Letter to the Editor

Prediction of Renal Function by Urinary Lead and Cadmium — Based on Classification Decision Tree and Logistic Regression Model*

LI Yang[1,2,&], WANG Qing Yao[1,2,&], TIAN Qing Hua[1,2], AN Qi[1,2], YANG Yu Tong[1,2], ZHANG Jia Chen[1], LI Shuang Jing[1], ZHOU Han[1], LIANG Yun Fen[1], SHEN Wei Tong[1], MU Li Na[3,#], and LEI Li Jian[1,2,#]

It is estimated that chronic kidney disease (CKD) will be the fifth leading cause of death in the world by 2040[1]. Early recognition and intervention for kidney damage are essential. Estimated glomerular filtration rate (eGFR) can be calculated by measuring blood creatinine to evaluate glomerular function, and urinary N-acetyl-β-δ-glucosaminidase (UNAG) level is generally recognized as a marker of renal tubular injury. Exposure to lead (Pb) and cadmium (Cd) can damage renal function, leading to a decrease in eGFR and an increase in UNAG[2]. Both Pb and Cd are easily accumulated in the body and then slowly released by the kidneys and excreted. So, urinary Pb (UPb) and Cd (UCd) levels can be used as biomarkers of exposure[3,4]. Therefore, predicting renal function based on UPb and UCd levels may provide scientific evidence for establishing toxicity thresholds and provide reasonable strategies for preventing kidney damage. Logistic regression and classification decision tree are commonly used as predictive models to identify the factors that affect diseases[5]. The logistic regression combined with classification decision tree models to predict renal impairment is currently lacking. In this study, eGFR and UNAG were used as effect indicators, and UPb and UCd levels were included as factors in the prediction model, to construct logistic regression and classification decision tree model, and analyzed the predictive performance of the model.

This study is a community-based cross-sectional survey in which we randomly selected residents from two communities in northern China as study participants through whole population sampling. The inclusion criteria for this study were as follows: age over 18 years, residing locally area for at least 5 years, and voluntarily participating in the survey. The exclusion criteria included individuals with secondary kidney diseases such as diabetic nephropathy, those taking medications that could affect renal function, occupationally exposure to heavy metals, and those with insufficient information or incomplete samples. Baseline information on the study participants was collected including gender, age, education level, per capita monthly household income, smoking, and alcohol consumption. The physical examination included height and weight, and the body mass index (BMI) was calculated based on height and weight. Urine specimens were collected from the study participants to measure urinary creatinine (UCr), UCd, UPb, and UNAG. And UCr levels were used to correct for urinary heavy metal levels and UNAG. Blood samples were collected from the study participants to measure high-density lipoprotein cholesterol (HDLC), total cholesterol (TC) levels, and serum creatinine (SCr), and SCr was used to calculate the eGFR. Dual data entry using EpiData version 3.0 to create the database. Divided the dataset into training and testing set in a ratio of 7:3. Before building the model, univariate analysis was performed by SPSS 26.0 software with eGFR and UNAG as dependent variables, respectively. Study participants were divided into a low eGFR group (eGFR ≤ 60.00 mL/min per 1.73 m^2) and a control group (eGFR > 60.00 mL/min per 1.73 m^2) based on clinically recognized eGFR diagnostic thresholds[6]. Divided into a high UNAG group (UNAG ≥ 16.28 U/g

doi: 10.3967/bes2024.037

*This study was supported by the National Natural Science Foundation of China [grant numbers 81872701, 81273040] and Shanxi Scholarship Council of China [grant number 2022–113].

1. Department of Epidemiology, School of Public Health, Shanxi Medical University, Taiyuan 030001, Shanxi, China; 2. Department of Epidemiology, School of Public Health, Shanxi Key Laboratory of Environmental Health Impairment and Prevention, NHC Key Laboratory of Pneumoconiosis, MOE Key Laboratory of Coal Environmental Pathogenicity and Prevention, Shanxi Medical University, Taiyuan 030001, Shanxi, China; 3. Department of Epidemiology and Environmental Health, School of Public Health and Health Professions, The State University of New York at Buffalo, Buffalo, NY, USA.

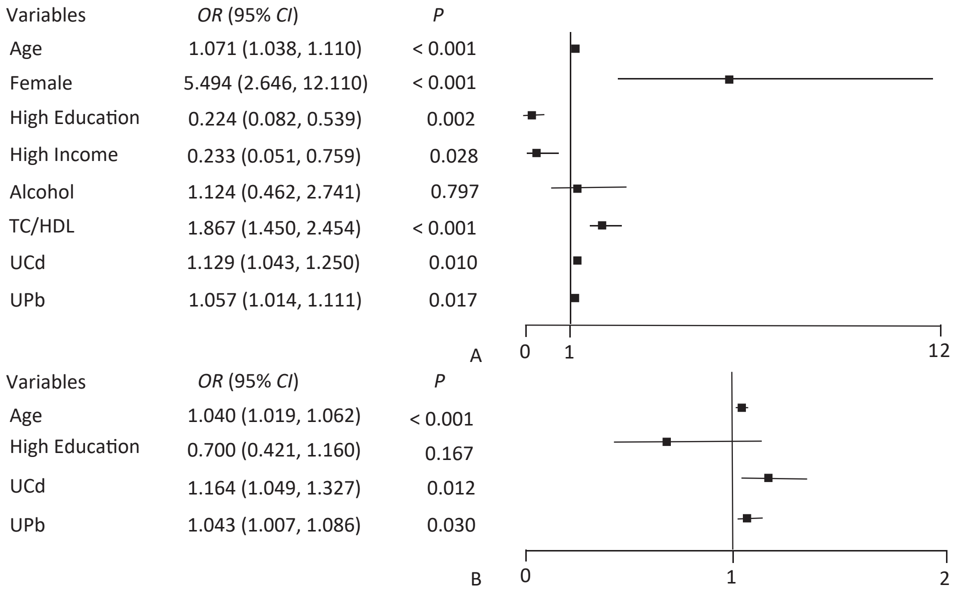
Cr) and a UNAG control group (UNAG < 16.28 U/g Cr) according to the median level of UNAG (16.28 U/g Cr). Statistically significant independent variables were included in logistic regression and classification decision tree. The variance inflation factor (VIF) was utilized to test for the multicollinearity between variables. R4.0.4 software was used to construct the model on the training set, and the predictive performance of the model was assessed by creating participant's working characteristic curve (receiver operating characteristic) and confusion matrix, calculating the area under the curve, sensitivity, specificity, accuracy, and precision. Two-sided P values < 0.05 were statistically significant.

Based on the inclusion and exclusion criteria, 532 participants were included in our study, of which 190 were males and 342 were females. The differences in per capita monthly household income, smoking and alcohol consumption were statistically significant between males and females. Further details on participant characteristics are shown in Supplementary Table S1 (available in www.besjournal.com). The results of univariate analysis showed that there were statistically significant differences (P < 0.001) between the low eGFR and the eGFR control groups in terms of gender, age, education level, per capita monthly income, alcohol consumption, TC/HDL ratio, UCd, and UPb levels (Supplementary Table S2, available in www.besjournal.com). And there were statistically significant differences between the high UNAG and the UNAG control groups in terms of age (P < 0.001), education (P < 0.001), UCd (P < 0.001), and UPb (P = 0.018) levels (Supplementary Table S3, available in www.besjournal.com). The variable assignments for the logistic regression are shown in Table 1, and the VIF for all variables were less than 5, indicating a low likelihood of multicollinearity among the variables in the model (Supplementary Tables S4–S5, available in www.besjournal.com).

The node variables included in the classification decision tree have both differences and similarities with the significant variables in logistic regression. The logistic regression results showed that UPb and UCd levels were risk factors for low eGFR (Figure 1A) and high UNAG (Figure 1B). The classification decision tree results showed the same results, with UPb level being the best classified feature for predicting the eGFR (Figure 2A) and UCd level being the best classified feature for predicting high UNAG (Figure 2B). Older age is also a risk factor for low eGFR and high UNAG (Figures 1–2), this is probably because renal function gradual decline with age, and changes in the structure and function of the kidneys, such as renal arteriosclerosis, glomerulosclerosis, and renal tubular atrophy with fibrosis[7]. When predicting eGFR, both logistic regression and classification decision tree incorporated the variables such as female and TC/HDL, suggesting that both female and TC/HDL levels were risk factors for lower eGFR (Figure 1A and Figure 2A), probably because the average age of the study participants in this study being around 60 years old, and it has been shown that a significant decrease in estrogen levels in older women lead to a decrease in the protective effect of the vascular endothelial system, resulting in a significant decrease in eGFR[8]. It has also been shown that lower TC and higher HDL levels are associated with a lower incidence of CKD, lower TC levels and higher HDL levels may have a protective effect on the kidneys[9]. However, logistic regression showed higher OR for female and TC/HDL (Figure 1A), suggesting that females and those with higher levels of TC/HDL are at greater risk of developing lower eGFR. In the classification decision tree, the two characteristic variables of female and TC/HDL level, were classified as leaf nodes rather than root node, with the root node was UPb level (Figure 2A). In addition, the logistic regression results also showed that higher income level and higher education level were protective factors for eGFR (Figure 1A), which may be due to the association between higher income and higher education level and better health management, which were not taken into account in the final classification decision tree model (Figure 2A). The similarities and differences between logistic regression and classification decision trees may be due to the different focuses of the two models. Logistic regression is a linear model that predicts outcomes by fitting a linear relationship to the data, focuses on explaining how the probability of the dependent variable changes when the independent variable changes. Its main purpose is to explore the risk factors of diseases and predict the probability of of disease occurrence based on risk factors. In contrast, the classification decision tree is a nonlinear model that divides the data based on the values of the predictors,

Table 1. Logistic regression analysis variables assignment table

Variable	Assignment
Sex	1 = male, 2 = female
Educational	0 = Junior high school and below, 1 = High school and above
Income (yuan)	0 = ≤ 3,000, 1 = > 3,000
Alcohol	0 = no, 1 = yes



Variables	*OR* (95% *CI*)	*P*
Age	1.071 (1.038, 1.110)	< 0.001
Female	5.494 (2.646, 12.110)	< 0.001
High Education	0.224 (0.082, 0.539)	0.002
High Income	0.233 (0.051, 0.759)	0.028
Alcohol	1.124 (0.462, 2.741)	0.797
TC/HDL	1.867 (1.450, 2.454)	< 0.001
UCd	1.129 (1.043, 1.250)	0.010
UPb	1.057 (1.014, 1.111)	0.017

Variables	*OR* (95% *CI*)	*P*
Age	1.040 (1.019, 1.062)	< 0.001
High Education	0.700 (0.421, 1.160)	0.167
UCd	1.164 (1.049, 1.327)	0.012
UPb	1.043 (1.007, 1.086)	0.030

Figure 1. Forest plot of eGFR and UNAG impact factors. (A) Forest plot of eGFR influencing factors; (B) Forest plot of UNAG influencing factors; eGFR, estimated glomerular filtration rate; UNAG, urinary N-acetyl-β-δ-glucosaminidase; HDL, high-density lipoprotein cholesterol; TC, total cholesterol; UCd, urinary cadmium; UPb, urinary lead.

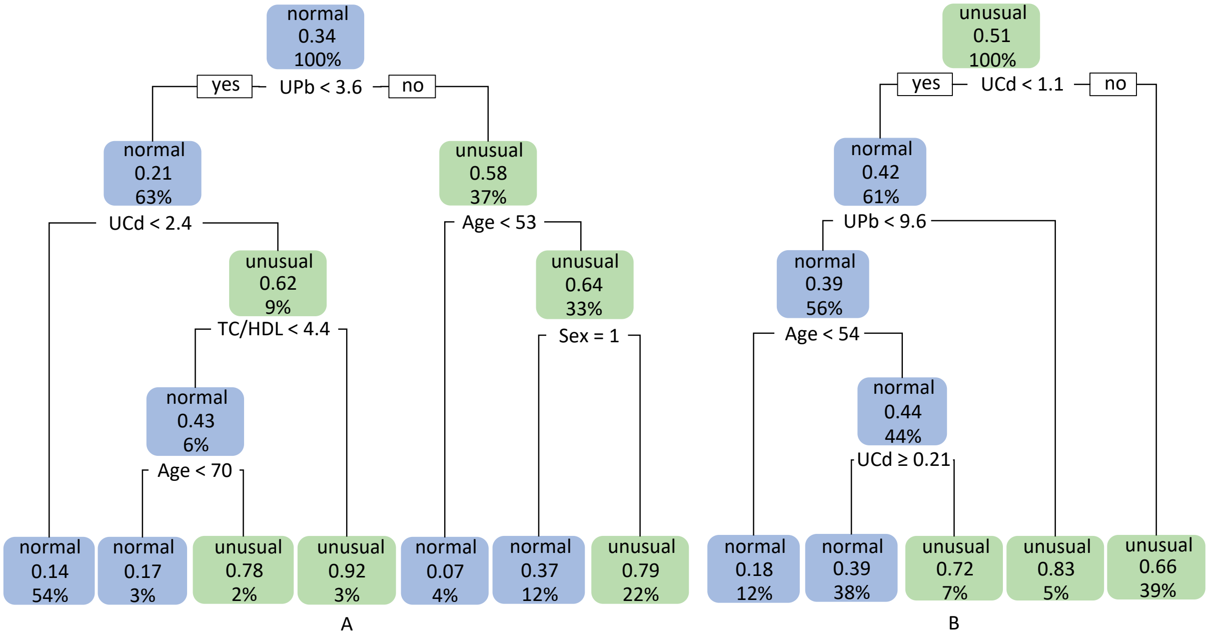


Figure 2. Classification decision trees for eGFR and UNAG. (A) Classification decision tree with eGFR as dependent variable; (B) Classification decision tree with UNAG as dependent variable; Blue leaf nodes indicated the probability that ≤ 50% of study participants have lower eGFR and higher UNAG, defined as normal eGFR and normal UNAG. Green leaf nodes indicate the probability that > 50% of study participants have lower eGFR and higher UNAG, defined as abnormal eGFR and abnormal UNAG. Decimal points indicate the probability that the classified population has lower eGFR and higher UNAG. Percentages indicate the percentage of the training set sample size that this categorized population represents. eGFR, estimated glomerular filtration rate; UNAG, urinary N-acetyl-β-δ-glucosaminidase; HDL, high-density lipoprotein cholesterol; TC, total cholesterol; UCd, urinary cadmium; UPb, urinary lead.

focuses on classifying all samples into the closest category by selecting the appropriate dividing characteristics at each node. Classification decision trees are able to find the optimal cutoff point for successive predictors in logistic regression. Both logistic regression and classification decision tree are more commonly used to categorize and predict models for dichotomous variables. In practice, the performance of these two models on a given dataset can be compared to determine which model is more suitable for a particular task, and the two models are used for different purposes and can be combined in practice to improve analytical performance.

The final model of the eGFR classification decision tree showed a 79% probability of lower eGFR in women aged over 53 years and had the UPb level ≥ 3.60 μg/g Cr. And a 92% probability of lower eGFR if the UPb level was < 3.60 μg/g Cr, the UCd level was ≥ 2.40 μg/g Cr, and the TC/HDL ratio was ≥ 4.4. The probability of low eGFR was 78% if the age was ≥ 70, the UPb level was < 3.60 μg/g Cr, the UCd level was < 2.40 μg/g Cr, and the TC/HDL ratio was < 4.4. (Figure 2A). The final UNAG classification decision tree results show that if the UCd level was ≥ 1.10 μg/g Cr, the probability of higher UNAG level was 66%; if the UCd level was < 1.10 μg/g Cr and the UPb level was ≥ 9.60 μg/g Cr, the probability of higher UNAG level was 83%; and if the UCd level was < 0.21 μg/g Cr , the UPb level was < 9.60 μg/g Cr, and the age was not less than 54 years, the probability of having a higher UNAG level was 72% (Figure 2B). And when the predictive variables were ranked by importance, UPb and UCd levels were the top two variables (Supplementary Figures S1–S2, available in www.besjournal.com), suggesting that reducing Pb and Cd exposure is more important than lowering the TC/HDL ratios or increasing education and income levels to protect kidney function. In addition, in this study population, the probability of eGFR abnormality was 14% when the UPb level was < 3.60 μg/g Cr and the UCd level < 2.40 μg/g Cr (Figure 2A). When the UCd level was < 1.10 μg/g Cr, the UPb level was < 9.60 μg/g Cr, and the age was less than 54 years, the probability of UNAG abnormality was 18% (Figure 2B). Suggest that kidney damage is possible when UCd level is > 1.10 μg/g Cr or UPb level is > 3.60 μg/g Cr, this indicates that UCd levels below the toxicity threshold levels recognized by the World Health Organization (5.24 μg/g Cr) may also trigger body damage[10]. For renal impairment, the toxicity threshold level of UPb could be < 3.60 μg/g Cr.

In this study, area under the curve, sensitivity, specificity, accuracy, and precision were calculated using ROC curves and confusion matrices, respectively, to assess the predictive performance of the models. eGFR was used as the effect indicator, the area under the ROC curve for the logistic regression and classification decision tree models were 0.779 (95% *CI*: 0.709–0.844) and 0.754 (95% *CI*: 0.697–0.844) (Supplementary Figure S3, available in www.besjournal.com). The sensitivity, specificity, precision, and accuracy of the logistic regression model were 56.1%, 84.5%, 66.7%, and 74.4% (Supplementary Figure S4A, available in www.besjournal.com), respectively, as well as the sensitivity, specificity, precision, and accuracy of the classification decision tree were 62.9%, 86.7%, 75.0% and 77.5% (Supplementary Figure S4B). And using UNAG as the effect indicator, the logistic regression model and the classification decision tree model showed that the area under the ROC curve were 0.624 (95% *CI*: 0.543–0.702) and 0.597 (95% *CI*: 0.53–0.675) (Supplementary Figure S5, available in www.besjournal.com). The logistic regression model had a sensitivity, specificity, precision, and accuracy of 50.7%, 56.6%, 56.1% and 58.1%, respectively (Supplementary Figure S6A, available in www.besjournal.com). The classification decision tree had a sensitivity, specificity, precision, and accuracy of 53.3%, 57.7%, 55.6% and 52.6%, respectively (Supplementary Figure S6B). The area under the ROC curve of both models was greater than 0.5, and the confusion matrix also showed that the sensitivity, specificity, precision and accuracy of both models were higher than 50%. It indicated that both models have certain predictive performance, and UPb levels and UCd levels can be used as indicators to predict lower eGFR and higher UNAG levels.

In this study, using cross-sectional data to construct the model has certain advantages. Firstly, the sample data can be obtained and the relationship between the variables can be evaluated in a shorter time. Secondly, the model constructed from cross-sectional data requires fewer parameters, which reduces the complexity of the model and the probability of error. This study also has some limitations. Firstly, participants were recruited from two selected random communities in Taiyuan, Shanxi Province, China. Therefore, to confirm the generalizability of these results, similar studies should be conducted in other regions. Secondly, because our study was a cross-sectional study, causality could not be established. The effect of UPb and UCd levels in predicting impaired renal function needs to be verified through cohort studies.

This study only uses internal validation and only tests the model's ability to predict sample data, and these results require further external validation.

In conclusion, we found that high levels of UPb and UCd, as well as older age, were associated with lower eGFR and higher UNAG. In this study, the UPb level was the best predictor of lower eGFR, while the UCd level was the best predictor of higher UNAG level. The ROC curves and confusion matrixes showed that the constructed model had good predictive performance. These results provide a potential scientific research basis for predicting renal function based on UPb and UCd levels, which may provide new research directions for the early diagnosis of kidney diseases.

We would like to thank all the study participants who volunteered to take part and all the community workers for their support of our research work.

The authors declare that they have no conflict of interest.

[&]These authors contributed equally to this work.

[#]Correspondence should be addressed to MU Li Na, E-mail: linamu@buffalo.edu; LEI Li Jian, E-mail: wwwdlijian@sxmu.edu.cn

Biographical notes of the first authors: LI Yang, female, born in 1998, Master of medicine, majoring in public health and preventive medicine; WANG Qing Yao, female, born in 1998, Master of medicine, majoring in public health and preventive medicine.

Received: September 15, 2023;

Accepted: January 22, 2024

REFERENCES

1. Rayego-Mateos S, Marquez-Expósito L, Rodrigues-Diez R, et al. Molecular mechanisms of kidney injury and repair. Int J Mol Sci, 2022; 23, 1542.
2. Chen X, Zhu GY, Wang ZQ, et al. The association between lead and cadmium co-exposure and renal dysfunction. Ecotoxicol Environ Saf, 2019; 173, 429−35.
3. Bai Y, Laenen A, Haufroid V, et al. Urinary lead in relation to combustion-derived air pollution in urban environments. A longitudinal study of an international panel. Environ Int, 2019; 125, 75−81.
4. Madrim MF, Ja'afar MH, Hod R. Prevalence of abnormal urinary cadmium and risk of albuminuria as a primary bioindicator for kidney problems among a healthy population. PeerJ, 2021; 9, e12014.
5. Lei Y, Nollen N, Ahluwahlia JS, et al. An application in identifying high-risk populations in alternative tobacco product use utilizing logistic regression and CART: a heuristic comparison. BMC Public Health, 2015; 15, 341.
6. Stevens PE, Levin A. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. Ann Intern Med, 2013; 158, 825−30.
7. Denic A, Glassock RJ, Rule AD. Structural and functional changes with the aging kidney. Adv Chronic Kidney Dis, 2016; 23, 19−28.
8. Wang HB, Lou YJ, Ma YJ, et al. Estimating the glomerular filtration rate and tubular dysfunction in an elderly population with normoalbuminuria in China. Clin Chim Acta, 2019; 495, 377−81.
9. Miao L, Min Y, Qi B, et al. Causal effect between total cholesterol and HDL cholesterol as risk factors for chronic kidney disease: a Mendelian randomization study. BMC Nephrol, 2021; 22, 35.
10. Joint FAO/WHO Expert Committee on Food Additives (JECFA). Summary report of the seventy-third meeting of JECFA. World Health Organization. 2010.

Supplementary Table S1. Basic characteristics of the study participants

Variables	**Male (*n* = 190)**	**Female (*n* = 342)**	**Z/χ^2**	***P***
Age (years), *n* (%)				
< 50	22 (11.6)	58 (17.0)	5.221	0.074
50–69	125 (65.8)	229 (67.0)		
≥ 70	43 (22.6)	55 (16.0)		
Educational level, *n* (%)				
Primary school and below	72 (37.9)	145 (42.4)	2.073	0.557
Junior high school	58 (30.5)	102 (29.8)		
Senior high school	50 (26.3)	84 (24.6)		
Bachelor's degree and above	10 (5.3)	11 (3.2)		
Income (yuan), *n* (%)[a]				
< 1,000	80 (42.1)	147 (43.0)	8.223	0.042
1,000–2,999	74 (38.9)	137 (40.1)		
3,000–4,999	23 (12.2)	51 (14.9)		
≥ 5,000	13 (6.8)	7 (2.0)		
BMI, *n* (%)				
< 24	64 (33.9)	118 (34.5)	0.138	0.933
24–27.9	76 (40.2)	133 (38.9)		
≥ 28	49 (25.9)	91 (26.6)		
Smoker, *n* (%)				
No	65 (34.2)	318 (93.0)	209.241	< 0.001
Yes	125 (65.8)	24 (7.0)		
Alcohol consumption, *n* (%)				
No	112 (58.9)	326 (95.3)	111.091	< 0.001
Yes	78 (41.1)	16 (4.7)		
TC/HDLC, *M* (IQR)	3.86 (3.27–4.64)	3.93 (3.14–4.77)	−0.321	0.749
UCd (μg/gCr), *M* (IQR)	0.89 (0.47–1.71)	0.84 (0.45–1.61)	−0.793	0.428
UPb (μg/gCr), *M* (IQR)	2.63 (0.91–5.53)	2.87 (0.97–5.98)	−0.982	0.326
eGFR mL/min/1.73 m^2, *M* (IQR)	70.74 (60.62–89.71)	75.26 (48.73–93.67)	−1.634	0.102
UNAG U/gCr, *M* (IQR)	16.28 (11.15–25.08)	16.28 (11.03–23.71)	−5.77	0.564

Note. [a]Monthly household income per capita. BMI, body mass index; HDL, high-density lipoprotein cholesterol; TC, total cholesterol; IQR, interquartile range; UCd, urinary cadmium; UPb, urinary lead; *M*, median.

Supplementary Table S2. Univariable analysis of factors associated with low eGFR

Variables	**eGFR control group (*N* = 342)**	**Low eGFR[a] group (*N* = 190)**	**Z/χ^2**	***P***
Age (years), *M* (IQR)	60 (51–65)	64 (61–70)	−6.919	< 0.001
Sex, *n* (%)			20.296	< 0.001
Male	146 (42.7)	44 (23.2)		
Female	196 (57.3)	146 (76.8)		
Educational level, *n* (%)			81.578	< 0.001
Junior high school and below	197 (57.6)	180 (94.7)		
High school and above	145 (42.4)	10 (5.3)		
Income (yuan), *n* (%)[b]			39.736	< 0.001
< 3,000	255 (74.6)	183 (93.6)		
≥ 3,000	87 (25.4)	7 (3.4)		
Smoker, *n* (%)			3.447	0.063
No	237 (69.3)	146 (76.8)		
Yes	105 (30.7)	44 (23.2)		
Alcohol consumption, *n* (%)			4.135	0.042
No	273 (79.8)	165 (86.8)		
Yes	69 (20.2)	25 (13.2)		
BMI (kg/m^2), $\bar{x} \pm s_x$	25.49 ± 3.50	25.62 ± 3.69	−1.228	0.219
TC/HDLC, *M* (IQR)	3.65 (3.03–4.30)	4.38 (3.70–5.12)	−7.542	< 0.001
UCd (μg/gCr), *M* (IQR)	0.76 (0.42–1.26)	1.11 (0.59–2.78)	−4.821	< 0.001
UPb (μg/gCr), *M* (IQR)	1.72 (0.68–4.00)	4.67 (2.43–7.88)	−7.506	< 0.001

Note. [a]eGFR < 60 mL/min/1.73 m^2; [b]Monthly household income per capita. BMI, body mass index; *M*, median; IQR, interquartile range; s_x, standard deviation; TC, total cholesterol; HDLC, high-density lipoprotein cholesterol; UCd, urinary cadmium level; UPb, urinary lead level.

Supplementary Table S3. Univariable analysis of factors associated with high UNAG

Variables	UNAG control group (*N* = 266)	High UNAG[a] group (*N* = 266)	*Z*/χ^2	*P*
Age (years), *M* (IQR)	60 (51–65)	64 (60–70)	−6.072	< 0.001
Sex, *n* (%)				
Male	95 (35.7)	95 (35.7)	0.000	> 0.999
Female	171 (64.3)	171 (64.3)		
Educational level, *n* (%)				
Junior high school and below	169 (63.5)	208 (78.2)	13.847	< 0.001
High school and above	97 (36.5)	58 (21.8)		
Income (yuan), *n* (%)[b]				
< 3,000	214 (80.5)	224 (84.2)	1.292	0.256
≥ 3,000	52 (19.5)	42 (15.8)		
Smoker, *n* (%)				
No	194 (72.9)	189 (71.1)	0.233	0.629
Yes	72 (27.1)	77 (28.9)		
Alcohol consumption, *n* (%)				
No	222 (83.5)	216 (81.2)	0.465	0.495
Yes	44 (16.5)	50 (18.8)		
BMI (kg/m^2), $\bar{x} \pm s_x$	25.47 ± 3.33	25.83 ± 3.82	−1.124	0.261
TC/HDLC, *M* (IQR)	3.83 (3.12–4.67)	3.98 (3.27–4.79)	−1.940	0.052
UCd (μg/gCr), *M* (IQR)	0.74 (0.42–1.28)	0.99 (0.54–2.12)	−3.764	< 0.001
UPb (μg/gCr), *M* (IQR)	2.40 (0.73–5.41)	3.09 (1.08–6.50)	−2.373	0.018

Note. [a]UNAG ≥ 16.28 U/gCr; [b]Monthly household income per capita. BMI, body mass index; *M*, median; IQR, interquartile range; s_x, standard deviation; TC, total cholesterol; HDLC, high-density lipoprotein cholesterol; UCd, urinary cadmium level; UPb, urinary lead level.

Supplementary Table S4. Results of multicollinearity testing for estimated glomerular filtration rate (eGFR)

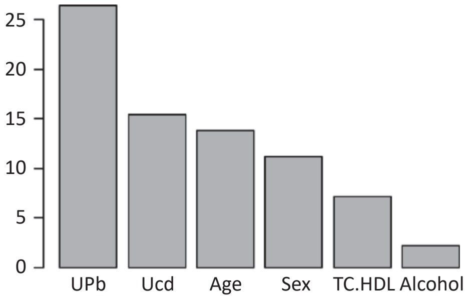
Variables	Non-standardized coefficient		Standard coefficient	*t*-Value	Sig.	Collinearity statistics	
	Beta	Standard error	Beta			Tolerance	VIF
(constant)	−0.470	0.159		−2.948	0.003		
Age	0.008	0.002	0.196	4.974	0.000	0.826	1.210
Sex	−0.203	0.040	−0.203	5.009	0.000	0.783	1.278
Educational level	−0.196	0.046	−0.186	−4.280	0.000	0.677	1.478
Income	0.132	0.051	−0.105	−2.590	0.010	0.782	1.278
Alcohol consumption	0.000	0.051	0.000	−0.004	0.997	0.780	1.282
TC/HDL	0.093	0.014	0.240	6.505	0.000	0.945	1.058
UPb	0.007	0.002	0.108	2.929	0.004	0.952	1.065
UCd	0.010	0.004	0.101	2.745	0.006	0.939	1.050

Note. HDLC, high-density lipoprotein cholesterol; TC, total cholesterol; UCd, urinary cadmium; UPb, urinary lead; VIF, variance inflation factor.

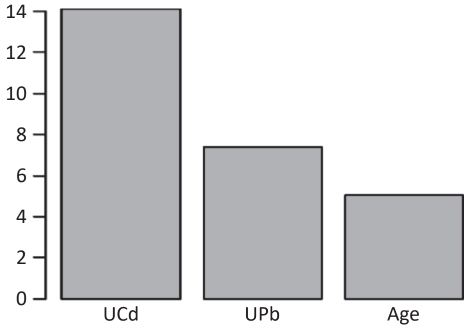
Supplementary Table S5. Results of multicollinearity testing for urinary N-acetyl-β-δ-glucosaminidase (UNAG)

Variables	Non-standardized coefficient		Standard coefficient	*t*-value	Sig.	Collinearity statistics	
	Beta	Standard error	Beta			Tolerance	VIF
(constant)	−0.050	0.149		−0.337	0.736		
Age	0.009	0.002	0.229	5.117	0.000	0.858	1.166
Educational level	−0.054	0.050	−0.049	−1.084	0.279	0.842	1.187
UPb	0.006	0.003	0.091	2.149	0.032	0.958	1.044
UCd	0.010	0.005	0.093	2.170	0.030	0.945	1.058

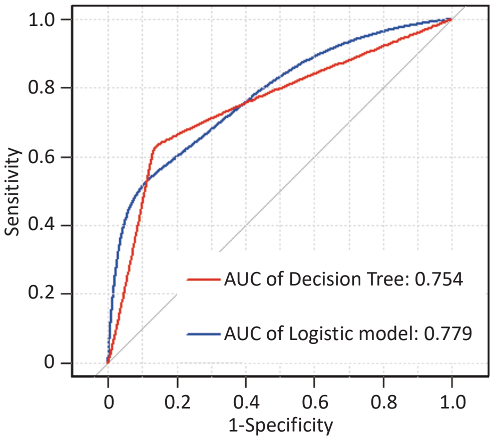
Note. UCd, urinary cadmium; UPb, urinary lead; VIF, variance inflation factor.



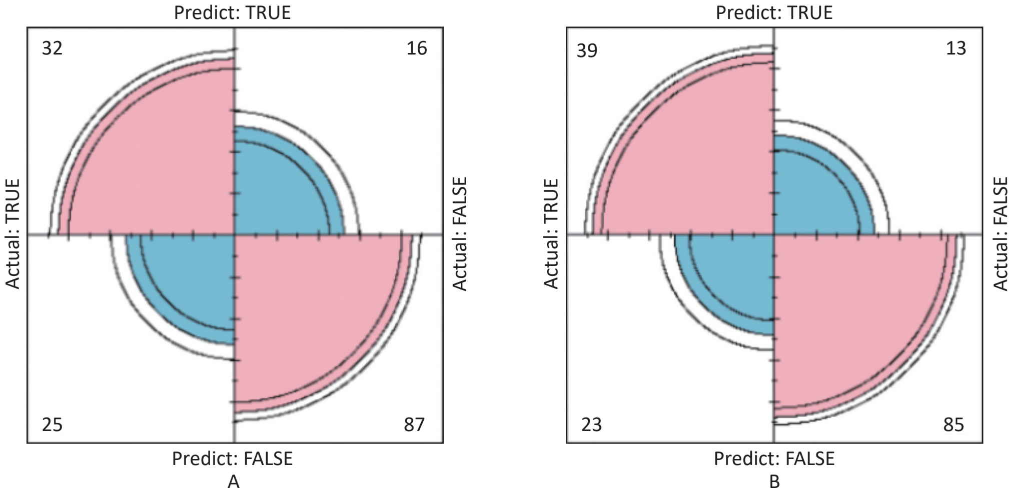
Supplementary Figure S1. Ranking of the Importance of Risk Factors for estimated glomerular filtration rate (eGFR). Using eGFR as a renal function indicator, UPb and UCd levels are more important predictive variables. eGFR, estimated glomerular filtration rate; HDLC, high-density lipoprotein cholesterol; TC, total cholesterol; UCd, urinary cadmium; UPb, urinary lead.



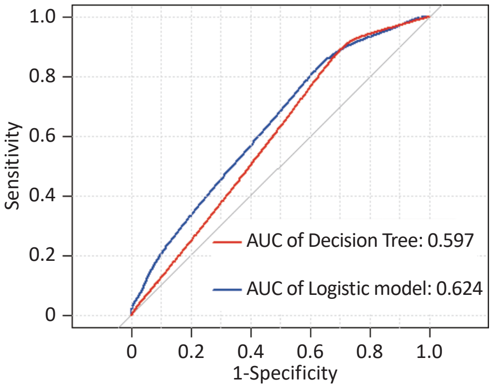
Supplementary Figure S2. Ranking of the Importance of Risk Factors for urinary N-acetyl-β-δ-glucosaminidase (UNAG). Using UNAG as a renal function indicator, UCd and UPb levels are more important predictive variables. UCd, urinary cadmium; UPb, urinary lead.



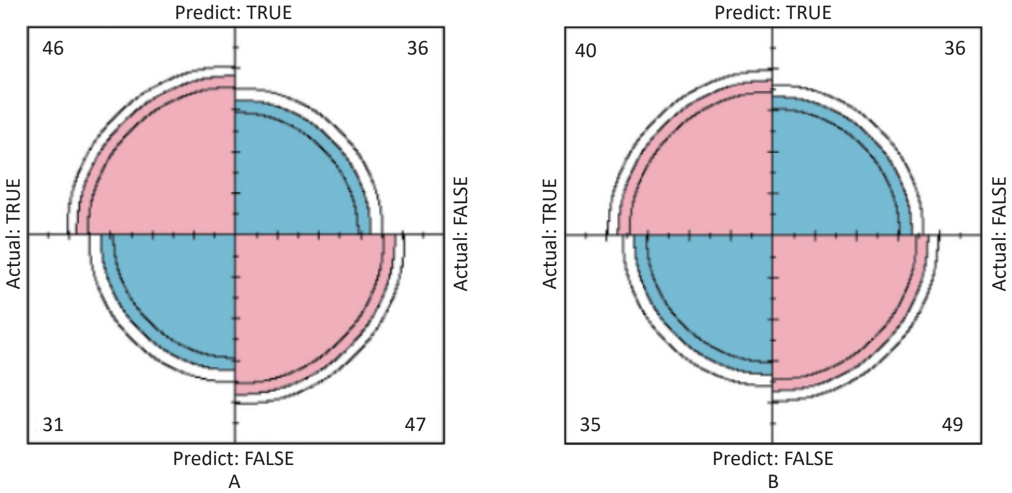
Supplementary Figure S3. Receiver operating characteristic (ROC) curve of the eGFR prediction model. The AUCs were all greater than 0.5, indicating that the model evaluation was effective. AUC, area under the curve; ROC, receiver operating characteristic; eGFR, estimated glomerular filtration rate.



Supplementary Figure S4. The prediction model confusion matrix of eGFR. (A) Confusion matrix for eGFR logistic regression; (B) Confusion matrix for eGFR classification decision tree. The pink part in the figure represents the correct prediction and the blue part represents the incorrect prediction. With eGFR as an indicator of renal function, classification decision tree models have higher sensitivity, specificity, precision and accuracy than logistic regression (True positive is shown by the number in the upper left corner, false positive is shown by the number in the upper right corner, false negative is shown by the number in the lower left corner, and true negative is shown by the number in the lower right corner). eGFR, estimated glomerular filtration rate.



Supplementary Figure S5. Receiver operating characteristic (ROC) curve of the UNAG prediction model. The AUCs were all greater than 0.5, indicating that the model evaluation was effective. AUC, area under the curve; ROC, receiver operating characteristic; UNAG, urinary N-acetyl-β-δ-glucosaminidase



Supplementary Figure S6. Confusion matrix of UNAG prediction model. (A) Confusion matrix for UNAG logistic regression; (B) Confusion matrix for UNAG classification decision tree. The pink part in the figure represents the correct prediction and the blue part represents the incorrect prediction. With UNAG as the indicator of renal function, the sensitivity and specificity of the classification decision tree model are higher than those of the logistic regression (True positive is shown by the number in the upper left corner, false positive is shown by the number in the upper right corner, false negative is shown by the number in the lower left corner, and true negative is shown by the number in the lower right corner). UNAG, urinary N-acetyl-β-δ-glucosaminidase