Big Data Research 25 (2021) 100234

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

# Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data

Yujie Yang [a,b], Ye Li [a,c], Runge Chen [a], Jing Zheng [d,*], Yunpeng Cai [a,**], Giancarlo Fortino [e]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen, China
[d] Shenzhen Health Information Center, Shenzhen, China
[e] Department of Informatics, Modeling, Electronics, and Systems, University of Calabria, Rende, Italy

ARTICLE INFO

*Article history:*
Received 14 December 2020
Received in revised form 21 February 2021
Accepted 11 April 2021
Available online 24 April 2021

*Keywords:*
Renal failure
Risk prediction
Electronic health record
Health big data
Machine learning

ABSTRACT

Renal failure is a fatal disease raising global concerns. Previous risk models for renal failure mostly rely on the diagnosis of chronic kidney disease, which lacks obvious clinical symptoms and thus is mostly undiagnosed, causing significant omission of high-risk patients. In this paper, we proposed a framework to predict the risk of renal failure directly from a big data repository of chronic disease population without prerequisite diagnosis of chronic kidney disease. The electronic health records of 42,256 patients with hypertension or diabetes in Shenzhen Health Information Big Data Platform were collected, with 398 suffered from renal failure during a 3-year follow-up. Five state-of-the-art machine learning methods are utilized to build risk prediction models of renal failure for chronic disease population. Extensive experimental results show that the proposed framework achieves quite well performance. Particularly, the XGBoost obtains the best performance with an area under receiving-operating-characteristics curve (AUC) of 0.9139. By analyzing the effect of risk factors, we identified that serum creatine, age, urine acid, systolic blood pressure, and blood urea nitrogen are the top five factors associated with renal failure risk. Compared with existing models, our model can be deployed into routine chronic disease management procedures and enable more preemptive, widely-covered screening of renal risks, which would in turn reduce the damage caused by the disease through timely intervention.



## 1. Introduction

Renal failure, also known as end-stage kidney disease (ESKD), is a pathological state of partial or total loss of renal function caused by the development of chronic kidney diseases (CKD) to the later stage. Patients with renal failure would soon suffer from uremia or even deadly consequence, and the treatment can only be dialysis or renal transplantation. The prevalence and total mortality of renal failure continue to increase [1]. In 2016, there were 720,000 patients with renal failure in the United States, and the hospital mortality rate of all dialysis patients was 0.5% [2]. In China, the number of renal failure patients was about 2.9 million and the mortality rate among dialysis patients was 28.42 per thousand years[3]. The difficulty of reversing renal damage increases steadily with the disease progression, thus early detection of high-risk groups for renal failure is particularly important to enable early interventions.

Currently, risk assertion and prevention of renal failure are mainly focused on CKD patients. However, the awareness rate of early CKD is low, which is less than 10% in developing and developed countries, and only 12.5% in China [1,3]. Most patients with CKD have no obvious symptoms in the early stage of onset, resulting in a very high rate of missing diagnosis among general population. A low awareness rate for doctors also exists, and nearly half of the country's attending and deputy doctors have a lower average understanding of CKD guidelines [1]. The high undiagnosed rate of CKD poses a severe challenge to renal failure prevention, as a large portion of high-risk patients were not monitored for disease risk in the early stages.

Several prospective cohort studies and cross-sectional studies have been conducted to develop CKD risk prediction models [4], such as SCORED score [5], ARIC/CHS score [6], Framingham score [7], QKidney score [8], Taiwan score [9], Japan/HIV score [10], and ADVANCE model [11]. The investigated risk factors mostly include

* Corresponding author at: Shenzhen Health Information Center, 2210 North Road, Luohu District, Shenzhen, China.
** Corresponding author at: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Blvd, Nanshan District, Shenzhen, China.
*E-mail addresses:* cnzhengj@126.com (J. Zheng), yp.cai@siat.ac.cn (Y. Cai).

https://doi.org/10.1016/j.bdr.2021.100234

age, gender, body mass index, blood pressure, diabetes status, serum creatinine, proteinuria, serum albumin, and total protein. In addition, some studies added novel biomarkers such as smoking, kidney stones, and family history of kidney disease, or genetic factors [12] to improve model performance. Subsequently, risk models for predicting progression to ESKD have been developed by meta-analysis, the most famous of which is the 4-variable Kidney Failure Risk Equation (KFRE), using gender, age, estimated glomerular filtration rate (eGFR), and urine albumin-to-creatine ratio (ACR) [13]. There are also two ESKD prediction equations based on 6 variables (adding diabetes mellitus and hypertension), or 8 variables (adding serum albumin, bicarbonate, calcium, and phosphate) [14]. Then many researches underwent on external validation of the KFRE in other groups or diverse patients [15–21]. However, these existing studies are mostly restricted to patients already diagnosed with chronic kidney disease. For general population, some studies focus on factor analysis, such as age [22], eGFR [23], obesity [24], gender [25], smoking and drink [26], but few attempts have been made on creating a risk prediction model. The major obstacle lies in the difficulty to acquire laboratory test results from non-CKD population. Without sufficient number of patient samples with complete laboratory test data covering all studied fields, it is infeasible to build a risk model using traditional cohort study approaches.

With the wide application of electronic medical record system, especially the establishment of regional health information platform for data exchange and sharing, large-scale clinical medical data have been accumulated, which provides a strong data support for medical health research [27,28]. Compared with traditional cohort study protocols, big data systems enable easier accumulation of large population dataset with much lower costs, which specifically boost the efficiency of observational studies. On the other hand, machine learning techniques are being used more and more widely for clinical analysis due to its strong potential to use complex mathematics operations to compute large amounts of data. Extracting and analyzing retrospective population data from electronical health record (EHR) big data platforms would largely extend the feasibility of many clinical studies in the scope of data availability, and we will demonstrate this in our renal failure study as well.

In this paper, we strove to extend the feasibility of renal risk prediction from CKD patients to general chronic disease populations. A total of 42,256 registered patients with hypertension or diabetes were selected from Shenzhen Health Information Big Data Platform. After rigorous population screening, only 5,974 patients were retained, of whom 398 had renal failure during a three-year follow-up. Five machine learning algorithms were used to establish the three-year risk models of renal failure, among which the integrated algorithm XGBoost achieved the optimal performance on the test set. Furthermore, we analyzed the univariate effect of renal failure and showed nine continuous variables that were nonlinearly correlated with renal failure risk.

The contribution of our work can be summarized into three scopes. Firstly, for the first time we extended risk modelling for renal failure to non-CKD patients by conducting a large-scale retrospective study, which was achieved by more efficient curation of target data through the aid of big data technologies. Secondly, with sophisticated machine learning methods, we were able to study a relatively large number of features simultaneously. As a result, we discovered some novel biomarkers of renal failure, including uric acid (UA), aspartate aminotransferase (AST), alanine transaminase (ALT), and total bilirubin (TBIL), which were not included in previous models, and identified their nonlinear role in renal function disorder. Thirdly, the proposed model was based on daily monitoring and physical examination data that are easy to acquire for both CKD and non-CKD chronic disease patients. Therefore, it can be deployed into chronic disease management systems to aid physicians to early identify high-risk population for timely intervention.

## 2. Materials and methods

### 2.1. Data resource

The data used in this paper are from Shenzhen Health Information Big Data Platform, which has access to more than 4,000 health institutions including 85 hospitals and more than 650 community health service centers. The platform covered medical service records including outpatient, inpatient, biochemical test, imaging examination, physical examination, and regular follow-up records of registered patients with hypertension, diabetes, cancer, and other diseases. At present, the platform has more than 5 billion medical service records and 598 million electronic medical records, covering a time span from 2010 to 2020. Medical records among different institutions of the same patient can be associated with a unique personal identification number. Due to the case that all medical records were collected in routine clinical activities and the anonymous nature of the obtained data, following the Guidelines of the WMA Declaration of Helsinki term 32, a waive-of-consent protocol was adopted and was approved by the SIAT IRB with No. SIAT-IRB-151115-H0084.

The causes of renal failure are complex, diabetic nephropathy (43.2%) and hypertension (23%) form the main causes of renal failure worldwide [2]. Moreover, a large portion of patients with diabetes and hypertension tend to receive periodic physical examinations, thus a large number of laboratory test result data needed for renal risk prediction have been accumulated, as in the case of the Shenzhen Health Information Big Data Platform. Therefore, this study mainly focused on predicting renal failure risk for these two types of chronic disease patients with high incidence and standardized management.

The main goal of this work is to establish a high-precision three-year short-term risk prediction model for the two major chronic disease population of hypertension and diabetes, based on the real-world population electronic medical record data and machine learning techniques, and thus to explore the risk factors of renal failure and support clinical decision making. The pipeline of the study is depicted in Fig. 1.

### 2.2. Study population

A total of 228,903 registered patients with hypertension or diabetes were selected from the platform, including 188,155 hypertension and 67,737 diabetes patients. The diagnostic of renal failure was extracted from the main diagnosis fields of the outpatient or inpatient records according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) [29] diagnostic codes, which includes N17(acute renal failure), N18(chronic renal failure), N19(unspecified renal failure), I12.0(hypertensive renal disease with renal failure), I13.1(hypertensive heart and renal disease with renal failure), and I13.2(hypertensive heart and renal disease with both (congestive) heart failure and renal failure). As a result, there were 5,649 cases of renal failure onset.

Based on the findings of previous researches and the needs for renal function evaluation, we limited our study to patients with biochemical tests after their registration date, and 186,284 samples were excluded. For positive cases (patients with renal failure), we required patients not to suffer from renal failure in the initial state and excluded 363 samples. To rule out the possible impreciseness of diagnostic time or delayed diagnosis, we required renal failure patients to have serum creatine (CREA) laboratory results at least six months ahead of the renal failure onset, which excluded 1,738

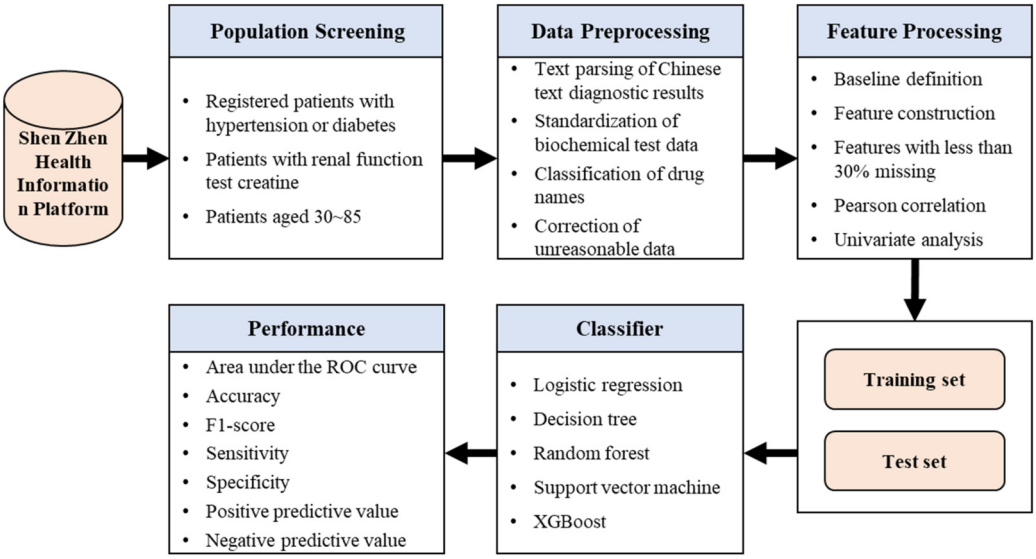


**Fig. 1.** The pipeline of the study. ROC: receiver operating characteristic.

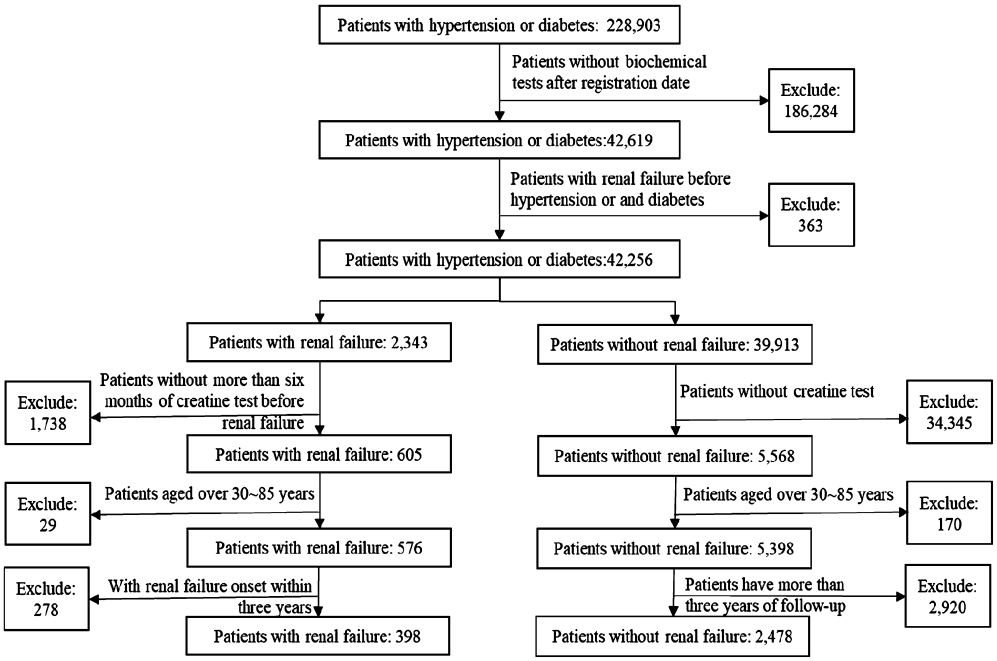


**Fig. 2.** The process of screening study population.

samples. In addition, only patients aged between 30 and 85 years were included in the study. Finally, only 5,974 samples met the above restriction, of which 398 samples had renal failure within 3 years of follow-up, and 2,478 samples without renal failure had more than 3 years of follow-up. The screening process of the study population is shown in Fig. 2.

### 2.3. Data preprocessing

The medical records of 42,256 samples were extracted from the platform, including 9.15 million biomedical test records, 3.37 million outpatient or inpatient records, and 620,000 follow-up records. As the medical records were collected from hundreds of health institutions with slightly different medical service systems, resulting in diverse data formats, poor data quality and even a large number of missing fields, the following steps were taken to clean the data:

Firstly, text parsing was performed. The diagnostic results in outpatient or inpatient records are a mixture of Chinese natural language text and multiple versions of ICD codes, which requires text processing. The unique characteristic of clinical text makes the traditional nature processing method difficult to be applied well. Previous studies have made many attempts and achieved some results [30,31]. In this study, we used the latest ICD-10 codes for text parsing. For ICD encoded text with other versions, conversion was carried out according to the corresponding relationship between versions of ICD codes. For the non-encoded diagnosis results, we used an internally designed lexical parsing code table to convert them into the corresponding ICD-10 codes, and the parsing process was iterated until the unparsed text was considered uninformative. The lexical parsing code table was built by adding a manually constructed matching list to the officially standard ICD-10 code classification table, in which the manually constructed matching list was generated by the following steps: a) extracting all diag-

**Table 1**
Text parsing examples of Chinese text diagnostic results.

| Original diagnostic results | | ICD-10 Codes after text parsing |
|---|---|---|
| DIAG_NAME | DIAG_CODE | DIAG_CODE_NEW |
| 冠状动脉粥样硬化性心脏病 (coronary heart disease) | ICD06999 | I25.103 |
| 冠心病 (coronary heart disease) | 12956 | I25.103 |
| 冠心病 (coronary heart disease) | 3495 | I25.103 |
| 冠心病 (coronary heart disease) | 1001974 | I25.103 |
| 冠状动脉粥样硬化性心脏病 (coronary heart disease) | I25.103 | I25.103 |
| 冠心病，2 型糖尿病 (coronary heart disease, type 2 diabetes) | / | I25.103, E14.900 |
| 房颤，冠状动脉性心脏病 (atrial fibrillation, coronary heart disease) | 192423 | I48.x01, I25.103 |
| 冠心病，高血压病 (coronary heart disease, hypertension) | 50683 | I25.103, I10.X02 |

nostic results with ICD-10 codes, b) removing characters such as spaces, tabs and newlines, c) removing duplicates with the same diagnostic results and codes, d) arranging diagnostic results in descending order of character length, e) repeating substring matching for longer strings and replacing successful matches with ICD-10 codes, until no more substring were contained. The final phrase and its ICD-10 codes were the constructed matching list. Table 1 shows a sample of text parsing.

Secondly, standardization was implemented to the biomedical test results. There are differences in the expression of biomedical test items in different institutions, including the item codes and units. We converted the corresponding test items into consistent code and unit. For example (Fig. 3a)), there are many kinds of item codes of serum creatine such as CRE, Crea, Cr, and CREA, and there are two units of μmol/L and mg/dL. We unified the code and unit into CREA and μmol/L, then carried out the numerical conversion of item results according to formula 1 mg/dL = 88.4umol/.

Thirdly, drug names were classified. The names of drugs in outpatient or inpatient prescriptions were diverse. For example, the commonly used antihypertensive drug irbesartan has a number of commodity names, such as Ambovey, Su shi, Yitaiqing and so on. Considering the characteristics of the study population, we only classified the drug-use into two categories of antihypertensive drugs and hypoglycemic drugs.

Finally, the unreasonable data was corrected. The records with obvious error items were deleted or assigned empty. For example (Fig. 3b)), systolic blood pressure below 40, diastolic blood pressure below 30, or body masa index below 10 were set to empty. In addition, records with more missing items were deleted.

### 2.4. Feature processing

Unlike most existing prospective studies, our study is a retrospective study based on real-world data that contains medical service records at multiple time points for each patient, thus it requires the definition of a baseline. In this study, the baseline was defined as the date of the first renal function test, and the corresponding test items and results were extracted as the baseline features. Here only the test items related to blood lipids, blood glucose, electrolytes, liver function and renal function were extracted. For items not checked at the baseline, the test records within three months before and after the baseline were extracted to fill in. If there were multiple records for the same item, the record closest to the baseline was selected. Second, the physiological parameters in the most recent follow-up record from the baseline were extracted as features, containing blood pressure, heart rate, and body mass index. Then, diseases and symptoms in the outpatient and inpatient records prior to the baseline were extracted, characterized by ICD-10 codes, and then binarized according to the existence. In addition, demographic characteristics, lifestyles (i.e., smoking and drink), and drug categories were extracted, where lifestyles and drug categories were binarized by presence.

Appropriate feature selection can reduce learning difficulty and improve model efficiency. First, feature with missing values above 30% were removed. Then, some feature selection techniques such as Pearson's correlation and univariate analysis were adopted to remove the redundant features. At the same time, some features based on existing research results and clinical experience were manually retained. The measurement scale of all retained features is shown in Table 2.

### 2.5. Predictive modeling

Five machine learning algorithms were implemented to establish the risk models between the extracted features and the occurrence of renal failure by using the Scikit-learn library in a Python programming environment.

XGBoost (Extreme Gradient Boosting):XGBoost is an efficient machine learning integration algorithm based on multiple decision trees under the gradient boosting framework [32]. Different from traditional gradient boosting decision tree methods, XGBoost supports column sampling, which can reduce overfitting and calculation. XGBoost also considers sparse values and supports missing value by default, which are naturally transformed to a sparse matrix containing only no missing value.

Logistic regression (LR): logistic regression is a classical classification model which is almost the most commonly used analytical method in epidemiology and medicine, often used in risk factors discovery, disease risk prediction and automatic disease diagnosis. It is a generalized linear regression analysis model [33], which introduces a sigmoid function to normalize dependent variables. Logistic regression is commonly used as a dichotomous model, and requires dependent variables to follow binomial distribution.

Decision tree (DT): decision tree is a typical classification and regression model for predicting a target in the form of tree struc-

| ITEM_CH_NAME | ITEM_EN_NAME | ITEM_RESULT_NUM | ITEM_RESULT_UNIT |
|---|---|---|---|
| 肌酐 | CRE | 66.3 | mg/dL |
| 肌酐[Cr] | Cr | 77.0 | umol/L |
| 肌酐 | CR | 68 | umol/L |
| 肌酐 | CR | 56 | umol/L |
| 肌酐(CR-S) | CR-S | 103.0 | umol/L |
| 肌酐 | Cr | 83.2 | umol/L |
| 肌酐 | CR | 81 | umol/L |
| 肌酐[CREA] | CREA | 55.00 | umol/L |
| 肌酐 | Cr | 73 | umol/L |
| 肌酐 | Cr | 53.0 | umol/L |
| 肌酐 | CRE | 53.5 | umol/L |
| 肌酐(Cr酶促动力学法) | CREA | 37 | umol/L |
| 肌酐 | Crea. | 42.1 | umol/L |

| FOLLOWUP_DATE | SBP | DBP | GLU | BMI |
|---|---|---|---|---|
| 2011/9/21 | 0 | 0 | | 0.00 |
| 2017/5/15 | 5 | 85 | 5.6 | 22.76 |
| 2017/9/6 | 12 | 79 | 5.0 | 32.89 |
| 2016/12/14 | 14 | 92 | 5.3 | 20.28 |
| 2016/1/17 | 108 | 64 | 4.2 | 2.13 |
| 2015/4/22 | 108 | 64 | 4.2 | 2.13 |
| 2015/2/15 | 110 | 70 | 5.0 | 2.13 |
| 2016/2/22 | 110 | 90 | 0.0 | 32.18 |
| 2015/7/17 | 112 | 78 | 4.6 | 2.13 |
| 2014/8/31 | 114 | | | |
| 2016/6/18 | 118 | 84 | 4.7 | 24.22 |
| 2017/4/11 | 119 | 86 | 6.5 | 23.44 |
| 2013/11/29 | 120 | | | |
| 2016/5/20 | 120 | 80 | 5.1 | 22.43 |
| 2016/9/4 | 120 | 1 | 4.5 | 24.22 |

**Fig. 3.** a) Examples of standardized biochemical tests: creatine; b) examples of exception value.

ture, which consists of nodes and edges [34]. Decision tree learning algorithm is usually a recursive selection of optimal features such that each subset has the best classification, including feature selection, decision tree generation and decision tree pruning process.

Random forest (RF): random forest is a popular ensemble algorithm, which determines the final prediction by combining the outcome of multiple weak classifiers [35]. In random forests, the base classifiers are trained independently, so the learning process is very fast. Moreover, random forests have the advantages of evaluating the importance of variables and resisting overfitting.

Support vector machine (SVM): support vector machine is a generalized linear classifier, which is characterized by the ability to minimize empirical errors and maximize geometric edge regions at the same time [36]. In addition, the stability and sparsity of support vector machine make it have good generalization ability, and the computation is small when using kernel functions.

We divided the data set into training set and test set in a ratio of 6:4. 10-fold cross validation was implemented on the training set and the performance of the models was evaluated on the test set. For missing values, no value imputation operation was performed for XGBoost as the missing value can be directly marked and only the samples without missing values were used for creating trees. For the other four methods, the missing values were filled with mean of each feature, and the data were standardized by the mean and variance of each feature. A series of evaluation criteria were employed to validate the models, including the area under the receiver operating characteristic curve (AUC), accuracy, F1-score, specificity, sensitivity, negative predictive value (NPV), and positive predictive value (PPV).

All experiments were performed under the environment manager Anaconda of Linux server in the isolated intranet, and a python3.6.5 kernel was used for data processing and modeling.

## 3. Results

### 3.1. Study cohort and characteristics description

A total of 5,974 patients were screened in the study cohort. After 3 years of follow-up, 398 patients (positive cases) had renal failure and 2,478 patients without renal failure (negative cases) were still being followed up. The distribution of the follow-up length is depicted in Fig. 4. After feature extraction and selection, there were 52 features pre-selected and the baseline characteristics are shown in Table 2. Among the 398 patients with renal failure, only 316 patients (79.39%) had chronic kidney disease at the baseline.

### 3.2. Model prediction performance

Considering the number of samples and follow-up length, we aimed to develop a 3-year risk prediction model of renal failure

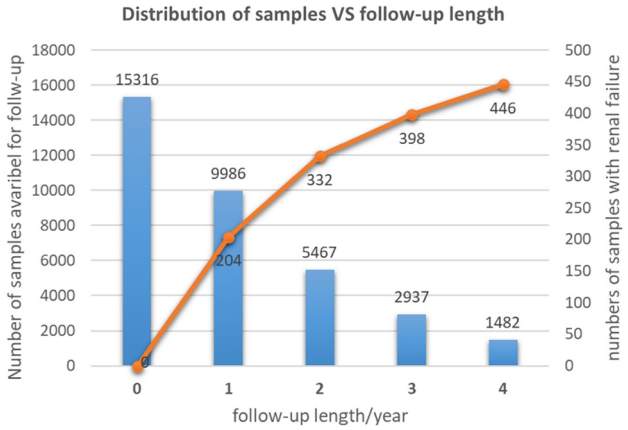


**Fig. 4.** Distribution of follow-up length of the study cohort. The histogram shows the number of samples still available as the follow-up length increasing. The line diagram shows the total number of samples with renal failure at the end of the follow-up period.

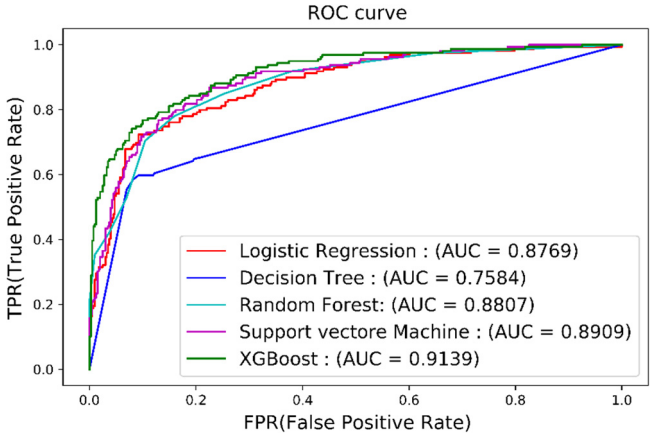


**Fig. 5.** The receiver operating characteristic curve of models.

based on the 2,876 samples, including 398 patients with renal failure and 2,478 controlled patients with more than three years of follow-up. 1,725 samples (60%) were used for model training, and the remaining 1,151 samples (40%) were used to validate the models. Table 3 shows the population profile of the dataset and we can see that there were slight differences in gender, age, and chronic disease types between the positive and negative cases.

The performance of each model is shown in Table 4. The method XGBoost achieved the best performance on the test set with AUC of 0.9139 and accuracy of 0.8643, followed by the three widely used traditional algorithms with comparable performance, which are SVM, RF, and LR. However, the performance of DT was

**Table 2**
Baseline characteristics of the study cohort.

| Characteristics | Positive cases | Negative cases | p-value* |
|---|---|---|---|
| Demographics | | | |
| Gender, male | 244(61.30%) | 1255(50.65%) | <0.001 |
| Age, mean (SD), year | 62.39(13.98) | 54.9(13.28) | <0.001 |
| Life style (current or previous) | | | |
| History of smoking | 19(4.77%) | 191(7.71%) | 0.037 |
| History of drink | 20(5.03%) | 246(9.93%) | 0.002 |
| Physical examination, mean (SD) | | | |
| Systolic blood pressure, mmHg | 132.68(13.42) | 130.79(10.11) | <0.001 |
| Diastolic blood pressure, mmHg | 81.21(10.95) | 79.69(8.18) | 0.063 |
| Pulse pressure, mmHg | 51.93(8.91) | 50.72(8.83) | 0.118 |
| Hear rate, times/minute | 76.45(6.98) | 74.07(5.88) | <0.001 |
| Body mass index, kg/mˆ2 | 23.45(2.87) | 23.11(2.54) | 0.143 |
| Medical history | | | |
| Hypertension | 391(98.24%) | 2430(98.06%) | 0.809 |
| Diabetes | 240(60.30%) | 1384(55.85%) | 0.096 |
| Chronic kidney disease | 316(79.39%) | 1196(48.26%) | <0.001 |
| Diabetic nephropathy | 113(28.39%) | 162(6.53%) | <0.001 |
| Hypertensive nephropathy | 42(10.55%) | 29(1.17%) | <0.001 |
| Glomerular nephritis | 10(2.51%) | 20(0.81%) | 0.002 |
| Chronic tubulointerstitial nephritis | 3(0.75%) | 5(0.20%) | 0.052 |
| Obstructive nephropathy | 148(37.19%) | 980(39.55%) | 0.371 |
| Coronary heart disease | 177(44.47%) | 1416(57.14%) | <0.001 |
| Stroke | 143(35.93%) | 960(38.74%) | 0.284 |
| Heart failure | 157(39.45%) | 691(27.89%) | <0.001 |
| Atrial fibrillation | 40(10.05%) | 152(6.134%) | 0.004 |
| Cardiovascular disease | 248(62.31%) | 1751(70.66%) | <0.01 |
| Albuminuria | 7(1.76%) | 8(0.32%) | <0.001 |
| Asthma | 11(2.76%) | 53(2.14%) | 0.438 |
| Dyssomnia | 26(6.53%) | 66(2.66%) | <0.001 |
| Palpitation | 7(1.76%) | 14(0.56%) | 0.009 |
| Chough | 14(3.52%) | 41(1.65%) | 0.012 |
| Chest pain | 10(2.51%) | 38(1.53%) | 0.157 |
| Malaise and fatigue | 8(2.01%) | 22(0.89%) | 0.041 |
| Nausea and vomiting | 4(1.01%) | 9(0.36%) | 0.076 |
| Abnormal respiration | 13(3.26%) | 78(3.14%) | 0.901 |
| Edema | 4(1.01%) | 16(0.65%) | 0.423 |
| Laboratory variables, mean (SD) | | | |
| Total cholesterol, mmol/L | 5.01(1.43) | 5.05(1.32) | 0.774 |
| Triglyceride, mmol/L | 2.86(8.37) | 1.84(2.16) | <0.001 |
| LDL-C, mmol/L | 2.91(1.14) | 2.94(1.04) | 0.708 |
| HDL-C, mmol/L | 1.19(0.36) | 1.23(1.32) | 0.269 |
| Serum Na, mg/dL | 138.87(5.07) | 140.11(5.54) | 0.002 |
| Serum K, mg/dL | 4.32(1.08) | 3.97(0.47) | <0.001 |
| Serum Ca, mg/dL | 2.21(0.18) | 2.34(2.73) | 0.542 |
| Blood glucose, mg/dL | 6.09(1.74) | 5.95(1.48) | 0.345 |
| Uric acid, mg/dL | 453.43(140.24) | 357.81(99.89) | <0.001 |
| Serum creatine, mg/dL | 272.55(676.35) | 123.34(842.48) | <0.001 |
| Blood urea nitrogen, mg/dL | 10.21(11.95) | 5.14(1.87) | <0.001 |
| Total bilirubin | 12.11(11.45) | 13.84(7.57) | 0.005 |
| Direct bilirubin | 3.36(6.86) | 3.31(3.09) | 0.854 |
| Indirect bilirubin | 8.46(4.36) | 10.23(5.29) | <0.001 |
| Total protein | 68.56(7.15) | 71.21(6.61) | <0.001 |
| Serum albumin | 39.42(5.17) | 42.7(4.62) | <0.001 |
| Alanine transaminase | 24.36(31.16) | 28.5(58.61) | 0.334 |
| Aspartate aminotransferase | 27.39(32.47) | 26.45(34.47) | 0.702 |
| Medications | | | |
| Antihypertensive drug | 75(18.84%) | 923(37.24) | <0.001 |
| Hypoglycemic drug | 43(10.8%) | 619(24.76%) | <0.001 |
| Outcome | | | |
| Renal failure events | 398 | - | |

HDL-C = high-density lipoprotein cholesterol; LDL-C = low-density lipoprotein cholesterol.

**Table 3**
The population profile of the training set and test set.

| Features | All | | Training set | | Test set | |
|---|---|---|---|---|---|---|
| | Positive | Negative | Positive | Negative | Positive | Negative |
| *Number* | 398(13.83%) | 2478(86.16%) | 239(13.85%) | 1486(86.14%) | 159(13.81%) | 992(86.18%) |
| Gender | 244(61.30%) | 1255(50.64%) | 151(63.17%) | 744(50.06%) | 93(58.49%) | 511(51.51%) |
| Hypertension | 391(98.24%) | 2430(98.06%) | 236(98.74%) | 1456(97.98%) | 155(97.48%) | 974(98.18%) |
| Diabetes | 240(61.38%) | 1384(55.85%) | 141(59.74%) | 829(55.78%) | 99(63.87%) | 555(55.94%) |
| Age | 61.59(14.32) | 60.17(12.48) | 61.53(14.41) | 60.49(12.46) | 61.68(14.23) | 59.69(12.49) |

**Table 4**
Performance of the five models on the test set.

| Method | AUC | Accuracy | F1-score | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|---|
| LR | 0.8769 | 0.8393 | 0.6715 | 0.5788 | 0.9426 | 0.8010 | 0.8494 |
| DT | 0.7584 | 0.8000 | 0.6291 | 0.5975 | 0.8803 | 0.6643 | 0.8465 |
| RF | 0.8807 | 0.8161 | 0.6199 | 0.5283 | 0.9302 | 0.7500 | 0.8326 |
| SVM | 0.8909 | 0.8393 | 0.6739 | 0.5849 | 0.9401 | 0.7949 | 0.8510 |
| XGBoost | 0.9139 | 0.8643 | 0.7467 | 0.7044 | 0.9277 | 0.7943 | 0.8878 |

LR = logistic regression; DT=decision tree; RF = random forest; SVM = support vector machine; AUC = Area under the receiver operating curve; PPV = positive predictive value; NPV = negative predictive value.

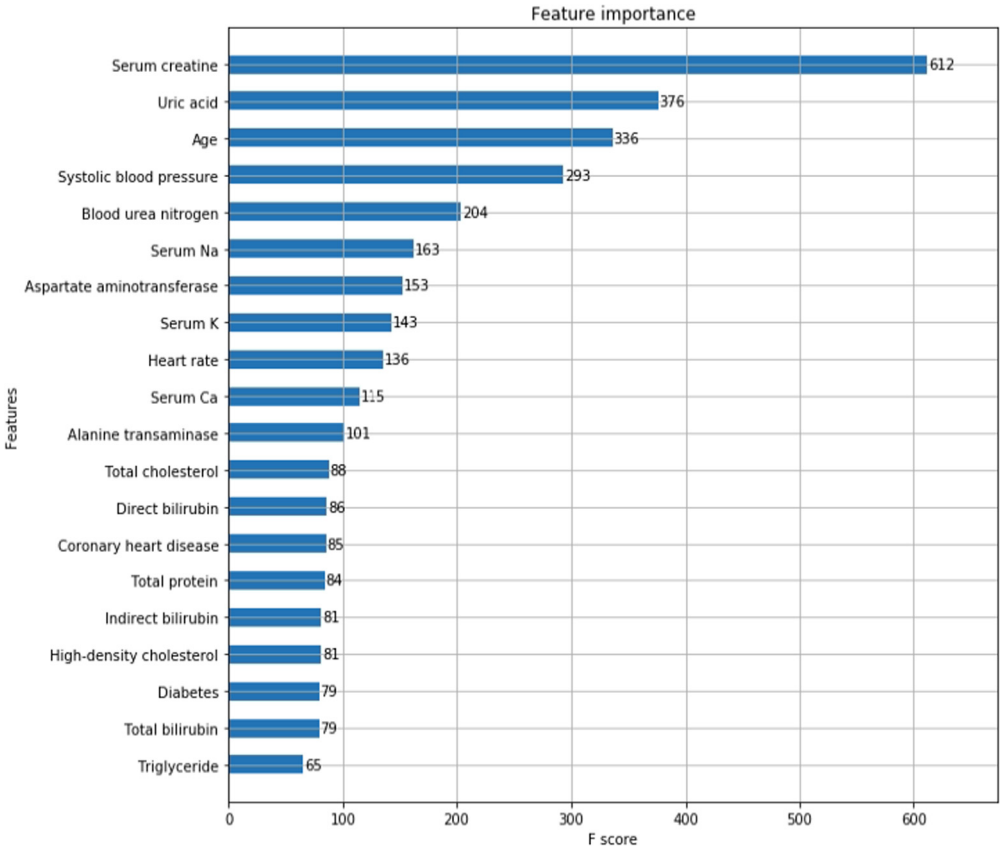


**Fig. 6.** The top 20 features selected by the XGBoost algorithm.

much inferior to the other four methods, as can be seen from the receiver operating characteristic (ROC) curve in Fig. 5.

### 3.3. Contributions of features to model prediction

The feature importance of the XGBoost method measures the relative contribution of the features in the process of building integration tree. The top-ranked 20 features selected by the XGBoost method are shown in Fig. 6, and the five features of CREA, UA, age, systolic blood pressure (SBP), and blood urea nitrogen (BUN) play important roles in risk prediction of renal failure. Our study also demonstrates that pulse pressure, heart rate, serum albumin (ALB), lower-density lipoprotein cholesterol (LDL-C), and AST are significantly associated with renal failure in chronic disease patients of hypertension and diabetes. Specifically, the roles of UA, AST, ALT or TBIL in renal failure were rarely studied and thus were not included in previous models.

### 3.4. Non-linear effect of risk factors

To further analyze the effect of risk factors, univariate trend analysis was implemented to describe the association between continuous variables and the morbidity of renal failure based on the 3-year risk prediction dataset. In this study, curve fitting was used to present the correlation and the morbidity was presented by the number of renal failure cases in a thousand samples. Gaussian function, polynomial function and exponential function were tested separately and the fitting effect was evaluated by discriminant coefficient $R^2$ [37]. The coefficient normally ranges from 0 to 1, and the closer it is to 1, the better the fitting effect. Fig. 7 shows the marginal effect of nine consecutive variables, in which CREA, UA, and age are the top 3 features in the feature importance list (Fig. 6) and their discriminant coefficient are 0.80, 0.68, and 0.86 respectively. We see that the marginal effects of some factors (e.g., age, CREA, BUN, and serum K) form a hinge-like sharp, where the marginal risk remains low when the factor falls within a range and increases steadily after it goes beyond a threshold. On the other hand, some factors (e.g., UA, AST, ALT) exhibit a U-shaped trend, where the marginal risk is minimized when the factor falls within a given range while increases both when it goes lower or higher. Unsurprisingly, the turn-points for most risk factors such as CREA and BUN are highly consistent with the recommended normal range for general population in routine check-ups. A lower level of serum Na is observed to correlated with elevated renal

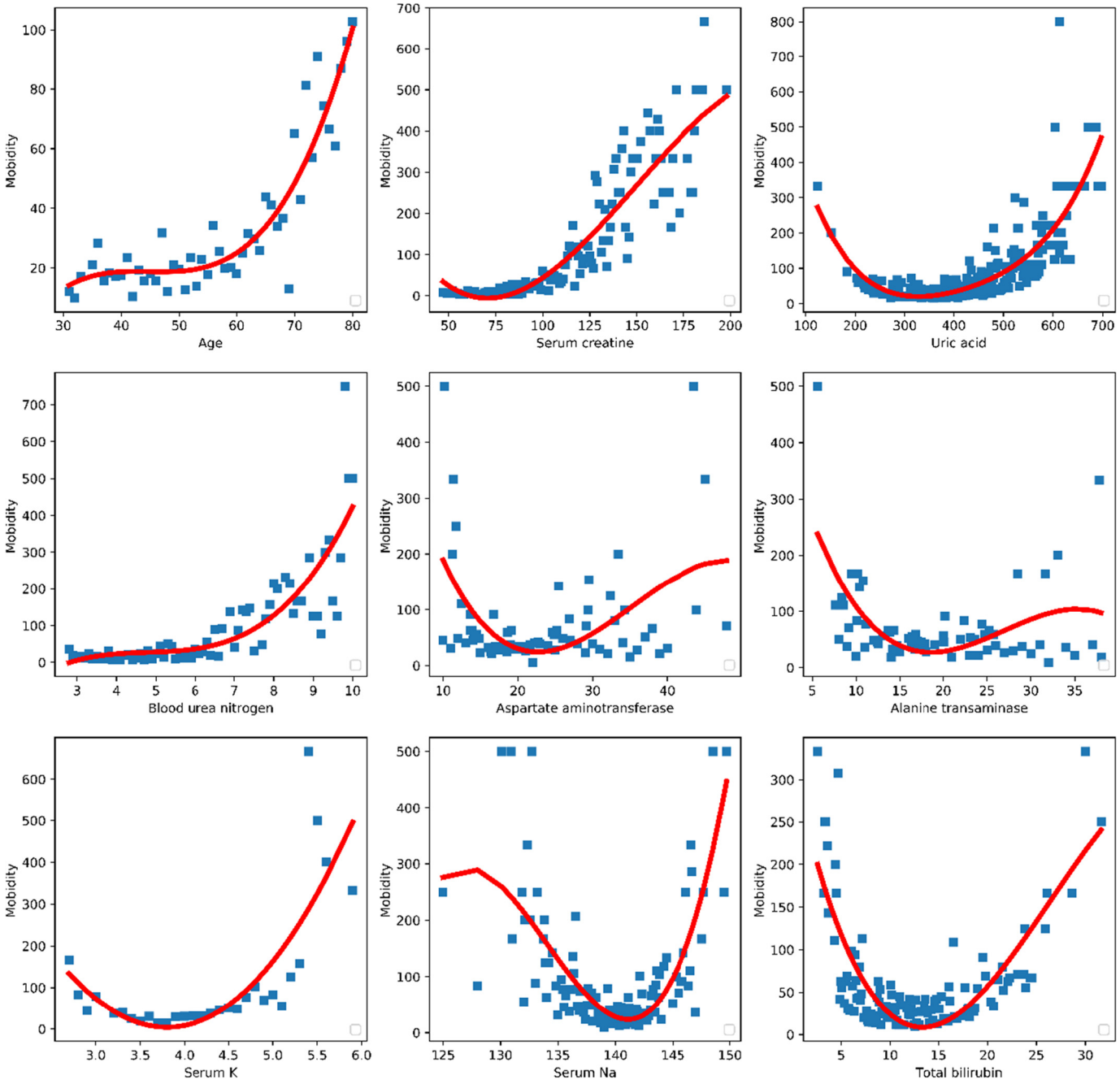


**Fig. 7.** Impact of given continuous variables on the morbidity of renal failure. Morbidity is expressed as the number of cases in a thousand samples.

failure risk, which is mainly due to its negative correlation with serum K, consistent with existing clinical findings. Interestingly, the marginal effect graphs show that lower levels of UA, AST, ALT or TBIL also contribute to higher risk of renal failure, which are rarely discussed in current literatures. Existing clinical studies have identified that lower UA or ALT are early signs of renal dysfunction [38], but few connect them with renal failure risk. However, the clinical implications for lower levels of the other two factors remain currently unclear. This would be worthy of further validations through a larger cohort study in the future.

## 4. Discussion and conclusion

We have developed a high-precision risk prediction model of renal failure for chronic disease patients with hypertension or diabetes based on electronic heath records from the Shenzhen Health Information Big Data Platform. Unlike existing studies, our model does not require patients to be diagnosed with CKD, which avoid the severe defect of low coverage for previous models led by the high undiagnosed rate of CKD patients in clinical practice. Collecting blood samples from large-scale non-CKD population and performing long-term follow-up have been difficult and costly. However, in our work, we manage to curate the data with the aid of big data technologies through extracting useful information from routine clinical records in the large-scale regional medical information platform, making it feasible to perform massive observational cohort studies more efficient.

Our findings partially overlap with some other early studies on patients with CKD. For example, ARIC/CHS score and Framingham score include age, gender, hypertension, diabetes, BMI, and HDL-C. Taiwan score and ADVANCE model include ACR, UA, glucose, and proteinuria. Also, the prediction model of CKD progression KFER

includes CREA, ALB, and history of CKD, stroke, heart failure, and arrhythmia. More importantly, we further identified several new prediction biomarkers such as AST, ALT, and TBIL with the power of sophisticated machine learning methods, and discovered their non-linear role in renal dysfunction. The effect of nonlinear correlation justifies the necessity of adopting sophisticated nonlinear machine learning models over traditional linear regressions. Furthermore, with non-linear ensemble algorithms such as XGBoost used in our work, there is no need to select variables in advance even when the number of potential variables is large, which is different from most traditional clinical studies and enables identification of novel biomarkers with both linear and non-linear effects during modeling process through mining large-scale population data. This is another advantage brought by big data technologies.

Our analysis has a few limitations. It is a retrospective study with data collected years ago, but this study indicates potential application of predicting risk of renal failure for chronic disease patients. In addition, the study cohort was imbalanced in view of the numbers of positive cases and negative cases, we performed randomly stratified sampling according to gender ratio and age stratification, in which the age stratification was 30–45, 45–60, 60–70, and 70–85. According to the proportion of positive cases in the four age ranges, the negative cases were randomly sampled in each age range. However, the cases randomly selected may not represent the rest of the patients accurately. We are currently collecting more patients with diverse basic diseases and trying to further validate and improve the model with recent data. External validation in multiple diverse disease cohorts and evaluation in clinical trials are also needed.

In conclusion, we have developed and validated a highly accurate risk model for predicting renal failure of chronic disease patients with hypertension or diabetes, without necessarily early diagnosis of kidney diseases, which advance the state-of-the-arts for renal failure prediction. The model uses routinely available physical and laboratory examination data and could predict the short-term risk of renal failure with high accuracy. Due to the ease of access to data, it could be easily implemented in laboratory information systems or EHR systems to help with a more pervasive, preemptive screening of renal failure risk, enabling higher efficiency of early disease prevention and intervention. Our works also justify the advantages of adopting big data technologies in public health as well.

## List of Abbreviations

ESKD: End-stage kidney disease
CKD: Chronic kidney disease
KFRE: Kidney Failure Risk Equation
eGFR: Estimated glomerular filtration rate
ACR: Urine albumin-to-creatine ratio
EHR: Electronical health record
UA: Uric acid
AST: Aspartate aminotransferase
ALT: Alanine transaminase
TBIL: Total bilirubin
CREA: Serum creatine
LR: Logistic regression
RF: Random forest
DT: Decision tree
SVM: Support vector machine
AUC: Area under the receiver operating characteristic curve
NPV: Negative predictive value
PPV: Positive predictive value
HDL-C: high-density lipoprotein cholesterol
LDL-C: Low-density lipoprotein cholesterol
ROC: Receiver operating characteristic
SBP: Systolic blood pressure
BUN: Blood urea nitrogen
ALB: Serum albumin

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences [grant number XDB38040200]; Shenzhen Science and Technology Program (grant number: KQTD20190929172835666); and Shenzhen Science and Technology Research Funding [grant numbers JCYJ20180703145202065, JCYJ20180703145002040].

## References

[1] GBD Chronic Kidney Disease Collaboration, Global, regional, and national burden of chronic kidney disease, 1990–2017: s systematic analysis for the Global Burden of Disease Study 2017, Lancet 10225 (395) (2020 February 29) 709–733, https://doi.org/10.1016/S0140-6736(20)30045-3, Epub 2020 Feb 13.
[2] Somnath Pal, Primary causes of end-stage renal disease, US Pharm. 41 (8) (2016) 6.
[3] L. Zhang, M.H. Zhao, L. Zuo, Y. Wang, F. Yu, H. Zhang, H. Wang, CK-NET Work Group, China Kidney Disease network (CK-NET) 2015 annual data report, Kidney Inter., Suppl. 9 (1) (2019 Mar) e1–e81, https://doi.org/10.1016/j.kisu.2018.11.001.
[4] J.B. Echouffo-Tcheugui, A.P. Kengne, Risk models to predict chronic kidney disease and its progression: a systematic review, PLoS Med. 9 (11) (2012) e1001344, https://doi.org/10.1371/journal.pmed.1001344.
[5] H. Bang, S. Vupputuri, D.A. Shoham, P.J. Klemmer, R.J. Falk, M. Mazumdar, D. Gipson, R.E. Colindres, A.V. Kshirsagar, SCreening for Occult REnal Disease (SCORED): a simple prediction model for chronic kidney disease, Arch. Intern. Med. 167 (4) (2007 Feb 26) 374–381, https://doi.org/10.1001/archinte.167.4.374.
[6] A.V. Kshirsagar, H. Bang, A.S. Bomback, S. Vupputuri, D.A. Shoham, L.M. Kern, P.J. Klemmer, M. Mazumdar, P.A. August, A simple algorithm to predict incident kidney disease, Arch. Intern. Med. 168 (22) (2008 Dec 8) 2466–2473, https://doi.org/10.1001/archinte.168.22.2466.
[7] C.S. Fox, P. Gona, M.G. Larson, J. Selhub, G. Tofler, S.J. Hwang, J.B. Meigs, D. Levy, T.J. Wang, P.F. Jacques, E.J. Benjamin, R.S. Vasan, A multi-marker approach to predict incident CKD and microalbuminuria, J. Am. Soc. Nephrol. 21 (12) (2010 Dec) 2143–2149, https://doi.org/10.1681/ASN.2010010085.
[8] J. Hippisley-Cox, C. Coupland, Predicting the risk of chronic kidney disease in men and women in England and Wales: prospective derivation and external validation of the QKidney scores, BMC Fam. Pract. 21 (11) (2010 Jun 21) 49, https://doi.org/10.1186/1471-2296-11-49.
[9] K.L. Chien, H.J. Lin, B.C. Lee, H.C. Hsu, Y.T. Lee, M.F. Chen, A prediction model for the risk of incident chronic kidney disease, Am. J. Med. 123 (9) (2010 Sep) 836–846.e2, https://doi.org/10.1016/j.amjmed.2010.05.010.
[10] M. Ando, N. Yanagisawa, A. Ajisawa, K. Tsuchiya, K. Nitta, A simple model for predicting incidence of chronic kidney disease in HIV-infected patients, Clin. Exp. Nephrol. 15 (2) (2011 Apr) 242–247, https://doi.org/10.1007/s10157-010-0393-x.
[11] M.J. Jardine, J. Hata, M. Woodward, V. Perkovic, T. Ninomiya, H. Arima, S. Zoungas, A. Cass, A. Patel, M. Marre, G. Mancia, C.E. Mogensen, N. Poulter, J. Chalmers, ADVANCE Collaborative Group, Prediction of kidney-related outcomes in patients with type 2 diabetes, Am. J. Kidney Dis. 60 (5) (2012 Nov) 770–778, https://doi.org/10.1053/j.ajkd.2012.04.025.
[12] C.M. O'Seaghdha, Q. Yang, H. Wu, S.J. Hwang, C.S. Fox, Performance of a genetic risk score for CKD stage 3 in the general population, Am. J. Kidney Dis. 59 (1) (2012 Jan) 19–24, https://doi.org/10.1053/j.ajkd.2011.08.030.
[13] N. Tangri, L.A. Stevens, J. Griffith, H. Tighiouart, O. Djurdjev, D. Naimark, A. Levin, A.S. Levey, A predictive model for progression of chronic kidney disease to kidney failure, JAMA 305 (15) (2011 Apr 20) 1553–1559, https://doi.org/10.1001/jama.2011.451.
[14] N. Tangri, M.E. Grams, A.S. Levey, J. Coresh, L.J. Appel, B.C. Astor, G. Chodick, A.J. Collins, O. Djurdjev, C.R. Elley, M. Evans, A.X. Garg, S.I. Hallan, L.A. Inker, S. Ito, S.H. Jee, C.P. Kovesdy, F. Kronenberg, H.J. Heerspink, A. Marks, G.N. Nadkarni, S.D. Navaneethan, R.G. Nelson, S. Titze, M.J. Sarnak, B. Stengel, M. Woodward, K. Iseki, CKD Prognosis Consortium, Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis, JAMA 315 (2) (2016 Jan 12) 164–174, https://doi.org/10.1001/jama.2015.18202.

[15] E. Winnicki, C.E. Mcculloch, M.M. Mitsnefes, S.L. Furth, B.A. Warady, E. Ku, Use of the kidney failure risk equation to determine the risk of progression to end-stage renal disease in children with chronic kidney disease, JAMA Pediatr. 172 (2) (2018 Feb 1) 174–180, https://doi.org/10.1001/jamapediatrics.2017.4083.

[16] M.J. Peeters, A.D. van Zuilen, J.A. van den Brand, M.L. Bots, P.J. Blankestijn, J.F. Wetzels, MASTERPLAN Study Group, Validation of the kidney failure risk equation in European CKD patients, Nephrol. Dial. Transplant. 28 (7) (2013 Jul) 1773–1779, https://doi.org/10.1093/ndt/gft063.

[17] S. Low, S.C. Lim, X. Zhang, S. Zhou, L.Y. Yeoh, Y.L. Liu, S. Tavintharan, C.F. Sum, Development and validation of a predictive model for Chronic Kidney Disease progression in type 2 diabetes mellitus based on a 13-year study in Singapore, Diabetes Res. Clin. Pract. 123 (2017 Jan) 49–54, https://doi.org/10.1016/j.diabres.2016.11.008.

[18] M.E. Grams, L. Li, T.H. Greene, A. Tin, Y. Sang, W.H. Kao, M.S. Lipkowitz, J.T. Wright, A.R. Chang, B.C. Astor, L.J. Appel, Estimating time to ESRD using kidney failure risk equations: results from the African American Study of Kidney Disease and Hypertension (AASK), Am. J. Kidney Dis. 65 (3) (2015 Mar) 394–402, https://doi.org/10.1053/j.ajkd.2014.07.026.

[19] M. Yamanouchi, J. Hoshino, Y. Ubara, K. Takaichi, K. Kinowaki, T. Fujii, K. Ohashi, K. Mise, T. Toyama, A. Hara, K. Kitagawa, M. Shimizu, K. Furuichi, T. Wada, Value of adding the renal pathological score to the kidney failure risk equation in advanced diabetic nephropathy, PLoS ONE 13 (1) (2018 Jan 16) e0190930, https://doi.org/10.1371/journal.pone.0190930.

[20] C.C. Lim, M.L. Chee, C.Y. Cheng, J.L. Kwek, M. Foo, T.Y. Wong, C. Sabanayagam, Simplified end stage renal failure risk prediction model for the low-risk general population with chronic kidney disease, PLoS ONE 14 (2) (2019 Feb 22) e0212590, https://doi.org/10.1371/journal.pone.0212590.

[21] S.L. Tummalapalli, M.M. Estrella, Predicting risk of kidney disease: is risk-based kidney care on the horizon?, JAMA 322 (21) (2019 Nov 8) 2079–2081, https://doi.org/10.1001/jama.2019.17378.

[22] M.S. Hommos, R.J. Glassock, A.D. Rule, Structural and functional changes in human kidneys with healthy aging, J. Am. Soc. Nephrol. 28 (10) (2017 Oct) 2838–2844, https://doi.org/10.1681/ASN.2017040421.

[23] A.R. Chang, M.E. Grams, S.H. Ballew, H. Bilo, A. Correa, M. Evans, O.M. Gutierrez, F. Hosseinpanah, K. Iseki, T. Kenealy, B. Klein, F. Kronenberg, B.J. Lee, Y. Li, K. Miura, S.D. Navaneethan, P.J. Roderick, J.M. Valdivielso, F.L.J. Visseren, L. Zhang, R.T. Gansevoort, S.I. Hallan, A.S. Levey, K. Matsushita, V. Shalev, M. Woodward, CKD Prognosis Consortium (CKD-PC), Adiposity and risk of decline in glomerular filtration rate: meta-analysis of individual participant data in a global consortium, BMJ 364 (2019 Jan 10) k5301, https://doi.org/10.1136/bmj.k5301.

[24] A. Whaley-Connell, J.R. Sowers, Obesity and kidney disease: from population to basic science and the search for new therapeutic targets, Kidney Int. 92 (2) (2017 Aug) 313–323, https://doi.org/10.1016/j.kint.2016.12.034.

[25] S.R. Silbiger, J. Neugarten, The impact of gender on the progression of chronic renal disease, Am. J. Kidney Dis. 25 (4) (1995 Apr) 515–533, https://doi.org/10.1016/0272-6386(95)90119-1.

[26] A. Shankar, R. Klein, B.E. Klein, The association among smoking, heavy drinking, and chronic kidney disease, Am. J. Epidemiol. 164 (3) (2006 Aug 1) 263–271, https://doi.org/10.1093/aje/kwj173.

[27] C. Safran, M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, D.E. Detmer, Expert panel. Toward a national framework for the secondary use of health data: an American medical informatics association white paper, J. Am. Med. Inform. Assoc. 14 (1) (2007 Jan-Feb) 1–9, https://doi.org/10.1197/jamia.M2273.

[28] J. Roski, G.W. Bo-Linn, T.A. Andrews, Creating value in health care through big data: opportunities and policy implications, Health Aff. (Millwood) 33 (7) (2014 July) 1115–1122, https://doi.org/10.1377/hlthaff.2014.0147.

[29] World Health Organization, ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision, 2nd ed., World Health Organization, 2004, https://apps.who.int/iris/handle/10665/42980.

[30] Azita Yazdani, Reza Safdari, Ali Golkar, Sharareh R. Niakan Kalhori, Words prediction based on N-gram model for free-text entry in electronic health records, Health Inf. Sci. Syst. 7 (2019) 6, https://doi.org/10.1007/s13755-019-0065-5.

[31] P.H. Wu, A. Yu, C.W. Tsai, J.L. Koh, C.C. Kuo, A.L.P. Chen, Keyword extraction and structuralization of medical reports, Health Inf. Sci. Syst. 8 (2020) 18, https://doi.org/10.1007/s13755-020-00108-6.

[32] T. Chen, Guestrin C. Xgboost, A scalable tree boosting system, in: KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2016 Aug, pp. 785–794.

[33] J.D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, 3rd edition, John Wiley & Sons, Canada, 2013 Apr.

[34] R.C. Barros, M.P. Basgalupp, A.C.P.L.F. de Carvalho, A.A. Freitas, A survey of evolutionary algorithms for decision-tree induction, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 42 (3) (2012 May) 291–312, https://doi.org/10.1109/TSMC.2011.2157494.

[35] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[36] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995 Sep) 273–297, https://doi.org/10.1007/BF00994018.

[37] A. Colin Cameron, Frank A.G. Windmeijer, An R-squared measure of goodness of fit for some common nonlinear regression models, J. Econom. 77 (2) (1997) 1790–1792, https://doi.org/10.1016/S0304-4076(96)01818-0.

[38] E. Elinav, Z. Ackerman, Y. Maaravi, I.Z. Ben-Dov, E. Ein-Mor, J. Stessman, Low alanine aminotransferase activity in older people is associated with greater long-term mortality, J. Am. Geriatr. Soc. 54 (11) (2006 Nov) 1719–1724, https://doi.org/10.1111/j.1532-5415.2006.00921.