Applying Random Forest Model Algorithm to GFR Estimation

Peijia Liu
The Third Affiliated Hospital of Sun Yet-sun University Department of Nephrology
Dong Yang
Guangzhou AD technology Co., Ltd
Shaomin Li
The Third Affiliated Hospital of Sun Yet-sun University Department of Nephrology
Yutian Chong
Third Affiliated Hospital of Sun Yat-Sen University
Wentao Hu
The Third Affiliated Hospital of Sun Yet-sun University Department of Nephrology
Bohan Wang
Guangzhou AD technology Co., Ltd
Xun Liu (naturestyle@163.com)
The Third Affiliated Hospital of Sun Yet-sun University Department of Nephrology

Research article

Keywords: Glomerular filtration rate, Prediction equation, Random forest regression, Choronic kidney disease

Posted Date: October 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-74843/v1

Abstract

Background The utilization of estimating-GFR equations is critical for kidney disease in the clinic. However, the performance of the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation has not improved substantially in the past eight years. Here we hypothesized that random forest regression(RF) method could go beyond revised linear regression, which is used to build the CKD-EPI equation

Methods 1732 participants were enrolled in this study totally (1333 in development data set from Tianhe District and 399 in external data set Luogang District). Recursive feature elimination (RFE) is applied to the development data to select important variables and build random forest models. Then same variables were used to develop the estimated GFR equation with linear regression as a comparison. The performances of these equations are measured by bias, 30% accuracy , precision and root mean square error(RMSE).

Results Of all the variables, creatinine, cystatin C, weight, body mass index (BMI), age, uric acid(UA), blood urea nitrogen(BUN), hematocrit(HCT) and apolipoprotein B(APOB) were selected by RFE method. The results revealed that the overall performance of random forest regression models ascended the revised regression models based on the same variables. In the 9-variable model, RF model was better than revised linear regression in term of bias, precision ,30%accuracy and RMSE(0.78 vs 2.98, 16.90 vs 23.62, 0.84 vs 0.80, 16.88 vs 18.70, all P<0.01). In the 4-variable model, random forest regression model showed an improvement in precision and RMSE compared with revised regression model. (20.82 vs 25.25, P<0.01, 19.08 vs 20.60, P<0.001). Bias and 30%accurancy were preferable, but the results were not statistically significant (0.34 vs 2.07, P=0.10, 0.8 vs 0.78, P=0.19, respectively).

Conclusions The performances of random forest regression models are better than revised linear regression models when it comes to GFR estimation.

Background

The application of the estimating-GFR equation in the clinic is critical, especially when it comes to diagnosing and prognosis of nephropathy, therapeutic interventions as well as drug dosing evaluation(1). It is impractical to get true GFR directly, because GFR is changing overtime. Instead, we could get a relatively reliable index by the measurement of serum or urine clearance of exogenous and endogenous filtration markers(2, 3). However, high expenditure, difficulty in obtaining and complicated procedures make these methods inconvenient for daily use. As a result, estimated GFR (eGFR) equations could be an optimal choice for routine GFR evaluation. The 2012 CKD-EPI equation built by linear regression with creatinine, cystatin C, sex, race and age as variables have been considered to be the best equation until now(4). Nevertheless, serum concentrations of creatinine and cystatin C are affected by GFR and non-GFR determinants, which will cause inaccurate GFR prediction(5). Age race and sex are clinical characteristics and demographics and can offset a proportion of the impacts of non-GFR

determinants(6). But they may be not sufficient. Because there are numerous conditions including UA(7), urinary protein(8), anaemia(9) and smoking(10) that can affect GFR and many metabolites are influenced by glomerular filtration rate(11, 12). Thus more variables may be needed to optimize the CKD equation.

On the other hand, Previous researches indicate that machine learning algorithms contribute to a specific improvement of eGFR Eq. (13–16). Random forest regression is a machine learning algorithm which can train and predict samples by ensembling regression trees(17). Since the randomness injection and the nature of bagging ensemble of regression trees, random forest regression is capable of dealing with relatively small samples with large amount of variables. As a result, this algorithm has been used instead of linear regression in various fields (18–20), especially when the number of variables is relatively large.

In this setting, we assumed that the random forest regression with more variables imported could enhance the performance of the evaluation equation for GFR.

Methods

Data sources

The study was based on the data from the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou. To be eligible for the study, participants should measure their GFR by the mean of radionuclide renal dynamic imaging or the dual plasma diethylenetriamine pentaacetic acid (DTPA) and should be older than 18 years old. Exclusion criterion: 1) amputation; 2) acute renal insufficiency; 3) hemodialysis, peritoneal dialysis, renal transplantation, and chronic kidney disease in patients with obstructive nephropathy; 4) severe oedema; 5) skeletal muscle atrophy; 6) hydrothorax or ascetic fluid; 7) dystrophy; 8) under 18 years of age; 9) heart failure; 10) ketoacidosis. Participants who are daily use cimetidine or trimethoprim would not be included in the study. Among them, data deletion criteria included(21): 1) previous alternative renal replacement treatment experience; 2) mGFR□5 ml/min/1.73m^2. 4504 participants were screened from January 2012 to April 2015, and 1732 individuals were included in the study (the details were demonstrated in table1). Data collected from 1333 participants (923 participants for modelling data set and 399 participants for internal validation data set) in Tianhe District of the Third Affiliated Hospital of Sun Yat-sen University were regarded as development set. External validation set Data set used the data from 399 individuals in Luogang District of the Third Affiliated Hospital of Sun Yat-sen University. Subjects who had been enrolled since July 25, 2017 signed written informed consent. For those who were previously enrolled, we contacted the participants or their families by phones or letters and obtained informed consents. For those who were unable to contact, Institutional Review Board at the Third Affiliated Hospital of Sun Yat-sen University provided the ethical review approval of an exemption from written informed consent.

Laboratory methods

The measurement of standard GFR (sGFR) or measured GFR (mGFR) was radionuclide renal dynamic imaging or 99mTc-DTPA, which was described in detail before (13-15, 22). Serum creatinine (SC) measurement was on Hitachi 7180 auto-analyzer (Hitachi, Tokyo, Japan) using reagents from Roche Diagnostics (Mannheim, Germany). The enzymatic method was used for SC measurement. Isotope dilution mass spectrometry reference method was performed on the calibration of SC level after the year 2010. Serum Cystatin C determination was with reference of standard reference material (SRM 967).

Mathematical methods

Variable selection

Nine variables, including creatinine, cystatin C, weight, BMI, age, UA, BUN, HCT and APOB, were selected by random forest RFE approach(23).

Development of the smooth line revised linear regression equation Model

Revised regression models of four variables(2012 CKD-EPI equation) and nine variables (selected by RFE approach) were implemented as comparisons. The process of equation development has carried out in detail in the paper we had published before(14, 24)and was shown in the Additional file.

Development of the New Prediction Model by Random Forest

The four and the nine variables mentioned before were included in the development of models by the method of random forest algorithm respectively. Tunning parameters of random forest were selected by grid search and 5-fold cross validation. Then the external validation set was used for the evaluation of the equation. The construction of the model was described in detail in Additional file.

Statistical methods

Accuracy, bias, precision and RMSE were the primary outcomes of the evaluation. Accuracy was the proportion of eGFR that was not more than 30% deviation from the mGFR. Bias was defined as the median difference of the mGFR and eGFR; precision was assessed as the interquartile range (IQR) of the difference. RMSE is the square root of the ratio of the sum of the square of the deviation of the observed value from the true value and the number of observations. Bootstrap method(2000 bootstraps) was used for the 95% confidence intervals (CIs) (25). Independent samples t-test or the Mann–Whitney test was used when comparing quantitative variables between two data-sets. Difference and accuracy were performed by Wilcoxon signed-rank test and McNemar test, respectively. Statistical significance level was $P < 0.05$. All calculations and statistics were conducted by SPSS 20.0, R 3.53 software and Python 3.7 software.

Results

Study population

There were significant differences between development datasets and external validation in some baseline characteristics. As it is shown in the table1, Blood lipid including CHOL, LDL and APOA were lower in development and internal validation set compared with the external validation set (4.82±1.44 vs 4.98±1.41, *P*=0.049; 2.95±1.14 vs 3.09±1.18, *P*=0.037; 1.28±0.25 vs 1.31±0.25, *P*=0.026; respectively). And the population in development and internal validation set used more glucocorticoids for treatment (*P*=0.04). mGFR, age, serum creatinine and other characteristics did not differ significantly.

GFR estimation models performance

Of all the variables, creatinine, cystatin C, weight, BMI, age, UA, BUN, HCT and APOB were selected by the RFE approach. The results revealed that the overall performance of random forest regression models ascended the revised regression models based on the same variables. 9-variable random forest regression model was optimal. In the 9-variable model, random forest regrssion model was better than revised linear regression in terms of bias, precision, 30%accuracy and RMSE (0.78 vs 2.98, 16.90 vs 23.62, 0.84 vs 0.80, 16.88 vs 18.70, all *P*<0.01). In the 4-variable model, precision and RMSE improved compared random forest regression model with revised regression model (20.82 vs 25.25, *P*<0.01, 19.08 vs 20.60, P<0.001, respetively). Bias and 30%accurancy were preferable, but the results were not statistically significant (0.34 vs 2.07, *P*=0.10, 0.8 vs 0.78, P=0.19, respectively).

Discussion

We performed revised linear regression and RF modelling for four and nine variables and performed a head-to-head comparison analysis of the same variable combinations. Our results demonstrate that eGFR equation based on RF model performed better than revised linear regression model on GFR estimation. Our findings can help clinicians better understand the patient's condition and optimize their treatment plan.

The advance of computer science and statistics has made machine learning an increasingly vital method. A series of machine learning methods contribute to a specific improvement for eGFR Eq. (13–15). However, no prediction equation can comprehensively exceed the CKD-EPI equation. Random forests are also a kind of ensemble-based machine learning method, which can find non-linear relationship that lies underneath data and make better classification or regression results. The random forest method has the advantage of producing a high-precision classifier, dealing with multi-class variables, reducing errors in balanced classification and preventing over-fitting. Many papers compared linear regression with random forest algorithms in other fields. As we can see, the random forest algorithm has not performed better in all field and aspects (26–28), this reveals that random forest method can only take advantage over linear regression in some data models. So a more significant number of multi-centre data are needed to validate our outcomes in the field of GFR estimation.

Non-GFR determinants affect the performance of the eGFR equation. In other words, the decrement of the influence of non-GFR determinants can elevate equation performance. An article points out that multimetabolite panels can improve performance of the eGFR equation even without age and gender as

variables (11, 29). Perhaps these factors can reduce the impact of non-GFR determinants. Possibly, we could incorporate more data dimensions to improve the prediction performance of the eGFR equation.

There were significant differences in CHOL, LDL.C, APOC and the use of glucocorticoids in the modelling and validation groups. According to the literature, the ratio of blood lipids and blood lipidosis indeed related to renal function [20, 21]. For glucocorticoids, it is well known that it is a treatment for many kidney diseases. Inevitably, these factors would interfere with the glomerular filtration rate. Therefore, these differences between modelling and validation groups can have an impact on the performance of all equations. Fortunately, these variables were not included in the prediction equation. Of note, sex did not include in the variables of RF selection, and maybe gender was not weighted enough among these variables. As for race, most of the participants are Han Chinese, which will limit its applicability.

Our study had several limitations. Firstly, the data collectionwas only one time. There might be measurement error in single data modelling, and medical data from the same patients should be collected and modelled repeatedly. Secondly, it may be impossible to obtain "true GFR", and Levey recommends plasma clearance of iohexol and urine clearance of iothalamate as mGFR for GFR Eq. (30, 31). In our study, mGFR was measured and calibrated by 99mTc-DTPA renal dynamic imaging. This measurement method will produce relatively obvious deviations both at high GFR and low GFR. (32, 33). Hence, the system and measurement errorcould not be avoided. Besides, mGFR should be verified to reduce the measurement error. Repeated measurement of serum of creatinine is also recommended (31, 34). Finally, samples of the validation set are from a single centre and are small in number. Furthermore, this study did not re-evaluate the predictive efficacy of subgroup classification for age, diabetes, and CKD staging.

Conclusions

Random forest RFE was used to select variables and to develop eGFR equations. The performances of random forest regrssion models are better than revised linear regression models based on the same variables. Much attention should be put to the random forest regression because it might contribute to a specific improvement to the eGFR evaluation equation.

List Of Abbreviations

ACEI	Angiotensin converting enzyme inhibitor
ALB	Albumin
APOA	Apolipoprotein A
APOB	apolipoprotein B
ARB	Angiotensin receptor blockers
BMI	Body mass index
BUN	Blood urea nitrogen
CCB	Calcium channel blocker
CIs	confidence intervals
CHD	Chronic heart disease
CHOL	total cholesterol
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
DBP	Diastolic blood pressure
DTPA	Dual plasma diethylenetriamine pentaacetic acid
eGFR	Estimated GFR
EPO	Erythropoietin
FBS	Fasting blood sugar
GFR	Glomerular filtration rate
HCT	hematocrit
HDL.C	HDL-C High-density lipoprotein cholesterol
HGB	Hemoglobin
IQR	interquartile range
LDL	low-density lipoprotein
LDL.C	Low-density lipoprotein cholesterol
LPA	Lipoprotein A
MCV	Mean corpuscular volume
MCHC	Mean corpuscular hemoglobin Concentration
mGFR	Measured GFR
PA	Prealbumin

RF	Random forest
RFE	Recursive feature elimination
SBP	Systolic blood pressure
SC	Serum creatinine
sGFR	Standard GFR
TRI	Triglyceride
UA	Uric acid

Declarations

Acknowledgement

We would like to thank all the doctors, nurses, technicians, and patients involved in this study for their cooperation. We would also thank Xiaoshuai Huang and Chenguang Shi for their kind suggestions of mathematic modeling and statistical analysis.

Funding

This study was supported by the National Key R&D Plan (Grant No.2018YFC1315400, the National Natural Science Foundation of China (Grant No. 81873631, Grant No.81370866 and Grant No. 81070612), the China Postdoctoral Science Foundation (Grant No. 201104335), the Third Affiliated Hospital of Sun Yat-Sen University Clinical Research Program (Grant No. YHJH201806).

Availability of data and material

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

Authors' contributions

Research idea and study design: XL; data acquisition: XL, YTC; data analysis/interpretation: PJL DY; statistical analysis: DY; supervision or mentorship: XL,WTH, SML, BHW and XL. Each author contributed important intellectual content during manuscript drafting or revision and accepts accountability for the overall work by ensuring that questions pertaining to the accuracy or integrity of any portion of the work are appropriately investigated and resolved. XL takes responsibility that this study has been reported honestly, accurately, and transparently; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained. All authors read and approved the final manuscript.

Competing interests

Authors declare no conflict of interest.

Consent for publication

Yes.

Ethics approval and consent to participate

Ethics has been authorized by the Ethics Committee of the Third Affiliated Hospital of Sun Yat-sen University and has been described in detail in previously published articles.

References

1. Stevens LA, Coresh J, Greene T, Levey AS. Assessing kidney function--measured and estimated glomerular filtration rate. The New England journal of medicine. 2006;354(23):2473-83.
2. Levey AS, Inker LA. GFR as the "Gold Standard": Estimated, Measured, and True. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2016;67(1):9-12.
3. Levey AS, Inker LA, Coresh J. GFR estimation: from physiology to public health. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2014;63(5):820-34.
4. Levey AS, Coresh J, Tighiouart H, Greene T, Inker LA. Measured and estimated glomerular filtration rate: current status and future directions. Nature reviews Nephrology. 2020;16(1):51-64.
5. Inker LA, Levey AS, Coresh J. Estimated Glomerular Filtration Rate From a Panel of Filtration Markers-Hope for Increased Accuracy Beyond Measured Glomerular Filtration Rate? Advances in chronic kidney disease. 2018;25(1):67-75.
6. Levey AS, Coresh J, Tighiouart H, Greene T, Inker LA. Measured and estimated glomerular filtration rate: current status and future directions. Nature reviews Nephrology. 2020;16(1):14.
7. Ficociello LH, Rosolowsky ET, Niewczas MA, Maselli NJ, Weinberg JM, Aschengrau A, et al. High-normal serum uric acid increases risk of early progressive renal function loss in type 1 diabetes: results of a 6-year follow-up. Diabetes Care. 2010;33(6):1337-43.
8. Hunsicker LG, Adler S, Caggiula A, England BK, Greene T, Kusek JW, et al. Predictors of the progression of renal disease in the Modification of Diet in Renal Disease Study. Kidney international. 1997;51(6):1908-19.
9. Gu L, Lou Q, Wu H, Ouyang X, Bian R. Lack of association between anemia and renal disease progression in Chinese patients with type 2 diabetes. Journal of diabetes investigation. 2016;7(1):42-7.
10. Xia J, Wang L, Ma Z, Zhong L, Wang Y, Gao Y, et al. Cigarette smoking and chronic kidney disease in the general population: a systematic review and meta-analysis of prospective cohort studies. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. 2017;32(3):475-87.

11. Coresh J, Inker LA, Sang Y, Chen J, Shafi T, Post WS, et al. Metabolomic profiling to improve glomerular filtration rate estimation: a proof-of-concept study. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. 2019;34(5):825-33.
12. Goek ON, Prehn C, Sekula P, Römisch-Margl W, Döring A, Gieger C, et al. Metabolites associate with kidney function decline and incident chronic kidney disease in the general population. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. 2013;28(8):2131-8.
13. Liu X, Li N, Lv L, Fu Y, Cheng C, Wang C, et al. Improving precision of glomerular filtration rate estimating model by ensemble learning. Journal of translational medicine. 2017;15(1):231.
14. Liu X, Li NS, Lv LS, Huang JH, Tang H, Chen JX, et al. A comparison of the performances of an artificial neural network and a regression model for GFR estimation. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2013;62(6):1109-15.
15. Liu X, Pei X, Li N, Zhang Y, Zhang X, Chen J, et al. Improved glomerular filtration rate estimation by an artificial neural network. PloS one. 2013;8(3):e58242.
16. Li N, Huang H, Qian H-Z, Liu P, Lu H, Liu X. Improving accuracy of estimating glomerular filtration rate using artificial neural network: model development and validation. Journal of translational medicine. 2020;18(1):120.
17. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
18. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT. Application of random forests methods to diabetic retinopathy classification analyses. PloS one. 2014;9(6):e98587.
19. Ai F-f, Bin J, Zhang Z-m, Huang J-h, Wang J-b, Liang Y-z, et al. Application of random forests to select premium quality vegetable oils by their fatty acid composition. Food Chemistry. 2014;143:472-8.
20. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC genetics. 2010;11:49.
21. Inker LA, Schmid CH, Tighiouart H, Eckfeldt JH, Feldman HI, Greene T, et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. The New England journal of medicine. 2012;367(1):20-9.
22. Liu X, Gan X, Chen J, Lv L, Li M, Lou T. A new modified CKD-EPI equation for Chinese patients with type 2 diabetes. PloS one. 2014;9(10):e109743.
23. Granitto P.M. FC, Biasioli F., Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems. 2006;82(2):8.
24. Chen J, Tang H, Huang H, Lv L, Wang Y, Liu X, et al. Development and validation of new glomerular filtration rate predicting models for Chinese patients with type 2 diabetes. Journal of translational medicine. 2015;13:317.
25. Johnson RW. An Introduction to the Bootstrap[J]. Teaching Statistics. 2010;23(2):49-54.

26. Zaorska K, Zawierucha P, Nowicki M. Prediction of skin color, tanning and freckling from DNA in Polish population: linear regression, random forest and neural network approaches. Hum Genet. 2019;138(6):635-47.
27. Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, et al. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ Int. 2019;130:104934.
28. Yuchi W, Gombojav E, Boldbaatar B, Galsuren J, Enkhmaa S, Beejin B, et al. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environ Pollut. 2019;245:746-53.
29. Freed TA, Coresh J, Inker LA, Toal DR, Perichon R, Chen J, et al. Validation of a Metabolite Panel for a More Accurate Estimation of Glomerular Filtration Rate Using Quantitative LC-MS/MS. Clinical chemistry. 2019;65(3):406-18.
30. Stevens LA, Levey AS. Measured GFR as a confirmatory test for estimated GFR. Journal of the American Society of Nephrology : JASN. 2009;20(11):2305-13.
31. Levey AS, Coresh J, Tighiouart H, Greene T, Inker LA. Strengths and limitations of estimated and measured GFR. Nature reviews Nephrology. 2019;15(12):784.
32. Itoh K, Tsushima S, Tsukamoto S. Comparison of methods for determination of glomerular filtration rate: 99mTc-DTPA renography, predicted creatinine clearance method and plasma sample method. Ann Nucl Med. 2003;7:561-5.
33. Natale G, Pirtro A, Massimo C. Measurement of glomerular Filtration Rate by the 99mTc-DTPA Renal dynamic imaging Is Less Precise than Measured and Predicted Creatinine Clearance. Nephron. 1999;81:136-40.
34. Kwong Y-TD, Stevens LA, Selvin E, Zhang YL, Greene T, Van Lente F, et al. Imprecision of urinary iothalamate clearance as a gold-standard measure of GFR decreases the diagnostic accuracy of kidney function estimating equations. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2010;56(1):39-49.

Tables

Table 1
Characteristics of participants

characteristic	Overall (1732)	Modeling group (n = 1333)	External validation group(n = 399)	P value
Number of participants	1732	1333	399	-
gender(male)[n(%)]	1047(60.5%)	805(60.4%)	242(60.7%)	*P = 0.925*
mGFR(ml/min/1.73 m^2)	87.01 ± 37.6	87.38 ± 38.18	85.80 ± 35.75	*P* = 0.463
age(years old)	57.15 ± 13.56	57.03 ± 13.60	57.53 ± 13.42	*P* = 0.520
SBP(mmHg)	137.67 ± 21.45	137.14 ± 21.55	139.47 ± 21.05	*P* = 0.057
DBP(mmHg)	79.90 ± 13.52	79.60 ± 13.38	80.90 ± 13.94	*P* = 0.078
BMI(Kg/m^2)	24.57 ± 3.60	24.50 ± 3.61	24.78 ± 3.53	*P* = 0.185
mGFR(ml/min/1.73^2)	87.01 ± 37.63	87.37 ± 38.2	85.80 ± 35.76	*P* = 0.463
Hemoglobin(g/L)	123.89 ± 23.34	124.04 ± 23.07	123.38 ± 24.22	*P* = 0.623
Hematocrit (%)	0.37 ± 0.06	0.37 ± 0.06	0.36 ± 0.07	*P* = 0.616
MCV(fL)	85.80 ± 7.77	85.73 ± 7.93	86.03 ± 7.20	*P* = 0.504
MCHC(g/L)	337.90 ± 18.51	338.28 ± 15.21	336.61 ± 26.71	*P* = 0.114
Kalium(mmol/L)	4.04 ± 0.48	4.04 ± 0.48	4.04 ± 0.48	*P* = 0.951
Natriumion(mmol/L)	140.59 ± 3.06	140.54 ± 3.15	140.77 ± 2.74	*P* = 0.184

Note: Unless noted, categorical variables are expressed as percentages, continuous variables are expressed as mean plus or minus standard deviation; "*"= P □ 0.05. Citation abbreviations in the table are described in detail in the acronym table. Low-density lipoprotein cholesterol, HDL-C High-density lipoprotein cholesterol, TRI Triglyceride, CHOL Total cholesterol, BMI Body mass index, CO_3^- bicarbonate radical, ALB albumin, FBS Fasting blood sugar, BUN Blood urea nitrogen, UA Uric acid, PA Prealbumin, APOA Apolipoprotein A, APOB Apolipoprotein B, HCT Hematocrit, ,PA Prealbumin, LPA Lipoprotein A, TRI Triglyceride, EPO Erythropoietin, CCB Calcium channel blocker, ARB Angiotensin receptor blockers, ACEI Angiotensin converting enzyme inhibitor, URI PRO Urine protein, HBP high blood pressure, CHD chronic heart disease, MCV Mean Corpuscular Volume, MCHC Mean Corpuscular Hemoglobin Concentration,

characteristic	Overall (1732)	Modeling group (n = 1333)	External validation group(n = 399)	P value
Chloridion(mmol/L)	104.08 ± 3.69	104.06 ± 3.75	104.14 ± 3.46	*P* = 0.672
Calcium(mmol/L)	2.30 ± 0.15	2.30 ± 0.15	2.29 ± 0.16	*P* = 0.171
Phosphate(mmol/L)	1.23 ± 0.25	1.23 ± 0.25	1.23 ± 0.26	*P* = 0.644
HCO_3^- (mmol/L)	23.25 ± 2.93	23.20 ± 2.97	23.44 ± 2.81	*P* = 0.150
ALB(g/L)	38.40 ± 4.91	38.38 ± 4.86	38.46 ± 5.09	*P* = 0.770
PA(g/L)	248.05 ± 68.70	247.86 ± 69.96	248.70 ± 65.78	*P* = 0.692
FBS(mmol/L)	7.51 ± 3.62	7.55 ± 3.67	7.36 ± 3.42	*P* = 0.348
BUN(mmol/L)	8.48 ± 5.80	8.43 ± 5.82	8.63 ± 5.74	*P* = 0.563
UA(umol/L)	423.18 ± 131.68	421.84 ± 134.39	427.66 ± 122.22	*P* = 0.439
CHOL* (umol/L)	4.86 ± 1.43	4.82 ± 1.44	4.98 ± 1.41	**P = 0.049**
HDL.C(umol/L)	1.07 ± 0.30	1.07 ± 0.30	1.09 ± 0.31	*P* = 0.250
LDL.C* (umol/L)	2.98 ± 1.15	2.95 ± 1.14	3.09 ± 1.18	**P = 0.037**
APOA* (g/L)	1.28 ± 0.25	1.28 ± 0.25	1.31 ± 0.25	**P = 0.026**
APOB(g/L)	1.19 ± 0.45	1.18 ± 0.45	1.22 ± 0.44	*P* = 0.203

Note: Unless noted, categorical variables are expressed as percentages, continuous variables are expressed as mean plus or minus standard deviation; "*"= P □ 0.05. Citation abbreviations in the table are described in detail in the acronym table. Low-density lipoprotein cholesterol, HDL-C High-density lipoprotein cholesterol, TRI Triglyceride, CHOL Total cholesterol, BMI Body mass index, CO_3^- bicarbonate radical, ALB albumin, FBS Fasting blood sugar, BUN Blood urea nitrogen, UA Uric acid, PA Prealbumin, APOA Apolipoprotein A, APOB Apolipoprotein B, HCT Hematocrit, ,PA Prealbumin, LPA Lipoprotein A, TRI Triglyceride, EPO Erythropoietin, CCB Calcium channel blocker, ARB Angiotensin receptor blockers, ACEI Angiotensin converting enzyme inhibitor, URI PRO Urine protein, HBP high blood pressure, CHD chronic heart disease, MCV Mean Corpuscular Volume, MCHC Mean Corpuscular Hemoglobin Concentration,

characteristic	Overall (1732)	Modeling group (n = 1333)	External validation group(n = 399)	P value
LPA(mmol/L)	240.65 ± 265.32	239.27 ± 263.84	245.27 ± 271.47	*P* = 0.692
TRI(mmol/L)	1.89 ± 1.76	1.89 ± 1.78	1.90 ± 1.70	*P* = 0.903
Cystatin C(mg/L)	1.55 ± 1.1.05	1.53 ± 1.02	1.60 ± 1.15	*P* = 0.225
Creatinine(mg/dl)	1.72 ± 1.80	1.69 ± 1.73	1.81 ± 2.03	*P* = 0.256
Medical treatment				
Albumin used the day of GFR measurement	20(1.2%)	15(1.1%)	5(1.3%)	*P* = 0.834
Diuretics used the day GFR measurement	234(13.5%)	181(13.6%)	53(13.3%)	*P* = 0.880
hormone*	35(2.0%)	32(2.4%)	3(0.8%)	**P = 0.04**
Immunosuppressant	10(0.6%)	8(0.6%)	2(0.5%)	*P* = 0.891
Uric acid lowering drug	198(11.4%)	155(11.6%)	43(10.8%)	*P* = 0.639
Lipid-lowering drugs	1045(60.3%)	794(59.6%)	251(62.9%)	*P* = 0.231
EPO	77(4.4%)	55(4.1%)	22(5.5%)	*P* = 0.238
Iron supplement	149(8.6%)	111(8.3%)	38(9.5%)	*P* = 0.564
Calcium supplement	228(13.2%)	164(12.3%)	64(16.0%)	*P* = 0.053

Note: Unless noted, categorical variables are expressed as percentages, continuous variables are expressed as mean plus or minus standard deviation; "*"= P □ 0.05. Citation abbreviations in the table are described in detail in the acronym table. Low-density lipoprotein cholesterol, HDL-C High-density lipoprotein cholesterol, TRI Triglyceride, CHOL Total cholesterol, BMI Body mass index, CO_3^- bicarbonate radical, ALB albumin, FBS Fasting blood sugar, BUN Blood urea nitrogen, UA Uric acid, PA Prealbumin, APOA Apolipoprotein A, APOB Apolipoprotein B, HCT Hematocrit, ,PA Prealbumin, LPA Lipoprotein A, TRI Triglyceride, EPO Erythropoietin, CCB Calcium channel blocker, ARB Angiotensin receptor blockers, ACEI Angiotensin converting enzyme inhibitor, URI PRO Urine protein, HBP high blood pressure, CHD chronic heart disease, MCV Mean Corpuscular Volume, MCHC Mean Corpuscular Hemoglobin Concentration,

characteristic	Overall (1732)	Modeling group (n = 1333)	External validation group(n = 399)	P value
Active vitamin D3	167(9.6%)	123(9.2%)	44(11.0%)	*P* = 0.283
Inactive vitamin D3	7(0.4%)	6(0.5%)	1(0.3%)	*P* = 0.582
Beta blocker	340(19.6%)	254(19.1%)	86(21.6%)	*P* = 0.270
Alpha blocker	54(3.1%)	44(3.3%)	10(2.5%)	*P* = 0.423
CCB	671(38.7%)	518(38.9%)	153(38.3%)	*P* = 0.853
ACEI	83(4.8%)	63(4.7%)	20(5.0%)	*P* = 0.814
ARB	866(50.0%)	667(50.0%)	199(49.9%)	*P* = 0.952
Diuretic	232(13.4%)	182(13.7%)	50(12.5%)	*P* = 0.660
Other variables				
Smoke	259(15.0%)	208(15.6%)	51(12.8%)	*P* = 0.166
Drink	159(9.2%)	115(8.6%)	44(11.0%)	*P* = 0.145
URI PRO	678(39.1%)	520(39.0%)	158(40.0%)	*P* = 0.825
HBP	1062(61.3%)	812(60.9%)	250(62.7%)	*P* = 0.531
Diabetes	1299(75.0%)	992(74.4%)	307(76.9%)	*P* = 0.307

Note: Unless noted, categorical variables are expressed as percentages, continuous variables are expressed as mean plus or minus standard deviation; "*"= P □ 0.05. Citation abbreviations in the table are described in detail in the acronym table. Low-density lipoprotein cholesterol, HDL-C High-density lipoprotein cholesterol, TRI Triglyceride, CHOL Total cholesterol, BMI Body mass index, CO_3^- bicarbonate radical, ALB albumin, FBS Fasting blood sugar, BUN Blood urea nitrogen, UA Uric acid, PA Prealbumin, APOA Apolipoprotein A, APOB Apolipoprotein B, HCT Hematocrit, ,PA Prealbumin, LPA Lipoprotein A, TRI Triglyceride, EPO Erythropoietin, CCB Calcium channel blocker, ARB Angiotensin receptor blockers, ACEI Angiotensin converting enzyme inhibitor, URI PRO Urine protein, HBP high blood pressure, CHD chronic heart disease, MCV Mean Corpuscular Volume, MCHC Mean Corpuscular Hemoglobin Concentration,

characteristic	Overall (1732)	Modeling group (n = 1333)	External validation group(n = 399)	P value
CHD	344(19.9%)	262(19.7%)	82(20.6%)	*P* = 0.694
Stroke	49(2.8%)	43(3.2%)	6(1.5%)	*P* = 0.069
mGFR ≥ 60 ml/min/1.73 m2	1298(74.9%)	999(74.9%)	299(74.9%)	
mGFR□60 ml/min/1.73 m2	434(25.1%)	334(25.1%)	100(25.1%)	*P* = 0.890
				P = 0.910

Note: Unless noted, categorical variables are expressed as percentages, continuous variables are expressed as mean plus or minus standard deviation; "*"= P □ 0.05. Citation abbreviations in the table are described in detail in the acronym table. Low-density lipoprotein cholesterol, HDL-C High-density lipoprotein cholesterol, TRI Triglyceride, CHOL Total cholesterol, BMI Body mass index, CO_3^- bicarbonate radical, ALB albumin, FBS Fasting blood sugar, BUN Blood urea nitrogen, UA Uric acid, PA Prealbumin, APOA Apolipoprotein A, APOB Apolipoprotein B, HCT Hematocrit, ,PA Prealbumin, LPA Lipoprotein A, TRI Triglyceride, EPO Erythropoietin, CCB Calcium channel blocker, ARB Angiotensin receptor blockers, ACEI Angiotensin converting enzyme inhibitor, URI PRO Urine protein, HBP high blood pressure, CHD chronic heart disease, MCV Mean Corpuscular Volume, MCHC Mean Corpuscular Hemoglobin Concentration,

Table 2
Comparison of random forest regression model and revised CKD-EPI model

Model	Num of Variables	Bias	Precision	Accuracy(30%)	RMSE
RF regression	9	**0.78** (-0.29, 1.66)	**16.90** (14.96, 19.39)	**0.84** (0.80, 0.88)	**16.88** (15.41, 18.89)
Revised regression	9	**2.98** (1.00, 5.38)	**23.62** (20.76, 26.20)	**0.80** (0.75, 0.83)	**18.70** (17.31, 20.48)
RF regression	4	0.34 (-1.71, 1.71)	**20.82** (18.16, 23.36)	0.80 (0.76, 0.84)	**19.08** (17.46, 22.15)
Revised regression	4	2.07 (-1.01, 4.25)	**25.25** (22.50, 28.26)	0.78 (0.74, 0.82)	**20.60** (19.06, 22.50)
Notes: Bias = median difference (95%CI), Precision = IOR of the difference (95%CI), Accuracy30 = 30% accuracy (95%CI), RMSE = Root Mean Squared Error (95%CI), '*' means P < 0.05, and '**' means P < 0.01 comparing with RF with linear regression. RF random forest, RMSE root mean square error.					
Nine variables: Creatinine, cystatin C, age, weight, BMI, UA, BUN, HCT and APOB					
Four variables: Creatinine, cystatin C, age, sex					

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile.docx