Research Square

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

Association Between Urinary Cadmium Levels and Tubular Injury Biomarkers with Development of a Machine Learning-Based Predictive Model in Non-Dialysis Patients with Chronic Kidney Disease: a retrospective study

xi chen
Yangpu Hospital, Tongji University School of Medicine
tian tang
Yangpu Hospital, Tongji University School of Medicine
yuxue liu
Yangpu Hospital, Tongji University School of Medicine
rong zhou
zhourongyp@126.com
Yangpu Hospital, Tongji University School of Medicine
xuejing guan
Yangpu Hospital, Tongji University School of Medicine

Research Article

Keywords: Chronic kidney disease, Urinary cadmium, Tubular injury biomarkers, Prediction model, Nomogram

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10584989/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Cadmium is a widely distributed nephrotoxic environmental pollutant, and long-term exposure can cause injury to renal tubular epithelial cells. Non-dialysis patients with chronic kidney disease (CKD) may be particularly vulnerable to cadmium nephrotoxicity due to reduced renal functional reserve. However, an individualised tool for assessing the risk of elevated urinary cadmium exposure remains unavailable for the non-dialysis CKD population. This study aimed to characterise the expression profile of tubular injury biomarkers in non-dialysis CKD patients with high urinary cadmium exposure, to identify independent predictors through the integration of machine learning and non-linear regression methods, and to construct a visualised clinical risk nomogram that could facilitate early detection of cadmium-related tubular injury and stratified management of high-risk individuals.

Methods

This retrospective study enrolled 278 patients with stage 1–5 non-dialysis CKD from Yangpu Hospital, Tongji University, between 1 January 2015 and 31 October 2022. Patients were divided into low-exposure (n = 180) and high-exposure (n = 98) groups according to the cut-off of 2 μg/g·Cr for urinary cadmium. Clinical data were collected, and urinary biomarkers including β_2-microglobulin (β_2-MG), retinol-binding protein (RBP), N-acetyl-β-D-glucosaminidase (NAG), α_1-microglobulin (α_1-MG) and microalbumin (U_mALB) were measured. Spearman rank correlation was used to assess inter-variable associations. Restricted cubic splines (RCS) were applied to evaluate dose–response relationships between continuous variables and high urinary cadmium exposure. LASSO regression was employed for variable selection, followed by multivariable logistic regression to construct the prediction model. Model discrimination was assessed using the area under the receiver operating characteristic curve (AUC), calibration was evaluated using the Hosmer–Lemeshow test and calibration plots, and clinical utility was assessed using decision curve analysis (DCA). Internal validation was performed using the bootstrap method with 1,000 resamples.

Results

The proportion of patients with diabetes was significantly higher in the high-exposure group than in the low-exposure group (42.86% vs. 28.33%, P = 0.02). Urinary levels of β_2-MG, RBP, NAG and U_mALB were all significantly elevated in the high-exposure group (all P < 0.001), whereas no statistically significant differences were observed between the two groups in terms of GFR, cystatin C, sex or smoking index (P > 0.05). RCS analyses demonstrated that NAG (P-overall < 0.001, P-nonlinear < 0.001) and RBP (P-overall < 0.001, P-nonlinear = 0.069) were significantly and positively associated with the risk of high urinary cadmium exposure, with NAG showing a non-linear saturation effect and RBP exhibiting an approximately linear relationship. β_2-MG displayed a J-shaped association (P-overall = 0.021, P-nonlinear

= 0.038). LASSO regression identified four predictive variables: T2DM, RBP, GFR and NAG. Multivariable logistic regression confirmed that NAG positivity (OR = 3.88, 95% CI 1.99–7.56, P < 0.001), T2DM (OR = 1.81, 95% CI 1.00–3.28, P = 0.050), RBP (OR = 1.03, 95% CI 1.01–1.06, P = 0.007) and GFR (OR = 1.01, 95% CI 1.00–1.02, P = 0.004) were independently associated with high urinary cadmium exposure, and a nomogram was constructed accordingly. The model achieved an AUC of 0.77 (95% CI 0.71–0.83). The Hosmer–Lemeshow test yielded a P value of 0.989, with a Brier score of 0.18, and the calibration curve demonstrated good agreement between predicted and observed probabilities. DCA showed a positive net benefit across threshold probabilities ranging from 0.05 to 0.82.

Conclusions

Elevated urinary cadmium levels in non-dialysis CKD patients are significantly associated with tubular injury biomarkers, among which NAG and RBP are effective indicators for identifying high-risk cadmium exposure, with distinct dose–response patterns. Patients with comorbid diabetes may represent a susceptible subgroup to cadmium nephrotoxicity. The prediction model developed in this study demonstrates satisfactory calibration and moderate discrimination, and may serve as a useful reference for early screening of high-risk individuals. However, further external validation in large-scale cohort studies is warranted before its clinical implementation.

1. Introduction

Cadmium (Cd) is a nephrotoxic environmental pollutant. Following dietary or environmental exposure, approximately 90% of absorbed cadmium accumulates in the renal proximal tubules, where its biological half-life extends from 10 to 30 years, posing a substantial global health concern[1, 2]. Soil and water contamination with cadmium remains a serious issue in parts of China, and chronic low-level exposure is prevalent among the general population[3]. The kidney is the primary target organ of cadmium toxicity. Epidemiological evidence has linked low-level cadmium exposure to an increased risk of chronic kidney disease (CKD), and cadmium is now recognised as a risk factor for CKD[4–6]. Chronic exposure predominantly affects the proximal tubules[3], impairing reabsorptive capacity. Early manifestations include low-molecular-weight proteinuria and enzymuria, while advanced injury may progress to irreversible interstitial fibrosis and end-stage renal disease. Because cadmium-induced kidney damage is clinically silent in its early stages, significant renal dysfunction is often already irreversible by the time of detection. Thus, identifying sensitive biomarkers that reflect early tubular injury and establishing valid risk prediction tools are of considerable public health and clinical importance.

Renal biopsy remains the reference standard for evaluating tubular injury, but its invasive nature precludes widespread screening application. Consequently, non-invasive urinary biomarkers have attracted growing interest. These biomarkers are broadly classified into two categories. The first includes tubular lysosomal enzymes, with N-acetyl-β-D-glucosaminidase (NAG) as the representative marker, which directly indicates tubular epithelial cell injury. The second comprises low-molecular-weight

proteins that are reabsorbed by the proximal tubules, with retinol-binding protein (RBP) and β_2-microglobulin (β_2-MG) as key examples, which reflect impaired proximal reabsorptive function. Accumulating evidence from epidemiological and toxicological studies has consistently shown that cadmium exposure significantly increases the urinary excretion of these markers, often preceding changes in conventional parameters such as serum creatinine or estimated glomerular filtration rate (eGFR). These biomarkers are therefore widely regarded as early indicators of cadmium-induced proximal tubular damage[7–9]. However, most existing studies have relied on single-marker association analyses[10, 11], with limited systematic evaluation of multiple markers, insufficient consideration of potential non-linear dose–response relationships, and a lack of clinically applicable risk prediction tools.

Machine learning methods offer distinct advantages over conventional statistical approaches in handling complex datasets, capturing non-linear associations, and managing high-dimensional data[12, 13]. In recent years, these techniques have been increasingly adopted in clinical research[14] and have shown promising predictive performance in kidney diseases, including acute kidney injury and CKD[15, 16]. LASSO regression enables variable shrinkage and model simplification by penalising redundant predictors, while restricted cubic splines (RCS) provide a flexible framework for visualising non-linear relationships between continuous variables and binary outcomes. The combined use of these methods represents a robust analytical strategy for variable selection and dose–response assessment in environmental epidemiology. To date, however, no study has integrated these approaches to systematically compare the predictive utility of multiple tubular biomarkers for high urinary cadmium exposure in the non-dialysis CKD population. Applying machine learning to biomarker research in cadmium-induced tubular injury may overcome the limitations of traditional methods and offer more reliable evidence to inform clinical practice.

To address these gaps, we conducted a multi-dimensional analysis in a cohort of 278 non-dialysis CKD patients, defining high cadmium exposure as a corrected urinary cadmium level ≥ 2 μg/g·Cr. The specific aims of this study were: (1) to compare baseline demographic, clinical, and tubular biomarker profiles between high- and low-cadmium exposure groups, (2) to use LASSO regression to identify key predictive biomarkers for high cadmium exposure, (3) to apply RCS models to characterise the non-linear dose–response relationships between candidate biomarkers and high cadmium exposure, and (4) to develop and internally validate a nomogram based on independent predictors, with the goal of providing a practical tool for early screening and risk stratification of cadmium-related tubular injury.

2. Methods

2.1 Study population

2.1.1 Study design and ethics

This was a single-centre retrospective observational study. Data were obtained from non-dialysis CKD patients who attended Yangpu Hospital, affiliated with Tongji University, between 1 January 2019 and 31 October 2022. The study protocol was approved by the Ethics Committee of Yangpu Hospital Affiliated to Tongji University (approval number: LL-2026-LW-014), and all procedures were performed in accordance with the Declaration of Helsinki. Due to the retrospective nature of the study, the requirement for informed consent was waived by the Ethics Committee.

2.1.2 Inclusion criteria

 Age ≥ 18 years, diagnosis of CKD stage 1–5 according to the KDIGO 2021 guideline, availability of concurrent measurements of corrected urinary cadmium, a complete panel of corrected urinary tubular biomarkers, serum cystatin C, as well as documented baseline medical history and smoking information.

2.1.3 Exclusion criteria

 History of acute kidney injury or nephrotoxic drug exposure within the preceding 3 months, undergoing maintenance haemodialysis or peritoneal dialysis.

A total of 300 patients were initially screened, 22 were excluded based on the above criteria, yielding a final cohort of 278 patients. Referring to published epidemiological cut-off values for cadmium exposure[17–19], we defined high cadmium exposure as a corrected urinary cadmium concentration ≥ 2 μg/g·Cr and low exposure as < 2 μg/g·Cr. Accordingly, patients were divided into low-exposure (n = 180) and high-exposure (n = 98) groups.

2.2 Clinical data and laboratory measurements

Baseline demographic and clinical data were collected, including sex, age, smoking index (calculated as daily cigarette consumption * years of smoking), and comorbidities of diabetes and hypertension. Renal function parameters included estimated glomerular filtration rate (eGFR) and serum cystatin C (CysC).

All urinary biomarkers were corrected by urinary creatinine and are referred to by their standard abbreviations throughout the text: β_2-microglobulin (β_2-MG), retinol-binding protein (RBP), N-acetyl-β-D-glucosaminidase (NAG), α_1-microglobulin (α_1-MG) and microalbumin (U_mALB).

2.3 Statistical analysis

2.3.1 Baseline comparisons

Categorical variables were presented as counts (percentages) and compared between groups using the Pearson chi-square test. Continuous variables were tested for normality using the Shapiro–Wilk test, all were found to be non-normally distributed and were therefore expressed as medians with interquartile ranges, with between-group differences assessed by the Wilcoxon rank-sum test.

2.3.2 Correlation analysis

Spearman's rank correlation coefficients were calculated separately for the low- and high-cadmium groups to evaluate pairwise associations among continuous variables and the ordinal variable (CKD stage). This analysis was used to identify potential multicollinearity (e.g., between smoking status and smoking index) to inform subsequent modelling.

2.3.3 Dose–response relationship assessment

Restricted cubic spline (RCS) regression models were used to evaluate the dose–response relationships between continuous variables and the outcome (high-cadmium group). Following Harrell's recommendations, four knots were placed at the 5th, 35th, 65th and 95th percentiles of each variable's distribution. The median value of each variable served as the reference (OR = 1.0). Odds ratios (ORs) with 95% confidence intervals (CIs) were calculated across the range of each variable. Both the overall association (P-overall) and non-linearity (P-nonlinear) were reported.

2.3.4 Predictor screening

LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression was applied for variable shrinkage and selection. The optimal penalty parameter λ was determined via 10-fold cross-validation.

2.3.5 Prediction model development and evaluation

Multivariable logistic regression and nomogram

Variables selected by LASSO were entered into a multivariable binary logistic regression model using the enter method. Regression coefficients (β), Wald statistics, ORs and 95% CIs were computed. A nomogram was constructed based on the regression coefficients to visualise individualised risk prediction. Discriminative ability was assessed using the receiver operating characteristic (ROC) curve, with the area under the curve (AUC) and its 95% CI reported. Internal validation was performed using the bootstrap method with 1,000 resamples, and the optimism-corrected C-index was calculated. The optimal cut-off value was determined using the Youden index, with corresponding sensitivity and specificity reported. Decision curve analysis (DCA) was performed to evaluate the net benefit of the model across a range of threshold probabilities, comparing it with "treat-all" and "treat-none" strategies. A clinical impact curve (CIC) was also plotted to illustrate the number of individuals classified as high-risk at different risk thresholds alongside the corresponding number of true positives.

2.3.6 Missing data handling

For continuous variables with missing values, median imputation was applied for Spearman correlation analyses to preserve sample size. For LASSO regression and multivariable logistic modelling, complete case analysis was adopted, with the effective sample size reported in the results.

2.3.7 Software

All statistical analyses and graphical presentations were performed using R software (version 4.2.3, R Foundation for Statistical Computing, Vienna, Austria). Logistic regression and basic statistical tests

were conducted using the stats package, LASSO regression and cross-validation using the glmnet package, restricted cubic splines and regression modelling using the rms package, nomogram construction and visualisation using the nomogram package, graphical output primarily using the ggplot2 package, and data cleaning and processing using the dplyr and tidyr packages. All tests were two-sided, and a P value < 0.05 was considered statistically significant.

3. Results

3.1 Baseline Characteristics

This study included 278 patients with chronic kidney disease (CKD), who were then assigned to low-cadmium (n = 180) or high-cadmium (n = 98) groups according to the median urinary cadmium concentration. The two cohorts were largely comparable across most clinical parameters, including CKD stage, estimated glomerular filtration rate (GFR), age, sex, hypertension prevalence, smoking index, and CysC levels (all P > 0.05). A notable exception was T2DM, whose prevalence was markedly higher in the high-cadmium group (42.86% vs. 28.33%, P = 0.02). For urinary markers, the high-cadmium group had pronounced elevations in β_2-MG, RBP, NAG, and U_mALB (all P < 0.05), whereas α_1-MG levels did not differ appreciably between groups (P = 0.13). (Table 1).

Table 1
Comparison of baseline characteristics of study participants stratified by urinary cadmium levels (low-cadmium group vs high-cadmium group)

Variable		Total(n = 278)	Lowgroup(n = 180)	Highgroup(n = 98)	statistics	p-value
Age		64.00 (54.25–70.00)	64.00 (54.00–72.00)	64.00 (55.00–69.00)	0.57	0.57
Sex(male)		151 (54.32%)	99 (55.00%)	52 (53.06%)	0.03	0.85
CKD_stage	1	76 (27.34%)	46 (25.56%)	30 (30.61%)	1.98	0.74
	2	54 (19.42%)	39 (21.67%)	15 (15.31%)		
	3	80 (28.78%)	51 (28.33%)	29 (29.59%)		
	4	35 (12.59%)	23 (12.78%)	12 (12.24%)		
	5	33 (11.87%)	21 (11.67%)	12 (12.24%)		
GFR		56.62 (31.06–92.61)	56.83 (30.67–90.70)	56.37 (32.61–93.80)	-0.22	0.83
T2DM	1	93 (33.45%)	51 (28.33%)	42 (42.86%)	5.38	**0.02**
Hypertension	1	199 (71.58%)	130 (72.22%)	69 (70.41%)	0.03	0.86
Smoking_index		0.00 (0.00–200.00)	0.00 (0.00–200.00)	0.00 (0.00–200.00)	0.14	0.87
Cystatin_C		1.48 (1.06–2.24)	1.51 (1.08–2.26)	1.38 (1.01–2.16)	1.18	0.24
Adjusted urinary biomarkers						
β_2_MG		0.84 (0.28–10.94)	0.57 (0.23–5.77)	2.00 (0.44–14.84)	-3.17	**< 0.001**
RBP		5.26 (1.08–14.24)	2.76 (0.70–10.49)	11.35 (4.21–25.09)	-4.99	**< 0.001**
NAG		28.02 (16.26–45.37)	21.63 (13.57–35.00)	36.81 (26.81–58.39)	-5.82	**< 0.001**
α_1_MG		31.82 (13.47–87.77)	29.07 (12.17–79.23)	36.28 (17.66–96.72)	-1.53	0.13

Variable	Total(n = 278)	Lowgroup(n = 180)	Highgroup(n = 98)	statistics	p-value
U_mALB	72.92 (12.73–224.27)	58.62 (10.61–190.86)	104.99 (25.40–319.19)	-2.68	**0.01**

3.2 Correlation Analysis

Spearman correlation analyses revealed several consistent relationships irrespective of group assignment. CKD stage, as expected, tightly correlated with GFR (r = − 0.97, P < 0.001) and CysC (r = 0.76–0.79, P < 0.001). The tubular injury markers: β_2-MG, α_1-MG, NAG, and RBP were moderately to strongly intercorrelated within both cohorts (r = 0.39–0.83, all P < 0.001), supporting their coherence as a composite panel for tubular assessment. Also, smoking status and smoking index proved nearly redundant (r = 0.97, P < 0.001). Between-group discrepancies emerged chiefly for the weak age- CysC and age- β_2-MG associations, which surfaced only in the low-cadmium arm, conversely, the moderate sex–smoking correlations attained significance exclusively in the high-cadmium arm. (Fig. 1a-d).

3.3 Restricted Cubic Spline (RCS) Analysis

We next fitted restricted cubic spline models to explore dose–response patterns linking continuous variables to high-cadmium group membership (Fig. 2a–i). After covariate adjustment, both NAG (P-overall < 0.001, P-nonlinear < 0.001) and RBP (P-overall < 0.001, P-nonlinear = 0.069) showed positive associations with the outcome: NAG in a non-linear fashion marked by a saturation threshold, and RBP in a roughly linear manner. β_2-MG exhibited a J-shaped non-linear relationship (P-overall = 0.021, P-nonlinear = 0.038). By contrast, the overall tests for GFR, age, smoking index, CysC, α_1-MG, and U_mALB all fell short of statistical significance (P-overall > 0.05).

3.4 LASSO Regression for Variable Selection

To pare down the candidate predictors, LASSO logistic regression with 10-fold cross-validation was applied (Fig. 3a-b). At λ.min (0.03), four predictors retained non-zero coefficients: T2DM, RBP, GFR, and NAG. Under the more conservative λ.1se (0.09), RBP stood alone as the sole retained variable. The remaining candidates: hypertension, age, sex, smoking, β_2-MG, α_1-MG, and U_mALB—had their coefficients compressed to zero, implying negligible independent predictive value for the outcome.

3.5 Prediction Model Development and Performance Evaluation

3.5.1 Multivariable Logistic Regression Model

Drawing on the LASSO outcomes and the non-linear patterns from RCS, we dichotomized NAG at its median for multivariable analysis. The final logistic model, which incorporated T2DM, RBP, GFR, and

binary NAG, indicated that each variable maintained an independent association with the outcome. (Table 2).

NAG positivity (OR = 3.88, 95% CI: 1.99–7.56, P < 0.001) and T2DM (OR = 1.81, 95% CI: 1.00–3.28, P = 0.050) conferred elevated risks, whereas RBP (OR = 1.03, 95% CI: 1.01–1.06, P = 0.007) and GFR (OR = 1.01, 95% CI: 1.00–1.02, P = 0.004) yielded modest but statistically significant effect estimates. A nomogram based on these regression coefficients was constructed for visualization (Fig. 4).

Table 2
Multivariable logistic regression analysis

Variable	β coefficient	Wald Z	P-value	OR (95% CI)
NAG (positive vs. negative)	1.3556	3.98	< 0.001	3.88 (1.99–7.56)
GFR	0.0112	2.89	0.004	1.01 (1.00–1.02)
RBP	0.0335	2.68	0.007	1.03 (1.01–1.06)
T2DM (yes vs. no)	0.5942	1.96	0.050	1.81 (1.00–3.28)

3.5.2 Model Discrimination and Calibration

The model yielded an AUC of 0.77 (95% CI: 0.71–0.83) and a C-index of 0.767, reflecting moderate discriminative power. (Fig. 5-a). Calibration curves showed close alignment between predicted and observed risks, corroborated by a Hosmer–Lemeshow test ($\chi^2 = 1.696$, P = 0.989) and a Brier score of 0.18 both indicative of excellent calibration. (Fig. 5-b). Internal validation via 1,000 bootstrap resamples confirmed the model's stability.

3.5.3 Clinical Utility

Decision curve analysis demonstrated that our model afforded a superior net benefit over both "treat-all" and "treat-none" strategies across threshold probabilities from 0.05 to 0.82. (Fig. 6-a). This clinical applicability was further supported by the clinical impact curve (Fig. 6-b). The Youden index identified an optimal cut-off of 0.46, at which sensitivity and specificity reached 54.0% and 88.0%, respectively.

Decision curve analysis (DCA) for the prediction model (a), Clinical impact curve (CIC) for the prediction model (b)

4. Discussion

In this single-centre cohort of 278 patients with CKD, we applied multiple statistical modelling strategies to systematically investigate the relationship between urinary cadmium exposure and tubular injury biomarkers, and constructed a visualised nomogram prediction model. We found that patients in the high-urinary-cadmium group (≥ 2 μg/g·Cr) had significantly elevated levels of urinary β_2-MG, RBP, NAG

and microalbumin, and these associations remained robust after adjusting for confounders. Restricted cubic spline analysis further revealed a non-linear saturation effect for NAG and a J-shaped risk curve for β_2-MG, while LASSO-Logistic regression identified NAG as the strongest independent predictor of cadmium-related injury in this high-risk CKD population (OR = 3.88). The model demonstrated good calibration (Hosmer–Lemeshow P = 0.989) and moderate discrimination (AUC = 0.77), with a positive net clinical benefit across a wide range of thresholds.

4.1 Association of cadmium exposure with tubular injury biomarkers

Cadmium is a well-recognised nephrotoxic environmental pollutant. Following intestinal absorption, it accumulates in proximal tubular epithelial cells, with a half-life of 10–30 years. Intracellular cadmium can damage tubular cells through multiple mechanisms, including disruption of the mitochondrial respiratory chain, induction of oxidative stress, and activation of apoptotic signalling pathways[20]. Tubular injury leads to impaired reabsorption of low-molecular-weight proteins (such as β_2-MG and RBP), resulting in increased urinary excretion of these proteins — the classic pattern of "tubular proteinuria." Our observation that RBP, NAG and β_2-MG levels were significantly higher in the high-cadmium group is consistent with the findings of Zhang et al. in Chinese population-based studies[21]. Notably, we found that tubular markers (RBP, NAG, β_2-MG) and the glomerular marker (U_mALB) were simultaneously elevated in the high-cadmium group, with a moderate positive correlation between the two sets of markers (r = 0.48–0.50). This supports the "glomerulo-tubular crosstalk" hypothesis — whereby severe tubulointerstitial injury may secondarily impair the glomerular filtration barrier through mechanisms such as tubuloglomerular feedback and activation of the renin–angiotensin system[22]. This finding suggests that in CKD patients with cadmium exposure, tubular injury should not be viewed in isolation, its synergistic interaction with glomerular damage may be an important driver of disease progression.

4.2 Differential dose–response patterns of biomarkers

In this study, RCS analysis revealed heterogeneous association patterns between the three biomarkers and cadmium exposure risk, which may have implications for clinical monitoring strategies.

RBP demonstrated a robust linear increasing effect. RCS analysis showed a strictly linear positive association between RBP and the outcome (P-overall < 0.001, P-nonlinear = 0.069), with OR values increasing progressively across the RBP concentration range without significant inflection or saturation. RBP is a low-molecular-weight carrier protein (~ 21 kDa) that is almost completely reabsorbed in the proximal tubules via the megalin/cubilin complex after glomerular filtration. Urinary RBP elevation directly and linearly reflects the extent of impaired proximal reabsorptive function, which is proportional to the cadmium burden accumulated in proximal tubular lysosomes, making RBP a sensitive marker of cadmium exposure[23, 24]. The linear feature of RBP supports its use as an ideal continuous quantitative indicator of cadmium-induced proximal tubular dysfunction.

NAG, in contrast, showed a different pattern. NAG is a high-molecular-weight enzyme located in proximal tubular lysosomes that does not readily cross the glomerular filtration barrier under physiological conditions[25]. When cadmium accumulation disrupts lysosomal membrane integrity or causes tubular epithelial cell necrosis, large amounts of NAG are released into the urine, thus, elevated urinary NAG reflects actual tubular cellular injury[9]. RCS analysis showed a significant non-linear positive association between NAG and the outcome (P-overall < 0.001, P-nonlinear < 0.001), with the reference point at x = 27.85, a rapid rise in OR at lower concentrations, and gradual saturation at higher concentrations. When NAG was included as a continuous variable in the multivariable logistic regression, it did not reach statistical significance (P = 0.221), however, after conversion to a binary variable, the regression coefficient became significant (β = 1.13, P = 0.0004). This apparent paradox suggests that the predictive effect of NAG for high cadmium status is not a simple linear increase, but rather involves a "threshold effect" — where the increase in NAG translates into a significantly elevated risk only after tubular injury accumulates beyond a certain critical level, consistent with findings from Muneko Nishijo et al[26]. The saturation trend observed in the RCS curve (slowing of OR increase at high concentrations) further supports this interpretation. These non-linear characteristics suggest that clinical interpretation of NAG levels should not rely solely on absolute values, but rather on appropriate cut-offs based on population characteristics to fully realise its early warning value.

β_2-MG exhibited a J-shaped association (P-nonlinear = 0.038), with both the lowest and highest quantiles associated with increased risk of high cadmium exposure. This counter-intuitive "low-level risk" phenomenon is consistent with the findings of Sakurai et al. in Japanese populations[27], and suggests that clinical interpretation of β_2-MG should consider GFR stratification rather than being viewed in isolation.

4.3 Interaction between diabetes and cadmium exposure: identification of a high-risk subgroup

In our baseline analysis, the proportion of patients with diabetes was significantly higher in the high-cadmium group (42.86% vs. 28.33%, P = 0.02). Diabetes remained an independent predictor in the multivariable logistic regression (OR = 1.81, P = 0.050), and was retained under the λ.min criterion in LASSO but excluded under the λ.1se criterion, reflecting its borderline predictive contribution. The nomogram also assigned a certain score to diabetes. This finding has important clinical and public health implications. Experimental studies have suggested that diabetic rats are more susceptible to cadmium nephrotoxicity, and that cadmium exposure may exacerbate diabetic kidney injury through inhibition of autophagy[28]. A population-based study by Jin et al. systematically reported a positive association between urinary cadmium and the risk of diabetic kidney disease (DKD) in patients with diabetes[29]. We speculate that patients with diabetes, due to underlying glucose metabolism disorders, enhanced oxidative stress, and glomerular hyperfiltration, may have higher renal susceptibility to environmental toxins. Therefore, for CKD patients with comorbid diabetes, greater attention should be given to the assessment and control of environmental cadmium exposure.

4.4 Methodological advantages of machine learning-assisted feature selection

The stepwise strategy adopted in this study- LASSO regression for screening, RCS for non-linear characterisation, and logistic regression for interpretable modelling- partially overcomes the limitations of any single method. LASSO introduces an L1 penalty term into the loss function to shrink some regression coefficients to zero, enabling simultaneous variable selection and parameter estimation with inherent tolerance to multicollinearity [13]. In our study, Spearman correlation analysis confirmed high collinearity between smoking status and smoking index (r = 0.97), as well as moderate to strong correlations among tubular markers (r = 0.39–0.83). In such highly collinear settings, traditional stepwise regression tends to yield unstable models, whereas LASSO stably shrinks redundant variable coefficients to zero. The discrepancy between the λ.min and λ.1se criteria (four vs. one variable) reflects the inherent trade-off between predictive accuracy and model parsimony, reporting both results helps readers assess the relative stability of each predictor.

4.5 Limitations

Several limitations of this study should be acknowledged. First, the cross-sectional design precludes causal inference- urinary cadmium and tubular markers were measured simultaneously, and the reabsorptive impairment inherent to CKD may itself affect urinary cadmium excretion. Causal relationships require prospective cohort studies for confirmation. Second, a single urinary cadmium measurement may not fully capture an individual's long-term cumulative exposure, and we did not adjust for dietary intake, soil contamination, or coexposure to lead/arsenic, leaving the possibility of unmeasured confounding. Third, the complete case analysis included only 257 patients (87 events), with an EPV of approximately 21.7- a small-to-moderate sample size that may introduce optimism bias in the model coefficients. RBP data were missing for 21 patients (7.6%) and were handled by listwise deletion, if the missingness was not random, this could have introduced bias. The model also did not account for medication use, CKD aetiology, or diabetes duration. Fourth, the model AUC was 0.77 (< 0.80), indicating moderate discrimination, and all evaluations were based on the training set — bootstrap internal validation cannot replace independent external validation, which represents the major methodological shortcoming. The single-centre design also limits generalisability. The positive coefficient for GFR, contrary to clinical intuition, may reflect residual confounding or coding direction and should be interpreted cautiously. Fifth, we did not perform subgroup analyses by CKD stage or diabetes duration due to insufficient sample size, precluding determination of optimal cut-offs for different populations. RBP and NAG are not routinely available in primary care settings, limiting model dissemination. The optimal cut-off of 0.46 yielded a sensitivity of only 54%, with a relatively high miss rate, for screening purposes, a lower threshold may be more appropriate. Furthermore, the model predicts high urinary cadmium status — an intermediate indicator- rather than hard endpoints such as ESRD or death, and its clinical value remains to be validated.

4.6 Clinical and public health implications

Based on our findings, we propose the following recommendations. For clinical screening, we suggest simultaneous measurement of corrected urinary RBP and NAG in high-risk CKD patients with comorbid diabetes, as the combined use of these two markers may offer superior sensitivity for early cadmium-related tubular injury compared with conventional GFR or cystatin C alone. For risk stratification, the nomogram developed in this study can be used to estimate the probability of high cadmium exposure, facilitating targeted environmental exposure intervention and long-term renal function follow-up in high-risk individuals. From a public health perspective, we recommend incorporating urinary tubular biomarkers into routine screening programmes in cadmium-contaminated regions, with particular emphasis on individuals with diabetes and CKD.

4.7 Future directions

Further research should focus on: conducting multicentre prospective cohort studies to validate the long-term predictive value of NAG and RBP for cadmium-related CKD progression, exploring the molecular mechanisms underlying the interactions among cadmium exposure, diabetes and tubular injury, and integrating dietary and environmental exposure data to construct a comprehensive multi-dimensional cadmium exposure risk prediction model.

5. Conclusions

In summary, this study demonstrates that elevated urinary cadmium levels are significantly associated with tubular injury biomarkers (RBP, NAG, β_2-MG and U_mALB) in non-dialysis CKD patients, among which RBP and NAG are the optimal predictors for identifying high-risk cadmium exposure, with NAG showing a non-linear relationship with cadmium exposure. Patients with comorbid diabetes may represent a susceptible subgroup to cadmium nephrotoxicity. The nomogram prediction model based on NAG, RBP, GFR and T2DM demonstrates good calibration and moderate discrimination, and may serve as an adjunctive risk stratification tool. We suggest incorporating urinary RBP and NAG into routine monitoring for high-risk non-dialysis CKD populations. Future multicentre prospective cohort studies are warranted to validate the predictive value of these biomarkers and to explore whether reducing cadmium exposure can slow CKD progression.

Declarations

Ethics declarations

The study protocol was approved by the Ethics Committee of Yangpu Hospital Affiliated to Tongji University (approval number: LL-2026-LW-014), and all procedures were performed in accordance with the Declaration of Helsinki. Due to the retrospective nature of the study, the requirement for informed consent was waived by the Ethics Committee.

Author information

Authors and Affiliations:

Xi Chen, Tian Tang, Yuxue Liu, Rong Zhou, Xuejing Guan. Department of Nephrology, Yangpu Hospital, Tongji University School of Medicine, Shanghai 200090, China.

Contributions:

All authors made substantial contributions to this work. CX designed the study, drafted and revised the manuscript, and handled submission and correspondence with the journal. TT performed statistical analyses and prepared the manuscript according to journal guidelines. LYX was responsible for case screening, and clinical data collection and curation. ZR supervised the research direction, critically reviewed the manuscript, and coordinated the review process to reach consensus among all authors. GXJ took charge of overall quality assurance, and critically reviewed the manuscript and responses to reviewers. All authors read and approved the final manuscript.

Corresponding author:

Rong Zhou,Chief Physician, MD

Department of Nephrology, Yangpu Hospital, Tongji University School of Medicine, Shanghai 200090, China.

Email address: zhourongyp@126.com

Xuejing Guan, MD

Department of Nephrology, Yangpu Hospital, Tongji University School of Medicine, Shanghai 200090, China.

Email address:Xuejing_guan@tongji.edu.cn

Consent for publication:

Not Applicable

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the General Project of Shanghai Municipal Health Commission, China, (Grant No. 202340058).

Author Contribution

All authors made substantial contributions to this work. CX designed the study, drafted and revised the manuscript, and handled submission and correspondence with the journal. TT performed statistical analyses and prepared the manuscript according to journal guidelines. LYX was responsible for case screening, and clinical data collection and curation. ZR supervised the research direction, critically reviewed the manuscript, and coordinated the review process to reach consensus among all authors. GXJ took charge of overall quality assurance, and critically reviewed the manuscript and responses to reviewers. All authors read and approved the final manuscript.

Data Availability

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

References

1. Nordberg GF, Nordberg M. Metallothionein and Other Factors Influencing Cadmium-Induced Kidney Dysfunction: Review and Commentary. Biomolecules. 15(8). Switzerland:MDPI (Basel, Switzerland),2025. 1083.
2. Satarug S, Vesey DA, Gobe GC. Kidney Cadmium Toxicity, Diabetes and High Blood Pressure: The Perfect Storm. Tohoku J Exp Med. 241(1). Japan:Tohoku University Medical Press,2017. 65–87.
3. Jiayi ZW. Advances of Epidemiological Study onHazard of EnvironmentalPopulation Exposure and HealthCadmium Pollution. Journal of Environmental Hygiene. 9(6). Beijing, China:Chinese Center for Disease Control and Prevention,2019. 621–627.
4. Zhao L, Liao M, Li L et al. Cadmium activates the innate immune system through the AIM2 inflammasome. Chem Biol Interact. 399Ireland:Elsevier Ireland Ltd,2024. 111122.
5. Satarug S. Challenges in Toxicological Risk Assessment of Environmental Cadmium Exposure. Toxics. 13(5). Switzerland:Multidisciplinary Digital Publishing Institute (MDPI); 2025.
6. Satarug S, C Gobe G, Vesey A. D, Cadmium and Lead Exposure, Nephrotoxicity, and Mortality. Toxics 8(4). Switzerland:Multidisciplinary Digital Publishing Institute (MDPI),2020.
7. Qing Y, Yang J, Zhu Y et al. Dose-response evaluation of urinary cadmium and kidney injury biomarkers in Chinese residents and dietary limit standards. Environ Health 20(1). England:BioMed Central,2021. 75.

8. Satarug S. Dietary Cadmium Intake and Its Effects on Kidneys. Toxics. 6(1). Switzerland:Multidisciplinary Digital Publishing Institute (MDPI),2018.
9. Pócsi I, Dockrell ME, Price RG. Nephrotoxic Biomarkers with Specific Indications for Metallic Pollutants: Implications for Environmental Health. Biomark Insights. 17United States:SAGE Publishing,2022. 11772719221111882.
10. Zhang YR, Wang P, Liang XX et al. Associations between Urinary Excretion of Cadmium and Renal Biomarkers in Nonsmoking Females: A Cross-Sectional Study in Rural Areas of South China. Int J Environ Res Public Health. 12(10). Switzerland:MDPI (Basel, Switzerland),2015. 11988 – 2001.
11. Kalahasthi R, Nagaraju R, Balachandar R et al. Association between Occupational Cadmium Exposure and Markers of Kidney Injury: A Systematic Review and Meta-Analysis. Chem Res Toxicol. 38(10). United States:American Chemical Society,2025. 1638–51.
12. Audureau E, Soubeyran P, Martinez-Tapia C et al. Machine Learning to Predict Mortality in Older Patients With Cancer: Development and External Validation of the Geriatric Cancer Scoring System Using Two Large French Cohorts. J Clin Oncol 43(12). United States:American Society of Clinical Oncology,2025. 1429–40.
13. Bai H, Wang X, Wang Q et al. Machine learning-based clinical decision support tool for advanced ESCC in the immunotherapy era: a multi-center study. Cancer Biol Med. China:Chinese Anticancer Association,2026.
14. Chang HH, Chiang JH, Tsai CC et al. Predicting hyperkalemia in patients with advanced chronic kidney disease using the XGBoost model. BMC Nephrol. 24(1). England:BioMed Central,2023. 169.
15. Gao T, Nong Z, Luo Y et al. Machine learning-based prediction of in-hospital mortality for critically ill patients with sepsis-associated acute kidney injury. Ren Fail 46(1). England:Informa Healthcare,2024. 2316267.
16. Jang EC, Park YM, Han HW et al. Machine-learning enhancement of urine dipstick tests for chronic kidney disease detection. J Am Med Inform Assoc. 30(6). England:Oxford University Press,2023. 1114–1124.
17. Cao Lixin HM, Tang Ying SJ, Yubin W. Epidemiological survey of urinary Cadmium in Shengyang urban children. Chinese Journal of Applied Clinical Pediatrics. 29(23). Xinxiang, Henan, China:Chinese Medical Association,2014. 1806–1809.
18. YANG Yadi GUY, Yan-sen XU. YY. Spatial-temporal Distribution Patterns of Urinary Cadmium Levels in the General Population in China. Journal of Ecology and Rural Environment. 40(10). Nanjing, Jiangsu, China:Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment,2024. 1269–1277.
19. Ruangyuttikarn W, Panyamoon A, Nambunmee K et al. Use of the kidney injury molecule-1 as a biomarker for early detection of renal tubular dysfunction in a population chronically exposed to cadmium in the environment. Springerplus 2Switzerland,2013. 533.
20. Wen ZHU, Lin-shen ZL. X. Advances in the toxic mechanisms of cadmium-induced kidney injury. Modern Preventive Medicine. 53(3). Chengdu, Sichuan, China:Chinese Preventive Medicine

Association, West China School of Public Health, Sichuan University,2026. 442–446,459.
21. Zhu H, Tang X, Gu C et al. Assessment of human exposure to cadmium and its nephrotoxicity in the Chinese population. Sci Total Environ. 918Netherlands:Elsevier,2024. 170488.
22. Wang J, Zhong J, Yang HC et al. Cross Talk from Tubules to Glomeruli. Toxicol Pathol. 46(8). United States:SAGE Publications Inc,2018. 944–948.
23. Christensen EI, Birn H. Megalin and cubilin: synergistic endocytic receptors in renal proximal tubule. Am J Physiol Ren Physiol 280(4). United States:American Physiological Society,2001. F562-73.
24. Thévenod F, Lee WK. Cadmium and cellular signaling cascades: interactions between cell death and survival pathways. Arch Toxicol. 87(10). Germany:Springer Berlin Heidelberg,2013. 1743-86.
25. Siddiqui K, Al-Malki B, George TP, et al. Urinary N-acetyl-beta-d-glucosaminidase (NAG) with neutrophil gelatinase-associated lipocalin (NGAL) improves the diagnostic value for proximal tubule damage in diabetic kidney disease. 3 Biotech. 2019;9(3):66.
26. Satarug S, Vesey DA, Nishijo M et al. The Effect of Cadmium on GFR Is Clarified by Normalization of Excretion Rates to Creatinine Clearance. Int J Mol Sci 22(4). Switzerland:MDPI (Basel, Switzerland),2021.
27. Ikeda M, Ezaki T, Tsukahara T et al. Threshold levels of urinary cadmium in relation to increases in urinary beta2-microglobulin among general Japanese populations. Toxicol Lett. 137(3). Netherlands:Elsevier BV,2003. 135 – 41.
28. Ma Y, Yue C, Sun Q, et al. Cadmium exposure exacerbates kidney damage by inhibiting autophagy in diabetic rats. Ecotoxicol Environ Saf. 2023;267:115674.
29. Zhang Y, Gong X, Li R et al. Exposure to cadmium and lead is associated with diabetic kidney disease in diabetic patients. Environ Health. 23(1). England:BioMed Central,2024. 1.

Figures

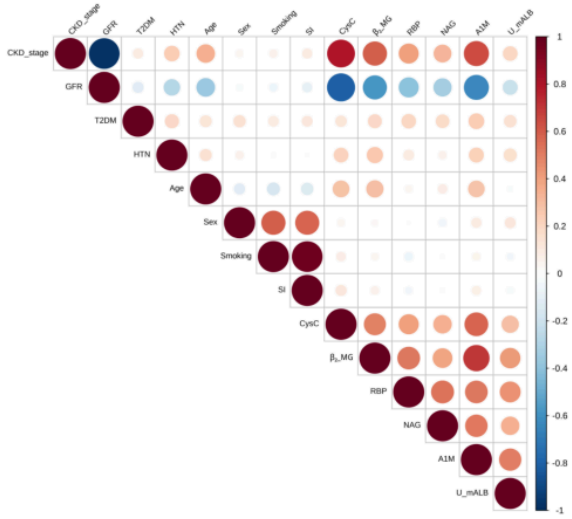


Figure 1-a Correlation coefficient heatmap for the low-cadmium group

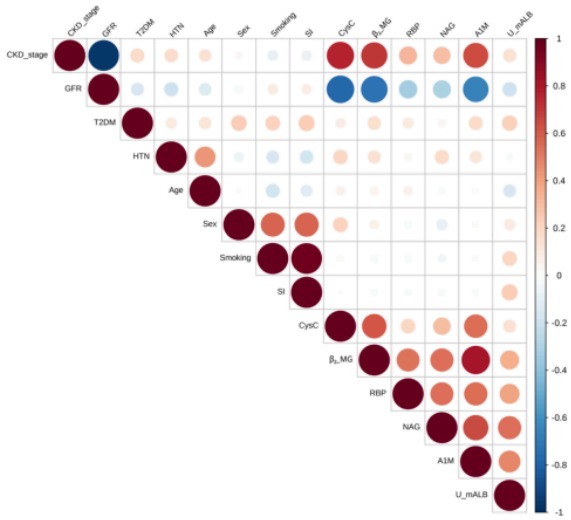


Figure 1-b Correlation coefficient heatmap for the low-cadmium group

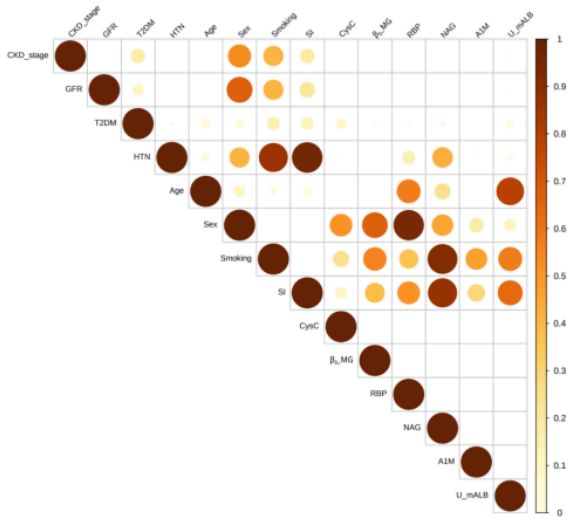


Figure 1-c P-value heatmap for the low-cadmium group

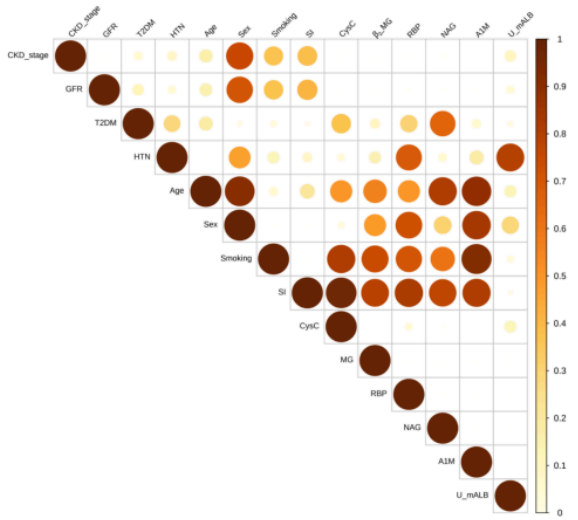


Figure 1-d P-value heatmap for the high-cadmium group

Figure 1

Heatmaps of Spearman's correlation coefficients (a, b) and P-values (c, d) in the low- and high-cadmium groups. Red/blue denote positive/negative correlations. *P < 0.05, **P < 0.01, ***P < 0.001. β_2-MG, RBP, NAG, U_mALB, α_1-MG, GFR, CysC.

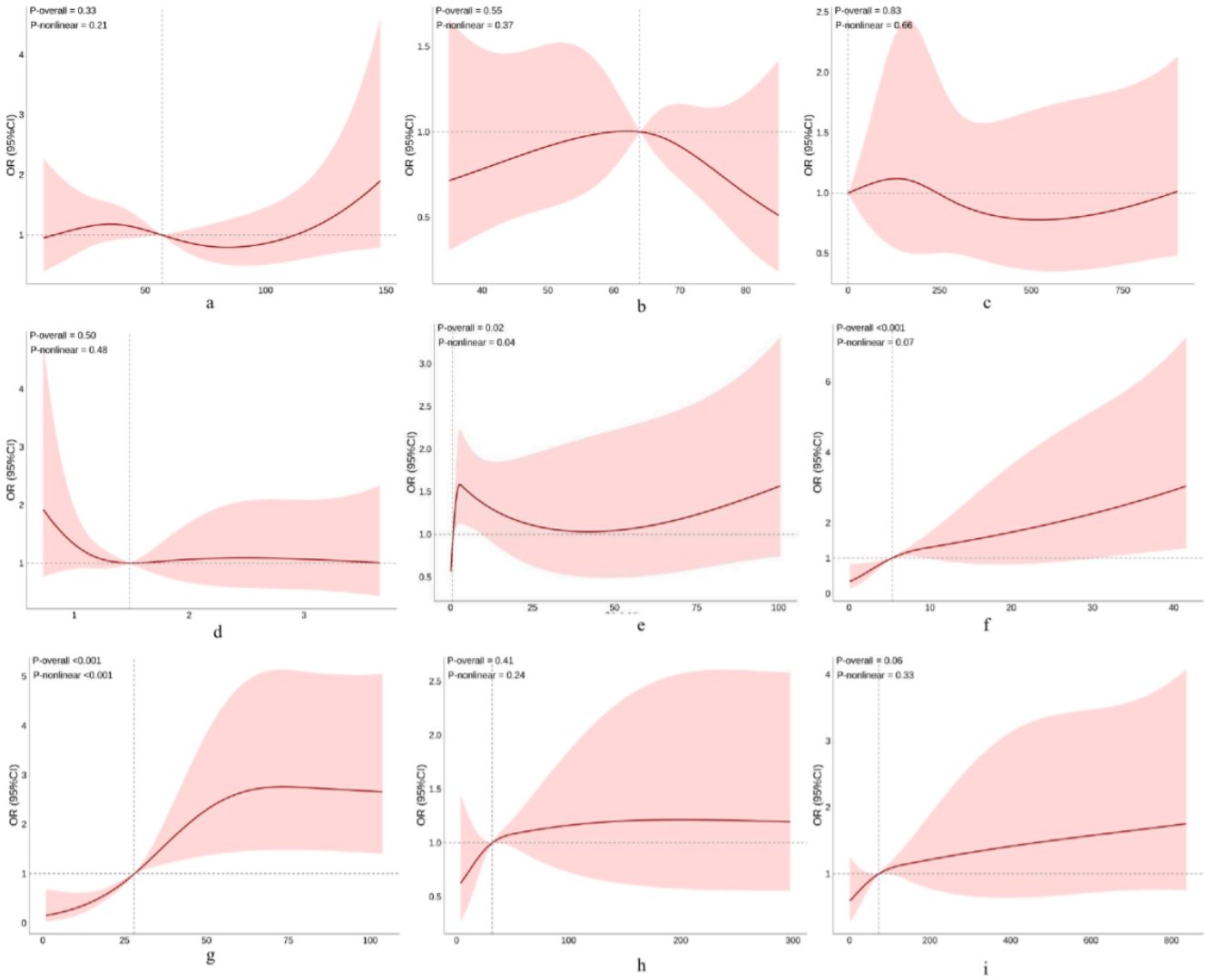


Figure2(a-i) Restricted cubic spline (a GFR, b Age, c SI, d CysC, e β2-MG, f RBP, g NAG, h A1M, i U_mALB)

Figure 2

Dose-response relationships between nine continuous variables and the outcome variable (Group) based on restricted cubic spline regression

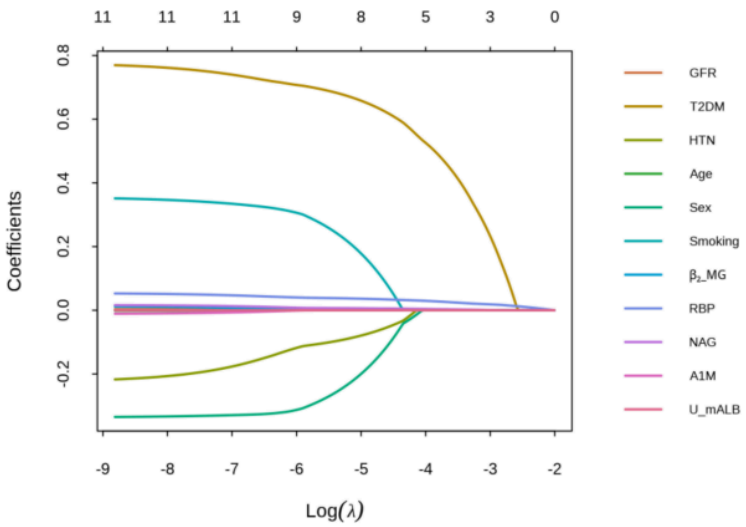


Figure 3-a: LASSO regression coefficient path plot

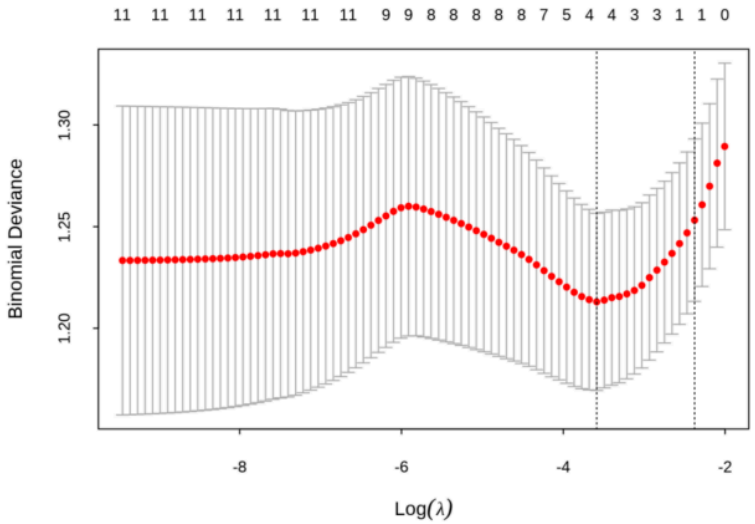


Figure 3-b: 10-fold cross-validation error curve for LASSO regression

Figure 3

LASSO regression coefficient path plot (a), 10-fold cross-validation error curve for LASSO regression (b)

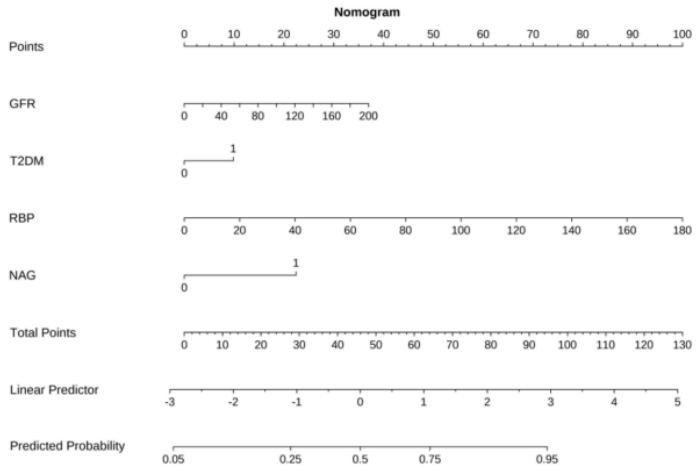


Figure 4 -a Nomogram for predicting the risk of the outcome event

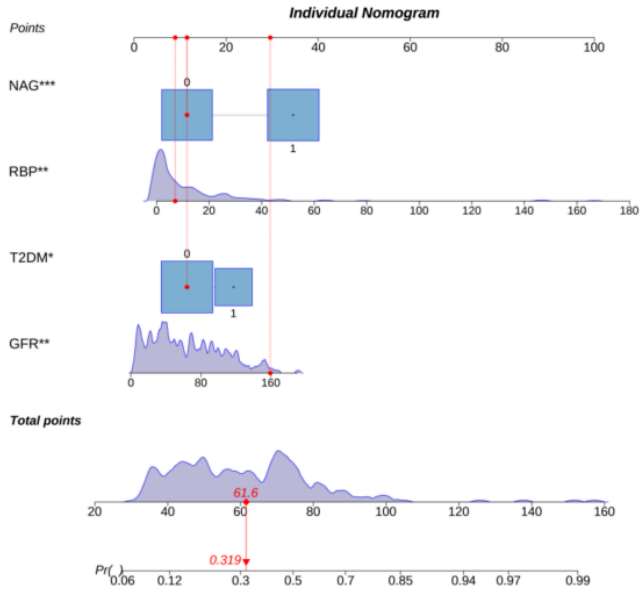


Figure 4 -b Example of nomogram application

Figure 4

Visualization and application of the nomogram: Nomogram for predicting the risk of the outcome event (a), Example of nomogram application (b)

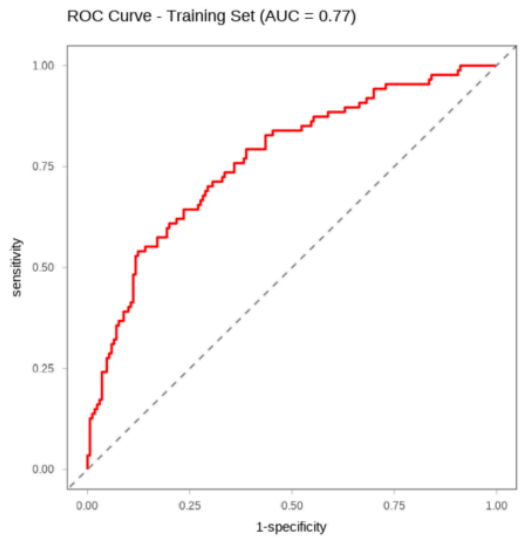


Figure 5-a Receiver operating characteristic (ROC) curve of the prediction model

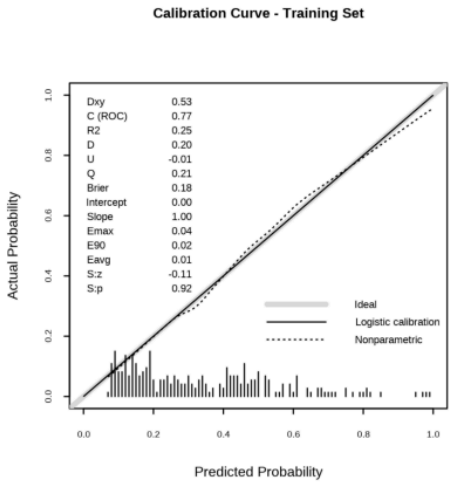


Figure 5-b Calibration curve of the prediction model

Figure 5

Assessment of discrimination and calibration of the nomogram. Receiver operating characteristic (ROC) curve of the prediction model (a), Calibration curve of the prediction model (b)

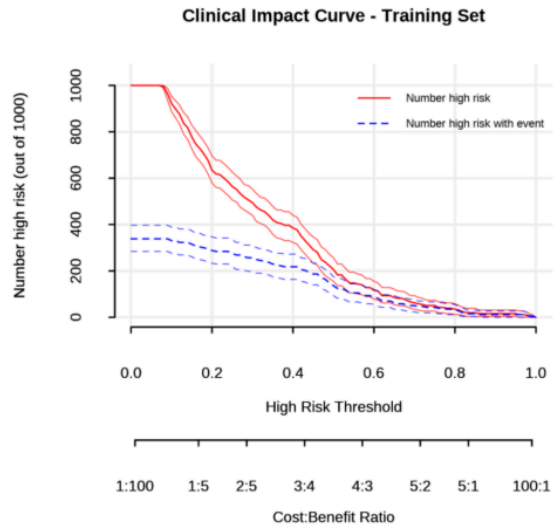


Figure 6-a Decision curve analysis (DCA) for the prediction model

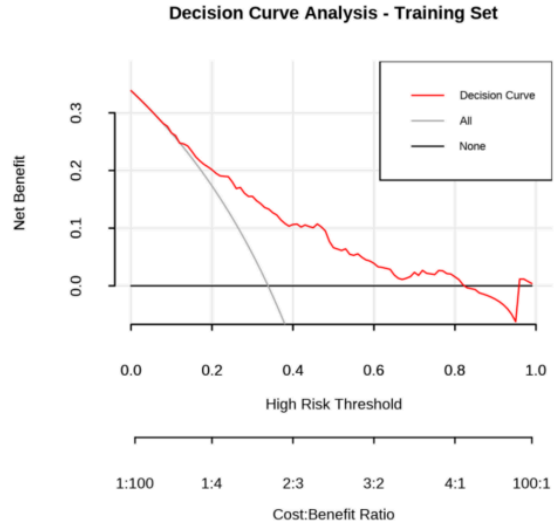


Figure 6-b Clinical impact curve (CIC) for the prediction model

Figure 6

Assessment of clinical utility of the nomogram.

Decision curve analysis (DCA) for the prediction model (a), Clinical impact curve (CIC) for the prediction model (b)