the Supplementary Files.

Received: 31 October 2024; Accepted: 29 January 2025
Published online: 15 March 2025

References

1. Angeli, P., Garcia-Tsao, G., Nadim, M. K. & Parikh, C. R. News in pathophysiology, definition, and classification of hepatorenal syndrome: A step beyond the International Club of ascites (ICA) consensus document. *J. Hepatol.* **71**(4), 811–822 (2019).
2. Alam, A. B., Wu, A., Power, M. C., West, N. A. & Alonso, A. Associations of serum uric acid with incident dementia and cognitive decline in the ARIC-NCS cohort. *J. Neurol. Sci.* **414**, 116866 (2020).
3. Qiao, M., Chen, C., Liang, Y., Luo, Y. & Wu, W. The influence of serum uric acid level on Alzheimer's Disease: A narrative review. *Biomed. Res. Int.* **2021**, 5525710 (2021).
4. Aiumtrakul, N., Wiputhanuphongs, P., Supasyndh, O. & Satirapoj, B. Hyperuricemia and impaired renal function: A prospective cohort study. *Kidney Dis.* **7**(3), 210–218 (2021).
5. Zhang, M. et al. Prevalence of hyperuricemia among Chinese adults: Findings from two nationally representative cross-sectional surveys in 2015-16 and 2018-19. *Front. Immunol.* **12**, 791983 (2022).
6. Borghi, C. et al. Expert consensus for the diagnosis and treatment of patient with hyperuricemia and high cardiovascular risk: 2021 update. *Cardiol. J.* **28**(1), 1–14 (2021).
7. Beunders, R. et al. Assessing GFR with proenkephalin. *Kidney Int. Rep.* **8**(11), 2345–2355 (2023).
8. Noghrehchi, F., Stoklosa, J., Penev, S. & Warton, D. I. Selecting the model for multiple imputation of missing data: Just use an IC! *Stat. Med.* **40**(10), 2467–2497 (2021).
9. Moons, K. G. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **162**(1), W1–W73 (2015).
10. Carracedo-Reboredo, P. et al. A review on machine learning approaches and trends in drug discovery. *Comput. Struct. Biotechnol. J.* **19**, 4538–4558 (2021).
11. Dunne, R. et al. Thresholding Gini variable importance with a single-trained random forest: An empirical Bayes approach. *Comput. Struct. Biotechnol. J.* **21**, 4354–4360 (2023).
12. Gao, L. & Ding, Y. Disease prediction via bayesian hyperparameter optimization and ensemble learning. *BMC Res. Notes* **13**(1), 205 (2020).
13. Kim, J. H. Multicollinearity and misleading statistical results. *Korean J. Anesthesiol.* **72**(6), 558–569 (2019).
14. Yang, W., Jiang, J., Schnellinger, E. M., Kimmel, S. E. & Guo, W. Modified Brier score for evaluating prediction accuracy for binary outcomes. *Stat. Methods Med. Res.* **31**(12), 2287–2296 (2022).
15. Copur, S., Demiray, A. & Kanbay, M. Uric acid in metabolic syndrome: Does uric acid have a definitive role? *Eur. J. Intern. Med.* **103**, 4–12 (2022).
16. Zhou, F. et al. Association of serum uric acid levels with the incident of kidney disease and rapid eGFR decline in Chinese individuals with eGFR > 60 mL/min/1.73 m2 and negative proteinuria. *Clin. Exp. Nephrol.* **23**(7), 871–879 (2019).

17. Wu, S. et al. Serum uric acid levels and health outcomes in CKD: A prospective cohort study. *Nephrol. Dial. Transpl.* **39**(3), 510–519 (2024).
18. Sato, Y. et al. The case for uric acid-lowering treatment in patients with hyperuricemia and CKD. *Nat. Rev. Nephrol.* **15**(12), 767–775 (2019).
19. Jordan, D. M. et al. No causal effects of serum urate levels on the risk of chronic kidney disease: A Mendelian randomization study. *PLoS Med.* **16**(1), e1002725 (2019).
20. Wu, S., Kong, M., Song, Y. & Peng, A. Ethnic disparities in bidirectional causal effects between serum uric acid concentrations and kidney function: trans-ethnic mendelian randomization study. *Heliyon* **9**(11), e21108 (2023).
21. Sellmayr, M. et al. Only hyperuricemia with crystalluria, but not asymptomatic hyperuricemia, drives progression of chronic kidney disease. *J. Am. Soc. Nephrol.* **31**(12), 2773–2792 (2020).
22. Badve, S. V. et al. Effects of allopurinol on the progression of chronic kidney disease. *N. Engl. J. Med.* **382**(26), 2504–2513 (2020).
23. Oh, T. R. et al. Hyperuricemia has increased the risk of progression of chronic kidney disease: Propensity score matching analysis from the KNOW-CKD study. *Sci. Rep.* **9**(1), 6681 (2019).
24. Li, Y. et al. Effect modification of hyperuricemia, cardiovascular risk, and age on chronic kidney disease in China: A cross-sectional study based on the China Health and Nutrition Survey Cohort. *Front. Cardiovasc. Med.* **9**, 853917 (2022).
25. Luo, Y. et al. Serum uric acid levels and their association with renal function decline and kidney disease progression in middle-aged and elderly populations: A retrospective cohort study. *J. Multidiscip. Healthc.* **16**, 3357–3366 (2023).
26. Lassus, J. & Harjola, V. P. Cystatin C: A step forward in assessing kidney function and cardiovascular risk. *Heart Fail. Rev.* **17**(2), 251–261 (2012).
27. Haines, R. W. et al. Comparison of cystatin C and creatinine in the assessment of measured kidney function during critical illness. *Clin. J. Am. Soc. Nephrol.* **18**(8), 997–1005 (2023).
28. Onopiuk, A., Tokarzewicz, A. & Gorodkiewicz, E. Cystatin C: A kidney function biomarker. *Adv. Clin. Chem.* **68**, 57–69 (2015).
29. Pottel, H., Delanaye, P. & Cavalier, E. Exploring renal function assessment: Creatinine, cystatin C, and estimated glomerular filtration rate focused on the European kidney function Consortium equation. *Ann. Lab. Med.* **44**(2), 135–143 (2024).
30. Wu, Y., Wang, S. & Xu, X. Correlation of serum cystatin C with renal function in gout patients with renal Injury. *J. Interferon Cytokine Res.* **41**(9), 329–335 (2021).
31. Guo, Y., Huang, H., Chen, Y., Shen, C. & Xu, C. Association between circulating cystatin C and hyperuricemia: A cross-sectional study. *Clin. Rheumatol.* **41**(7), 2143–2151 (2022).

Acknowledgements

The authors gratefully acknowledge the support provided by the families of the main authors (Wenquan Wu, Yue Dai, and Yiyang Ding) during the course of this research.

Author contributions

Study design: W.K. and T.W.; Data analysis and interpretation: T.W., W.K., H.Y., J.C.; Important knowledge content: T.W., W.K.; Figure preparation and organization: H.Y., T.W.; Critical modification: W.K., H.Y.; Manuscript revision: H.Y., T.W., W.K.; T.W. and H.Y. contributed equally to this work. W.K. and T.W. are the guarantors of this work, so they can access all the data in the study fully and are responsible for the integrity of the data and the accuracy of the data analysis. All authors have read and approved the final manuscript.

Funding

This work was supported by the Nanjing Medical Science and Technology Development Special Fund Project (Grant No. YKK20170), the Ningwei Traditional Chinese Medicine project (No. 6, 2024), the Natural Science Research of Jiangsu Higher Education Institutions of China (No. 24KJB360012) and the National Natural Science Foundation of China (No. 81904112).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-88632-x.

Correspondence and requests for materials should be addressed to W.K.