Clinical and Experimental Nephrology (2020) 24:865–875
https://doi.org/10.1007/s10157-020-01909-5

ORIGINAL ARTICLE

Development of prognostic model for patients at CKD stage 3a and 3b in South Central China using computational intelligence

Qiongjing Yuan[1] · Haixia Zhang[1,2] · Yanyun Xie[1] · Wei Lin[3] · Liangang Peng[4] · Liming Wang[5] · Weihong Huang[6] · Song Feng[7] · Xiangcheng Xiao[1]

Received: 3 December 2019 / Accepted: 10 March 2020 / Published online: 1 August 2020

Abstract

Background Chronic kidney disease (CKD) stage 3 was divided into two subgroups by eGFR (45 mL/ min 1.73 m^2). There is difference in prevalence of CKD, racial differences, economic development, genetic, and environmental backgrounds between China and Western countries.

Methods We used a computational intelligence model (CKD stage 3 Modeling, CSM) to distinguish CKD stage 3 with CKD stage 3a/3b by data distribution rules, pearson correlation coefficient (PCC), spearman correlation (SCC) analysis, logistic regression (LR), random forest (RF), support vector machine (SVM), and neural network (Nnet) to develop Prognostic Model for patients with CKD stage 3a/3b in South Central China. Furthermore, we used RF to discover risk factors of progression of CKD stage 3a and 3b to CKD stage 5. 1090 cases of CKD stage 3 patients in Xiangya Hospital were collected. Among them, 455 patients progressed to CKD stage 5 in a median follow-up of 4 years (IQR 4.295, 4.489).

Results We found that the common risk factors for progression of CKD stage 3a/3b to CKD stage 5 included albumin, creatinine, total protein, etc. Proteinuria, direct bilirubin, hemoglobin, etc. accounted for the progression from stage CKD stage 3a to stage 5. The risk factors for CKD stage 3b progression to stage 5 included low-density lipoprotein cholesterol, diabetes, eosinophil percentage, etc.

Conclusions CSM could be used as a point-of-care test to screen patients at high risk for disease progression, might allowing individualized therapeutic management.

Keywords CKD stage 3 modeling · Chronic kidney disease · Computational intelligence · End-stage renal disease

Introduction

Chronic kidney disease (CKD), a major public health problem with an increasing incidence and prevalence year by year, affects 700 million people globally [1]. A recent study

Qiongjing Yuan and Haixia Zhang contribute equal as first authors.

✉ Song Feng
fs205@sina.com

✉ Xiangcheng Xiao
xiaoxc@csu.edu.cn

1 Department of Nephrology, Xiangya Hospital, Central South University, 87 Xiangya Road, Changsha 410008, Hunan, China

2 Department of Nephrology, Second Affiliated Hospital of Soochow University, 1055 Sanxiang Road, Suzhou 215000, Jiangsu, China

3 Department of Pathology, Xiangya Hospital, Central South University, 87 Xiangya Road, Changsha 410008, Hunan, China

4 Changsha Aeronautical Vocational and Technical College, Changsha 410014, Hunan, China

5 Bitvalue Technology (Hunan) Company Limited, Xiangjiang Road, Changsha 410082, China

6 Mobile Health Ministry of Education-China Mobile Joint Laboratory, Xiangya Hospital, Central South University, Changsha 410008, China

7 Network Information Center, Xiangya Hospital, Central South University, Xiangya Road, Changsha 410008, China

projecting the future burden of CKD in the United States estimated that the prevalence of CKD (here defined as CKD stages 1–4) among those aged 30 years or older would increase from 13.2% in 2010 to 16.7% in 2030 [2]. In 2012, a survey including 47,204 adults from 13 provinces and cities in China showed that the total prevalence of CKD was 10.8% [3]. It is estimated that the number of dialysis patients in China will increase at a rate of 20–30% per year. Namely, over 400, 000 Chinese patients will develop end-stage renal disease (ESRD) every year, which comprises a large part of the world's ESRD population. This development will bring a heavy burden to public health and society, leading to a considerable challenge to medical and health undertakings. The cost of dialysis treatment alone for one patient would be approximately $14,300 per year, whereas the per capita disposable income is $1210 in urban areas and $375 in rural areas in China [4]. Early recognition and prevention of potential ESRD is therefore of significant importance.

In 2012, the Kidney Disease: improving Global Outcomes (KDIGO) guidelines recommended reclassifying CKD [5]. The classification divided CKD stage 3 into two subgroups by applying a cutoff point of the estimated glomerular filtration rate (eGFR) (45 mL/ min 01.73 m^2). Hence, subjects with CKD stage 3a were considered low risk compared with patients with CKD stage 3b. This new classification was based on a meta-analysis performed in 45 cohorts involving over 1.5 million participants mainly from developed countries [5]. The incidence, prevalence, and progression of CKD vary within countries by ethnicity and social determinants of health, possibly through epigenetic influence [6]. The prevalence of CKD and the prevalence of diabetic CKD have both stabilized in the United States since the early 2000 s, signaling a change in the epidemiology of CKD [7]. There are several populations around the world with an emerging risk of increasing CKD, including China (in the context of rapid urbanization and a rising incidence of diabetes) [7]. Therefore, we cannot simply follow the guidelines from the data from developed countries. In addition, only a few studies have indicated that KDIGO staging is applicable to patients with CKD stage 3 in China [8] and have not observed differences in the prognosis between patients with stage 3a and 3b CKD. It is important to know whether the division of CKD stage 3 is suitable for Chinese patients.

Electronic medical records provide large-scale real-world clinical data for the use in developing clinical decision systems. However, sophisticated methodology and analytical skills are required to handle the large-scale datasets necessary for the optimization of prediction accuracy [9]. With government incentives offered to clinical organizations to transition from paper-based patient information to well-structured and managed digital form, there has been a tremendous explosion in the availability of patient-centric healthcare data. Such data can be leveraged to open new avenues in advancing healthcare by improving patient care and creating new efficiencies in delivering care [10]. Besides, early prediction of deterioration can play an important role in supporting health care professionals, as an estimated 11 percent of hospital deaths follow a failure to promptly recognize and treat deteriorating patients [11]. Machine learning algorithms are well suited to analyze large, complex dataset [12], which can identify information quickly, effectively and explore intrinsic relationship. To verify the staging for Chinese patients with CKD stage 3 and built up an alerting model, we built our CKD3 staging modeling (CSM) approach and evaluated its reliability in a retrospective study involving CKD patients treated at Xiangya Hospital, one of the largest hospitals in South Central China. The CSM approach computes the cutoff point of the eGFR for staging of patients with CKD stage 3 and possible risk factors for progression to ESRD based on the following three components: (1) identifying the cutoff points according to the data distribution; (2) verifying the demarcation point using an eGFR of 40–48 as the dividing points for stage 3a/3b CKD; and (3) assessing the risk factors of stage 3a/3b CKD by using RF analysis. This study is based on the Central South University medical big data project subject platform [13], using Spearman correlation coefficient (SCC) analysis, algorithms including LR, RF, SVMs, and Nnets, to explore whether the KDIGO stage criteria for patients in South Central China with stage 3a/3b CKD are suitable. Moreover, we explored factors that influenced the prognosis of patients with stage 3a/3b CKD with the new criteria.

Materials and methods

Study design

We conducted a retrospective cohort study using the full text of clinical notes in the year when the patients first met the criteria for CKD stage 3 ($30 \leq$ eGFR < 60 mL/ min 1.73 m^2). All the clinical data were extracted from the electronic medical records system (EMRS). The data were analyzed using the CSM system. All identified events were adjudicated through chart review.

CKD stage 3 modeling (CSM)

The CSM approach is a prediction model based on an artificial intelligence core intended to identify a new cutoff point of 43 in patients with CKD stage 3 and to distinguish the different factors related to progression of stage 3a/b CKD to ESRD (Fig. 1). The predictive model at the CSM core

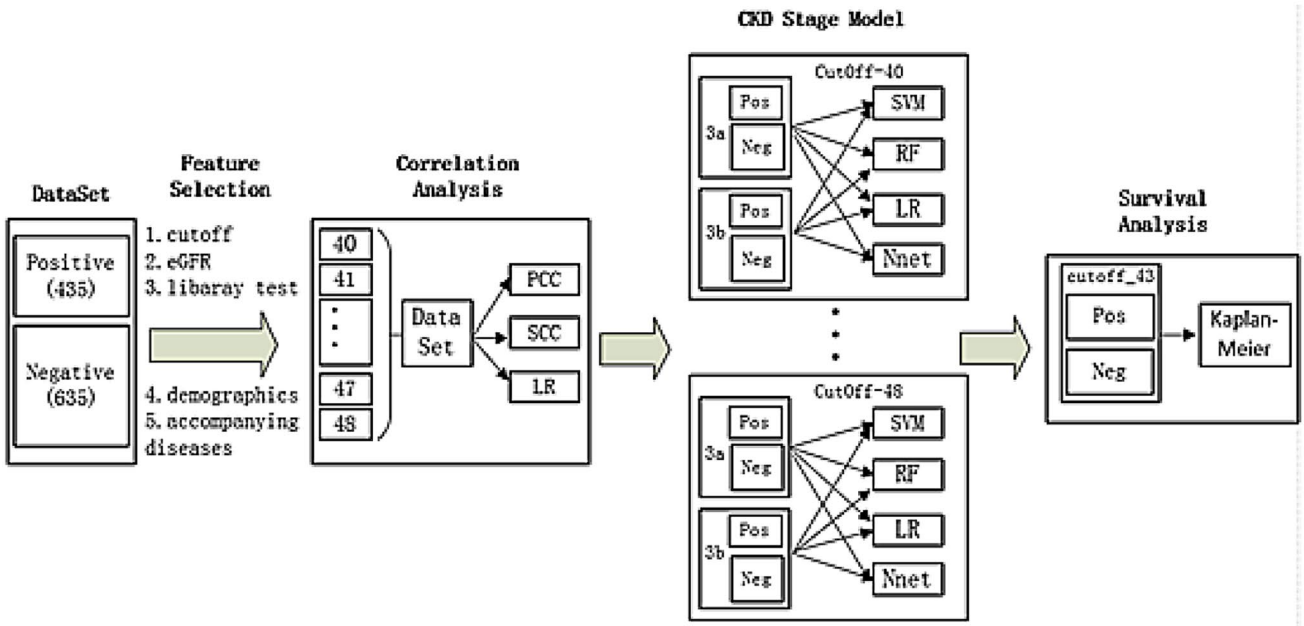

Fig. 1 Flow diagram of CKD stage model

is a machine learning model that was trained on a set of almost 50,049 clinical records; records were extracted from the clinical information system. The artificial neural network was trained to predict a suitable cutoff point for CKD stage 3 among patient in South Central China and to identify influential factors for progression of CKD stage 3a/3b to CKD stage 5 based on the parameters listed in (Table 1).

Population studied

Inpatients and outpatients with an eGFR between 30 and 60 mL/min 1.73 m^2 who were treated at Xiangya Hospital between August 1, 2010, and April 1, 2018 were included. One of the criteria for patient's enrollment was that patients were followed up at least once a year. If the patient had multiple records in a year, each record would be obtained. The time of onset of CKD stage 3 was recorded as the first time CKD stage 3 was diagnosed. The time when the eGFR decreased to less than 15 mL/min 1.73 m^2 was also recorded. The eGFR was determined with the CKD Epidemiology Collaboration equation (CKD-EPI) for Chinese patients with CKD [8]. We screened patients in with stage 3 and CKD stage 5, recorded the eGFR and the date they were first included in each cohort, extracted the intersection of both cohorts, and set the date they were first diagnosed with CKD stage 3 before progression to CKD stage 5. Concrete exclusion criteria included the following: acute kidney injury (AKI) (2012 KDIGO guidelines); age < 18 or > 70 years; the first time CKD stage 3 was diagnosed that was later than the time of diagnosis of CKD stage 5; Incomplete clinical data; and hemodialysis, peritoneal dialysis, and kidney transplantation patients.

Study outcomes

ESRD was defined as the initiation of irreversible development of an eGFR < 15 mL/min 1.73 m^2. The ultimate ascertainment of eGFR is based on the values from a central laboratory. ESRD events were adjudicated by an independent committee consisting of relevant specialist physicians.

Data collection

We obtained data from the EMRS of Xiangya hospital. We collected information on patient demographics (name, ID, age, sex), diagnosis, accompanying diseases (diabetes, hypertension, and cardiovascular disease) and the laboratory data urine nitrite(NIT), urobilinogen(URO), bilirubinuria(BiL), urine specific gravity (SG), urine white blood cells(WBC), urine vitamin C(Vitamin C), glucosuria(Glu), proteinuria, ketonuria(Ket), urine pH(PH), neutrophil percentage(NeuTP), neutrophil count(NeuT), monocyte percentage(MONOP), monocyte count(MONO), basophil percentage(BASOP), basophil count(BASON), eosinophil percentage (EOP), eosinophil count(EON), mean corpuscular volume (MCV), mean platelet volume (MPV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), white blood cell count (WBC), red blood cell volume distribution width (RDW),

Table 1 Patient characteristics in the two categories of patients in the facility level analysis

Characteristic	All	POS	NEG	*p* value
Total no. of patients	1090	455	635	
Chronic nephritis	518(47.5)	252(55.4)	266(41.9)	$<0.001^{a}$
Diabetic kidney disease	229(21)	143(31.4)	86(13.5)	$<0.001^{a}$
Others	343 (31.5)	60 (13.2)	283 (44.6)	$<0.001^{a}$
Urine nitrite (pos), no. (%)	18 (2.96)	10 (3.23)	8 (2.68)	0.81^{a}
urobilinogen (pos), no. (%)	6 (0.99)	2 (0.65)	4 (1.34)	0.443^{a}
Urine bilirubin (pos), no. (%)	19 (3.12)	9 (2.9)	10 (3.34)	0.818^{a}
Urine specific gravity of urine, median (IQR)	1.017 (0.01)	1.016 (0.01)	1.018 (0.01)	$<0.001^{c}$
Urine white blood cell. (pos), no. (%)	120 (20.24)	48 (16.16)	72 (24.32)	0.014^{a}
Urine vitamin C (pos), no. (%)	66 (13.69)	34 (13.82)	32 (13.56)	1^{a}
Urine glucose (pos), no. (%)	80 (15.47)	64 (29.36)	16 (5.35)	$<0.001^{a}$
Urine protein (pos), no. (%)	378 (62.07)	268 (86.45)	110 (36.79)	$<0.001^{a}$
Urine ketone bodies (pos), no. (%)	20 (3.28)	8 (2.58)	12 (4.01)	0.368^{a}
Urine_pH, median (IQR)	5.5 (1)	5.59 (1)	5.59 (1)	0.002^{c}
Blood_HDL/TC, mean ± SD	0.26 ± 0.08	0.25 ± 0.08	0.27 ± 0.08	0.002^{b}
Blood neutrophils percentage, mean ± SD	65.21 ± 11.15	66.58 ± 10.89	64.04 ± 11.25	0.001^{b}
Blood neutrophils ($\times 10^9$/L), mean ± SD	5.03 ± 2.62	5.25 ± 2.53	4.85 ± 2.69	0.028^{b}
Serum LDL (mmol/L), mean ± SD	3.36 ± 1.35	3.58 ± 1.56	3.15 ± 1.05	$<0.001^{b}$
Blood monocyte percentage, mean ± SD	6.82 ± 3.03	6.75 ± 3.38	6.88 ± 2.7	0.57^{b}
Blood monocyte ($\times 10^9$/L), mean ± SD	0.5 ± 0.29	0.51 ± 0.32	0.5 ± 0.27	0.459^{b}
Blood basophils percentage, mean ± SD	0.5 ± 0.44	0.51 ± 0.39	0.49 ± 0.48	0.534^{b}
Blood basophils ($\times 10^9$/L), median (IQR)	0.02 (0.06)	0.02 (0.05)	0.02 (0.07)	0.279^{b}
Blood eosinophils percentage, mean ± SD	2.23 ± 2.28	2.46 ± 2.62	2.03 ± 1.94	0.009^{b}
Blood eosinophils ($\times 10^9$/L), mean ± SD	0.16 ± 0.2	0.18 ± 0.24	0.14 ± 0.15	0.003^{b}
Blood mean corpuscular volume (MCV), mean ± SD	90.72 ± 6.77	89.86 ± 6.63	91.46 ± 6.81	0.001^{b}
Blood mean platelet volume (MPV), mean ± SD	9.02 ± 1.49	9.05 ± 1.47	8.99 ± 1.51	0.548^{b}
Blood mean corpuscular hemoglobin (MCH), mean ± SD	30.29 ± 2.38	29.93 ± 2.15	30.6 ± 2.52	$<0.001^{b}$
Blood mean corpuscular hemoglobin concentration (MCHC), mean ± SD	333.6 ± 13.4	332.5 ± 14.2	334.6 ± 12.3	0.033^{b}
Serum urea (mmol/L), mean ± SD	7.72 ± 2.9	8.73 ± 3.43	6.99 ± 2.17	$<0.001^{b}$
Serum uric acid (umol/L), mean ± SD	419 ± 111.9	437 ± 111.9	406.1 ± 110.3	$<0.001^{b}$
Blood lymphocyte percentage, mean ± SD	25.25 ± 9.8	23.62 ± 9.13	26.63 ± 10.14	$<0.001^{b}$
Blood lymphocyte ($\times 10^9$/L), mean ± SD	1.77 ± 0.74	1.73 ± 0.78	1.81 ± 0.7	0.174^{b}
Blood white blood cell (WBC) ($\times 10^9$/L), mean ± SD	7.5 ± 2.92	7.71 ± 2.83	7.32 ± 2.98	0.056^{b}
Blood red blood cell volume distribution width (RDW), mean ± SD	14.1 ± 3.85	14.65 ± 5.27	13.65 ± 1.97	0.001^{b}
Blood hematocrit (HCT), mean ± SD	37.1 ± 6.51	35.35 ± 6.56	38.57 ± 6.1	$<0.001^{b}$
Blood red blood cell (RBC) ($\times 10^{12}$/L), mean ± SD	4.09 ± 0.75	3.94 ± 0.76	4.23 ± 0.72	$<0.001^{b}$
Blood platelet volume distribution width (PDW), mean ± SD	16.74 ± 1.43	16.59 ± 1.73	16.89 ± 1.04	0.011^{b}
Blood thrombocytocrit, median (IQR)	0.18 (0.07)	0.19 (0.07)	0.17 (0.07)	0.001^{b}
Blood platelet ($\times 10^9$/L), mean ± SD	199.38 ± 81.26	209.91 ± 79.53	190.42 ± 81.73	0.001^{b}
Blood hemoglobin (g/L), mean ± SD	123.58 ± 22.22	117.56 ± 22.29	128.7 ± 20.86	$<0.001^{b}$
Serum total protein (TP) (g/L), mean ± SD	68.68 ± 9.74	64.79 ± 9.46	71.68 ± 8.86	$<0.001^{b}$
Serum chlorine (mmol/L), mean ± SD	104.7 ± 4.8	105.16 ± 5.2	104.01 ± 4.3	0.008^{b}
Serum total bile acid (TBA) (umol/L), mean ± SD	5 ± 9.5	4.44 ± 5.53	5.51 ± 12	0.096^{b}
Serum total bilirubin (TBIL) (umol/L), mean ± SD	10.47 ± 14.09	9.18 ± 18.96	11.49 ± 8.36	0.025^{b}
Serum globulin (g/L), mean ± SD	28.73 ± 5.61	28.21 ± 4.74	29.13 ± 6.17	0.011^{b}
Serum triglyceride (TG) (mmol/L), mean ± SD	2.38 ± 2.07	2.6 ± 2.4	2.15 ± 1.65	0.008^{b}
Serum A/G, mean ± SD	1.43 ± 0.33	1.33 ± 0.32	1.51 ± 0.31	$<0.001^{b}$
Serum albumin (ALB) (g/L), mean ± SD	39.95 ± 7.31	36.57 ± 7.44	42.56 ± 6.04	$<0.001^{b}$
Serum direct bilirubin (umol/L), mean ± SD	4.12 ± 7.03	3.71 ± 9.85	4.43 ± 3.46	0.166^{b}
Serum creatinine (sCr) (umol/L), mean ± SD	143.9 ± 35.41	161.9 ± 38.94	131.1 ± 25.93	$<0.001^{b}$

Table 1 (continued)

Characteristic	All	POS	NEG	*p* value
Serum cholesterol (mmol/L), mean ± SD	5.56 ± 1.68	5.82 ± 1.92	5.28 ± 1.33	< 0.001[b]
Serum glucose (mmol/L), mean ± SD	6.39 ± 2.98	7 ± 3.97	6.09 ± 2.3	0.007[b]
Serum ALT (U/L), mean ± SD	28.77 ± 59.35	27.98 ± 63.88	29.38 ± 55.58	0.729[b]
Serum AST (U/L), mean ± SD	29.58 ± 47.57	29.08 ± 42.16	30.04 ± 52.05	0.77[b]
Serum calcium (mmol/L)	2.24 ± 0.19	2.21 ± 0.18	2.28 ± 0.2	< 0.001[b]
Serum sodium (mmol/L)	141.45 ± 3.79	141.74 ± 3.97	141.02 ± 3.46	0.033[b]
Serum potassium (mmol/L)	4.1 ± 0.54	4.11 ± 0.59	4.1 ± 0.48	0.844[b]
Serum high-density lipoprotein (HDL) (mmol/L), mean ± SD	1.42 ± 0.49	1.42 ± 0.54	1.42 ± 0.44	0.912[b]
eGFR (ml/min/1.73m2), median (IQR)	45.6 (16.16)	40.79 (13.74)	49.04 (12.77)	< 0.001[c]
Male (sex), no. (%)	614 (56.3)	266 (58.4)	348 (54.8)	0.24[a]
Hypertension, no. (%)	333 (30.6)	183 (40.2)	150 (23.7)	< 0.001[a]
Cardiovascular disease, no. (%)	29 (2.7)	16 (3.5)	13 (2)	0.18 [a]
Age (year), mean ± SD	50.01 ± 11.39	47.9 ± 12.3	51.57 ± 10.41	< 0.001[b]

POS positive group, progression to ESRD

NEG negative group, no progression to ESRD

[a]Fisher's exact test

[b]Unpaired *t* test

[c]Wilcoxon test

hematocrit (HCT), red blood cell count (RBC), lymphocyte percentage (LYMPHP), lymphocyte count (LYMPHN), platelet volume distribution width (PDW), thrombocytocrit (PCT), platelet (PLT), hemoglobin (HGB), total bile acids (TBA), total bilirubin (TBIL), total protein (TP), albumin (ALB), globulin(GLB), albumin-to-globin ratio (A/G), direct bilirubin (DBIL), blood high-density lipoprotein cholesterol-to-total cholesterol (HDL/TC), low-density lipoprotein(LDL), alanine aminotransferase (ALT), aspartate aminotransferase (AST), cholesterol(TC), chloride(CL), triglycerides (TG), high-density lipoprotein (HDL), serum creatinine (sCr), urea(UREA), uric acid(URIC), glucose(Glu), calcium(Ca), sodium(Na), potassium(K), and eGFR. We defined baseline laboratory values for each laboratory test as the first available result on or after the first diagnosis of CKD stage 3. Hypertension was defined as a systolic blood pressure (BP ≥ 140 mmHg and/or a diastolic BP ≥ 90 mmHg, or diagnosis of hypertension. Patients were considered to have diabetes mellitus if they had a fasting glucose ≥ 7.0 mmol/L; an HbA1c ≥ 6.5%; or diagnosis of diabetes. If ESRD did not occur by the average time of progression among patients with CKD stage 3, the observation was censored. This study was approved by the Ethics Committee of Xiangya Hospital, and the need for informed consent was waived. We adhered to the Declaration of Helsinki.

Data processing

Firstly, correlation analysis between different CKD stage 3a/3b cutoff point and time progress in the study period was carried out. Secondly, the function cor. test () was used to calculate the PCC, SCC and their related p values. The LR also was used to calculate their related *p* values. Later, four models were built, including the linear and nonlinear models—LR, RF, SVM, and Nnet, for each CKD stage3a and CKD3b group. Logistic regression was another generalized linear model (GLM) procedure using the same basic formula, but instead of the continuous Y, it was regressing for the probability of a categorical outcome. RF were conducted using the functions random forest () in the package "random forest". SVM were conducted using the functions svm () in the package "e1071". Nnet were conducted using the functions nnet () in the package "nnet". Fivefold cross-validation were used to evaluate the statistical models, LR, RF, SVM, and Nnet. In the fivefold cross-validation, the sample data are randomly partitioned into five equal groups. Each time, one group of data was retained as the validation data for testing the model, and the remaining four groups were used as training data. This process was then repeated five times, with each group used exactly once as the validation data. To reduce variability, five rounds of cross-validation were performed using different partitions, and the validation results were averaged over the rounds. The performance of the model was evaluated based on the comparison between predicted and observed number of patients whether progressed into CKD stage 5. Thirdly, CSM continued to search risk factors of progression to ESRD in CKD stage 3a/3b patients by RF.

Results

After applying the inclusion and exclusion criteria, we identified 1090 patients who constituted the analytic cohort (Table 1). Among them, 455 were confirmed to have developed ESRD during follow-up (positive group). The median follow-up time was 4.0 years [95% confidence interval (CI), 4.295–4.489]. This work focuses on the use of machine learning to predict disease risk and model the contributing factors learned from an electronic health record dataset.

43 mL/ min 1.73 m^2 may be the new cutoff point for predicting CKD stage 3 progression to ESRD among patients in South Central China

Hypothesis tests between the patients who progressed CKD stage 5 and those who did not were carried out for each feature. We analyzed the relationship between the eGFR and progression time by a scatter density map and contour line and found that two high density regions could be distinguished when the eGFR was 43 mL/min·1.73 m^2 and when the eGFR was 45 mL/min·1.73 m^2 (Fig. 2).

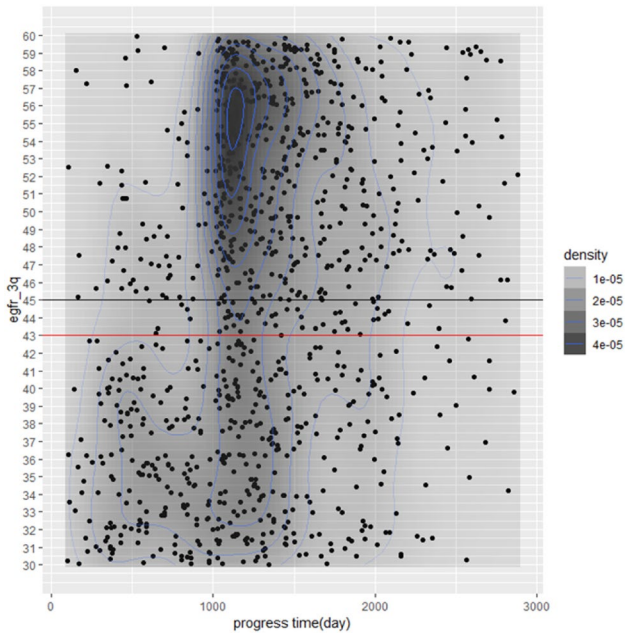

Fig. 2 Scatter plot for finding the eGFR cutoff point. On the horizontal axis was the time of progression (days), and on the vertical axis was the patient's first eGFR value. The distribution rule for the eGFR value and the time of progression was given through the density curve, and it could be seen that the region with the highest density was close to the region with a distance of 45, and the region with the highest density was 43, which could better distinguish the density region

Furthermore, all patients' samples were split into CKD stage 3a and CKD stage 3b groups by different eGFRs (range from 40 to 48 mL/min·1.73 m^2). PCC and SCC were used to find the best values of eGFR to distinguish CKD stage 3a from 3b to progression to CKD stage 5. It is showed that when the eGFR is 43 mL/ min·1.73 m^2, the correlation coefficient is the largest (Table 2). Further, we used a logistic regression model to measure the eGFR divided CKD3 patients into two groups and the time of progression to CKD stage 5. We found that when CKD3 patients were classified by an eGFR of 43 mL/ min·1.73 m^2, the regression coefficient and significance were prominent (Fig. 3, Table 3).

According to an eGFR of 40 and 48 mL/min·1.73 m^2 as the dichotomy for stage 3a and CKD stage 3b, respectively, using four types of algorithms to distinguish stage 3a and 3b CKD, a classification model and model performance comparison reference appendix were established. Based on the eGFR cutoff point of 43 mL/min·1.73 m^2, Random Forest model performed the best for distinguishing stage 3a and 3b CKD patients who would progress to ESRD. As shown in the figure below, for stage 3a and 3b CKD, this model had an accuracy of 85% and 77%, respectively, and an area under the curve (AUC) value of 88% and 83%, respectively, which includes all the variable values in (Table 1) (Fig. 4a, b).

Screening predictors of CKD stage 3a/3b progression to CKD stage 5 by RF

After establishing a reliable forecast model, we used an RF to clarify the different risk factors for progression of stage 3a/3b CKD to CKD stage 5. The risk factors for stage 3a/3b CKD were explored by modeling with an RF at an eGFR cutoff point of 43 mL/min 1.73 m^2, and the results of the importance assessment and analysis of the model parameters are given (Fig. 5). The higher the value of mean decrease accuracy or mean decrease Gini score was, the higher the importance of the variable in the model.

Table 2 Correlation between the eGFR and progression time by pearson and spearman correlation analysis

eGFR (mL/ min·1.73 m^2)	40	41	42	43	44	45	46	47	48
SCC	0.230	0.226	0.223	0.233	0.210	0.197	0.193	0.191	0.171
PCC	0.230	0.226	0.224	0.232	0.210	0.198	0.197	0.192	0.17

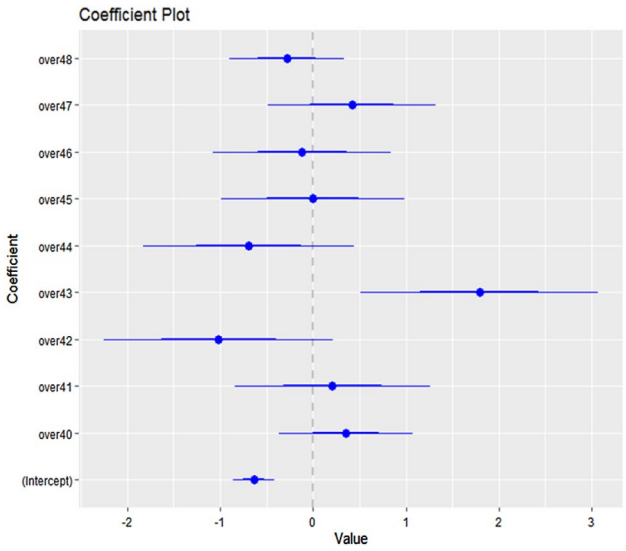

Fig.3 Correlation between the eGFR and progression time with logistic regression measures. The figure shows the regression results. The estimated value of each coefficient is a point, the bold line represents a confidence interval for the standard error, and the thin line represents a confidence interval for twice the standard error. The vertical line is zero. To evaluate statistical significance, we evaluated whether the double confidence interval contained 0; if it did not, the result was statistically significant

The common influencing factors of stage 3a/3b CKD progression to CKD stage 5 included serum SCR, eGFR, serum TP, serum TC, serum urea, serum albumin, serum TBIL, serum HDL proteinuria, age. Furthermore, serum DBIL, serum A/G, blood HGB, serum Ca, blood HCT, serum ALT, urine SG accounted for the progression of CKD stage 3a to CKD stage 5. The contributing factors for CKD stage 3b progression to CKD stage 5 included blood EOP, blood MCH), diabetic kidney disease, blood EON, serum Na, serum Cl.

Incidence rates of ESRD events according to the cutoff of the eGFR of 43 mL/min·1.73 m^2

The incidence rates of ESRD events according to the cutoff of the eGFR of 43 mL/min·1.73 m^2 are shown in (Table 4). During the median follow-up of 4.0 years (95% CI, 4.295–4.489), higher incidence rates of ESRD events were observed in CKD with a decreased eGFR (Table 4, Fig. 6, p for log-rank test < 0.001).

Discussion and conclusion

The KDIGO guidelines on CKD represent an extraordinary effort to summarize and synthesize evidence together with a thoughtful expression of the best practices and opinion [14]. One of the meaningful suggestions was the division of stage 3a and 3b CKD. It was suggested that it would be clinically sound to subdivide CKD stage 3 into stages 3a (45–59 mL/ min 1.73 m^2) and 3b (30–44 mL/ min 1.73 m^2), as these two ranges may be associated with different clinical patterns and risks. It has recently been shown that patients with CKD and an eGFR < 45 mL/ min 1.73 m^2, particularly older patients, experience faster disease progression [15]. Patients with CKD stage 3b should probably be referred earlier for specialized renal care [16]. Some have recommended that people with an eGFR category CKD stage 3a without associated markers of kidney damage (proteinuria or hematuria) should not necessarily be considered to have CKD and should be considered for further evaluation and referral according to the clinical judgment of the health care provider [17, 18].

CKD is a global health challenge, especially in low- and middle-income countries. China is a large developing country with different health care and primary care structures, and some recommendations by the international guidelines' groups might not be relevant to the Chinese population. First, the prevalence of CKD stage 3 was 1.6% in China compared with 7.7% in the USA and 4.2% in Norway [3]. The findings described rise in the prevalence of diabetes in China [19, 20], a signal strongly forewarning a growing epidemic of CKD in China in the upcoming years to decades, perhaps analogous to trends seen in the United States from the 1980 s to early 2000 s [21]. Wen et al. reported the prevalence of CKD and its stages among the general population in Taiwan [22], where the ethnicity and living habits were the same as in Mainland China, but the economic development was better, and they found a higher proportion of lower eGFR (CKD stage 3 or worse) than that reported among the population in Mainland China [3]. With respect to CKD in China, there were twice as many people with proteinuria than those with a low eGFR, while in the US, the prevalence difference in a low eGFR and proteinuria was much smaller than in China [23]. Taken together, with the different prevalence of CKD, racial differences, economic development, genetic, and environmental backgrounds between China and Western countries, we should evaluate the guidelines according to our actual situation rather than simply adhering to the recommendations.

Table 3 Correlation between the eGFR and progression time by logistic regression measures

eGFR (mL/min·1.73 m^2)	40	41	42	43	44	45	46	47	48
p value	0.328	0.693	0.100	0.005**	0.221	0.070	0.806	0.356	0.359

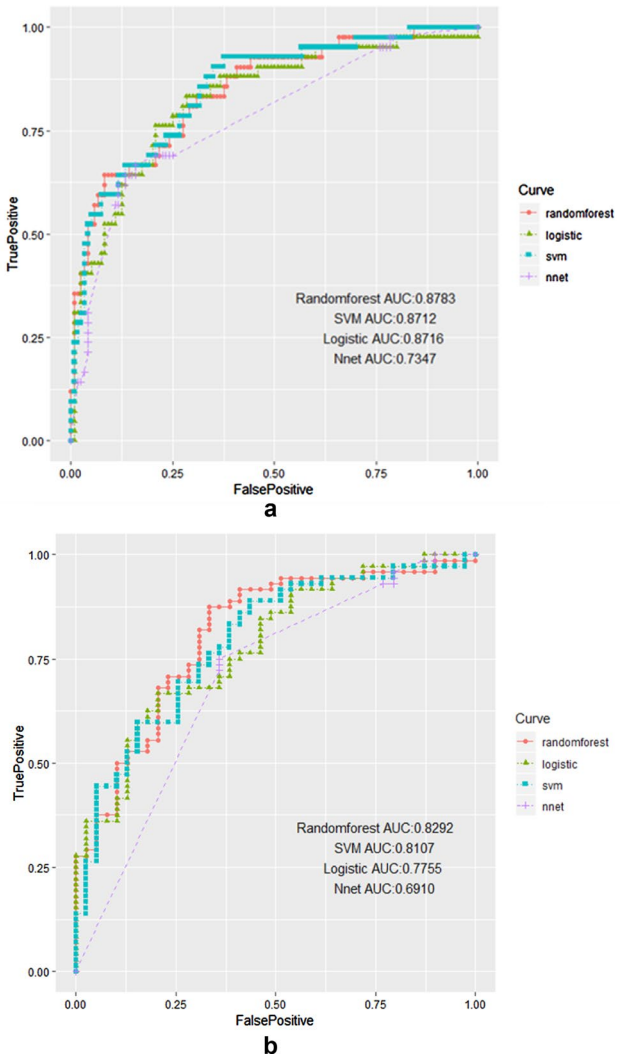

Fig. 4 **a** Comparison of unified models with different algorithms in CKD stage 3a patients. **b** Comparison of unified models with different algorithms in CKD stage 3b patients. The figure shows the regression results. The estimated value of each coefficient is a point, the bold line represents a confidence interval for the standard error, and the thin line represents a confidence interval for twice the standard error. Random forest, support vector machine, logistic regression, and neural network algorithms were used to construct a classification model to predict whether patients with 43 < eGFR < 60 mL/min·1.73 m^2 would progress to CKD stage 5. The random forest model had the largest AUC value (0.8783), indicating that the model has the best prediction effect. **b** The estimated value of each coefficient is a point, the bold line represents a confidence interval for the standard error, and the thin line represents a confidence interval for twice the standard error. Random forest, support vector machine, logistic regression and neural network algorithms were used to construct a classification model to predict whether patients with 30 ≤ eGFR < 43 mL/min·1.73 m^2 would progress to CKD stage 5. The random forest model had the largest AUC value (0.8292), indicating that the model had the best prediction effect

As the adoption of electronic health records continues to rise and a generation of individuals has their entire health histories stored electronically, this approach provides a novel way to gain potential insights about the disease risk as a natural byproduct of care delivery and electronic health record documentation [24]. Mathematical and statistical tools developed in the field of artificial intelligence (AI) and machine learning are well poised to assist clinical researchers in deciphering complex predictive patterns in healthcare data [25]. It is challenging for humans to directly analyze these massive data; this is not only because of the massive time required and cares needed to avoid human errors, but also the ability to derive the insights or information in depth. Clearly, machine learning holds nonparallel advantages over humans in these domains [26]. Unlike the previous CKD stage 3 classification studies, this is the first study to use an unbiased machine learning approach using text from clinical notes to identify appropriate cutoff points for patients with CKD stage 3, determine different risk factors for CKD stage 3a and 3b, more importantly, build a model to predict the possibility of progression to ESRD in a predetermined period. The face validity of this approach was confirmed by different calculation methods of AI. This study also conducts proposed methods to extract insights about performance trends that cannot be easily extrapolated using standard analyses and treats various influencing factors according to the model set by the CSM approach.

In this computer-based retrospective analysis, we confirmed that it is clinically significant to divide CKD stage 3 patients into CKD stage 3a and 3b. More importantly, machine learning, when applied to predictive modeling, can determine patterns of risk factors useful for improving prediction quality [27]. In this study, the identification of several well-established risk factors for ESRD in CKD stage 3 patients, including age [1], proteinuria [1], diabetic kidney disease [1], eGFR [1], serum ALB [2], creatinine [28], blood urea [29], hematocrit [2], serum cholesterol [30], HDL cholesterol [31], HGB [27], TBIL [32], DBIL [33], serum ALT [34], serum Na [35], serum Cl [35],and serum calcium [36] were indicated by machine learning. In addition, the machine learning method also identified some risk factors that have not been previously described, such as A/G, MCH, urine SG, TP, EOP, and EON future research is needed to determine the possible role of these factors in the progression of CKD.

In addition, there are different factors associated with progression from stage 3a and 3b CKD to CKD stage 5. Apart from common factors of CKD stage 3 progression to CKD stage 5, serum DBIL, serum A/G, blood HGB, serum Ca, blood HCT, serum ALT, urine SG accounted for the progression from CKD stage 3a to CKD stage 5. The contributing factors for CKD stage 3b progression to CKD stage

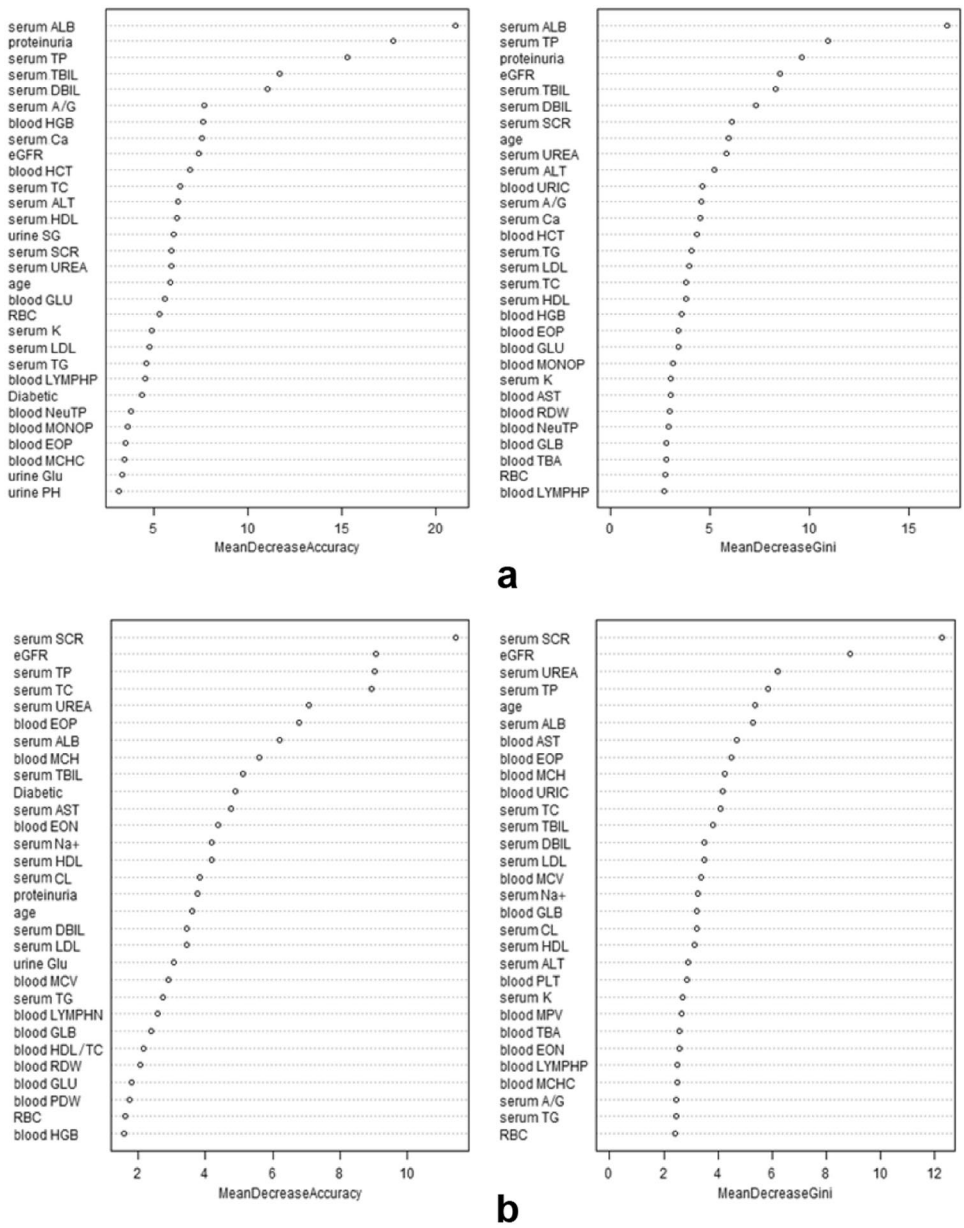

Fig. 5 **a** Important variables in the model of patients with CKD stage 3a who progressed to CKD stage 5 by random forest. **b** Important variables in the model of patients with stage 3b CKD who progressed to CKD stage 5 by random forest**.** Serum albumin, proteinuria, serum TP, serum TBIL, serum DBIL serum A/G, blood HGB, serum Ca, eGFR, blood HCT, serum TC, serum ALT, serum HDL, urine SG, serum SCR, serum urea, age accounted for the progression from stage 3a CKD to CKD stage 5. Serum SCR, eGFR, serum TP, serum TC, serum urea, EOP, serum ALB, blood MCH, serum TBIL, Diabetes, blood EON, serum Na, serum HDL, serum Cl, proteinuria, age accounted for the progression from stage 3b CKD to CKD stage 5

Table 4 Relationship between the cutoff point of the eGFR for stage 3 CKD patients of 43 ml/min·1.73 m^2 and ESRD event rates

eGFR cutoff (mL/min 1.7 m^2)	Number of events	*p* for log-rank
ESRD events		< 0.0001
Stage 3a CKD (43 < eGFR < 60) (*N* = 647)	167 (25.8%)	
Stage 3b CKD (30 < eGFR < 43) (*N* = 443)	288 (65.0%)	
Total	455 (41.74%)	

5 include blood EOP, blood MCH), diabetic kidney disease, blood EON, serum Na, serum Cl. These findings may remind clinicians to pay attention to different factors in patients with stage 3a and 3b CKD.

This is the first study to use an unbiased approach using text from clinical notes to identify predictors of progression to ESRD among CKD stage 3 patients. Our work confirmed that it is reasonable to divide CKD stage 3 into stage 3a and 3b. Besides, eGFR cutoff point of 43 mL/min 1.73 m^2 is a suitable cutoff point by predicting progression to ESRD in Central South Chinese patients using different machine learning methods. More important, our

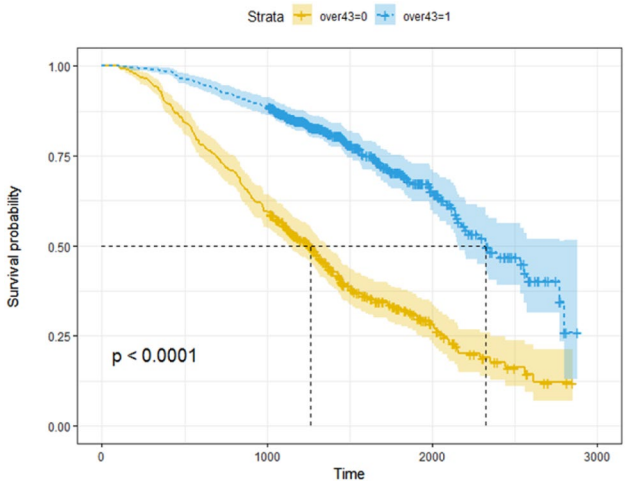

Fig. 6 Kaplan–Meier curve for ESRD events according to the cutoff of eGFR for stage 3 CKD patients of 43 mL/min 1.73 m^2**.** The survival curve of patients with stage 3a CKD was smoother than that of patients with stage 3b CKD, indicating that the prognosis of patients with stage 3a CKD was better than that of patients with stage 3b CKD

findings may provide clinical proof of the beneficial effects of deploying the CSM approach in everyday practice as part of routine nephrological practice. As the adoption of electronic health records continues to rise and a generation of individuals has their entire health histories stored electronically, this approach provides a novel way to gain potential insights about disease risk as a natural byproduct of care delivery and electronic health record documentation [24]. As systems analytics, big data, and machine learning, among others, come online and become more widely available, we may be able to tackle CKD more holistically, efficiently, and satisfactorily.

This study has some limitations. The primary limitation of this study is that its findings are drawn from a single tertiary hospital, which may have idiosyncrasies in documentation style and patient characteristics that may differ from other institutions. Validating this analysis in other cohorts is needed. This approach was successful in translating the clinical narrative into a tool for the discovery of possible predictors that have not been previously linked to kidney failure. Second, if low-risk patients were systematically excluded from these cohorts due to lack of follow-up creatinine testing, then estimates from the resulting models could overestimate risk of advanced chronic kidney disease. Third, the kidney disease outcome evaluated in this paper was progression to ESRD, future prospective studies may also include death or cardiovascular events as other outcomes either. Fourth, based on the retrospective study, we cannot collect the treatment information correctly. We will conduct prospective research to collect more detailed data to replicate our findings and approach in multicenters and determine the cutoff point of different stage of CKD in the future.

In summary, our findings confirm a new cutoff point for CKD stage 3 by computational intelligence, which is different from a previous study. The CSM approach provides a novel tool to identify the different influencing factors for stage 3a/3b CKD progression to CKD stage 5. The CSM approach may be adapted and used in the management of other chronic diseases in which international guidelines require confirmation in different populations.

Acknowledgment This work was supported in part by the Hunan Provincial Natural Science Foundation of China under Grant no. 2018JJ5056 and by the QUALCOMM university-sponsored program and Xiangya Clinical Big Data Project of Central South University. We thank Yidu Cloud (Beijing) Technology Co., Ltd. for technical assistance.

Compliance with ethical standards

Conflict of interest The authors have declared that no conflict of interest exists.

Human and animal rights All procedures performed in studies involving human participants were in accordance with the ethical standards of the Ethics Committee (Xiangya Hospital, Central South University, No. 2018121290) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Coresh J, Turin TC, Matsushita K, Sang Y, Ballew SH, Appel LJ, Arima H, Chadban SJ, Cirillo M, Djurdjev O, Green JA, Heine GH, Inker LA, Irie F, Ishani A, Ix JH, Kovesdy CP, Marks A, Ohkubo T, Shalev V, Shankar A, Wen CP, de Jong PE, Iseki K, Stengel B, Gansevoort RT, Levey AS. Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. JAMA. 2014;25:12–4.
2. Li L, Chang A, Rostand SG, Hebert L, Appel LJ, Astor BC, Lipkowitz MS, Wright JT, Kendrick C, Wang X, Greene TH. A within-patient analysis for time-varying risk factors of CKD progression. J Am Soc Nephrol. 2014;25(3):606–13.
3. Zhang L, Wang F, Wang L, Wang W, Liu B, Liu J, Chen M, He Q, Liao Y, Yu X, Chen N, Zhang JE, Hu Z, Liu F, Hong D, Ma L, Liu H, Zhou X, Chen J, Pan L, Chen W, Wang W, Li X, Wang H. Prevalence of chronic kidney disease in China: a cross-sectional survey. Lancet. 2012;379(9818):815–22.
4. Zhang L, Wang H. Chronic kidney disease epidemic: cost and health care implications in China. Semin Nephrol. 2009;29(5):483–6.
5. Levey AS, de Jong PE, Coresh J, El Nahas M, Astor BC, Matsushita K, Gansevoort RT, Kasiske BL, Eckardt KU. The definition, classification, and prognosis of chronic kidney disease: a KDIGO controversies conference report. Kidney Int. 2011;80(1):17–28.
6. Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease. Lancet. 2017;389(10):1238–52.

7. Mills KT, Xu Y, Zhang W, Bundy JD, Chen CS, Kelly TN, Chen J, He J. A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010. Kidney Int. 2015;88(5):950–7.
8. Peng Z, Wang J, Yuan Q, Xiao X, Xu H, Xie Y, Wang W, Huang L, Zhong Y, Ao X, Zhang L, Zhao M, Tao L, Zhou Q. Clinical features and CKD-related quality of life in patients with CKD G3a and CKD G3b in China: results from the Chinese cohort study of chronic kidney disease (C-STRIDE). BMC Nephrol. 2017;18(1):311.
9. Lin H, Long E, Ding X, Diao H, Chen Z, Liu R, Huang J, Cai J, Xu S, Zhang X, Wang D, Chen K, Yu T, Wu D, Zhao X, Liu Z, Wu X, Jiang Y, Yang X, Cui D, Liu W, Zheng Y, Luo L, Wang H, Chan CC, Morgan IG, He M, Liu Y. Prediction of myopia development among Chinese school-aged children using refraction data from electronic medical records: a retrospective, multicentre machine learning study. PLoS Med. 2018;15(11):100–2674.
10. Sun J, McNaughton CD, Zhang P, Perer A, Gkoulalas-Divanis A, Denny JC, Kirby J, Lasko T, Saip A, Malin BA. Predicting changes in hypertension control using electronic health records from a chronic disease management program. J Am Med Inform Assoc. 2014;21(2):337–44.
11. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, Mottram A, Meyer C, Ravuri S, Protsyuk I, Connell A, Hughes CO, Karthikesalingam A, Cornebise J, Montgomery H, Rees G, Laing C, Baker CR, Peterson K, Reeves R, Hassabis D, King D, Suleyman M, Back T, Nielson C, Ledsam JR, Mohamed S. A clinically applicable approach to continuous prediction of future acute kidney injury [J]. Nature. 2019;572(7767):116–9.
12. Li C, Yao Z, Zhu M, Lu B, Xu H. Biopsy-Free Prediction of Pathologic Type of Primary Nephrotic syndrome using a machine learning algorithm [J]. Kidney Blood Press Res. 2017;42(6):1045–52.
13. Li B, Li J, Jiang Y, Lan X. Experience and reflection from China's Xiangya medical big data project [J]. J Biomed Inform. 2019;93:1–6.
14. Inker LA, Astor BC, Fox CH, Isakova T, Lash JP, Peralta CA, Tamura MK, Feldman HI. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. Am J Kidney Dis. 2014;63(5):713–35.
15. O'Hare AM, Choi AI, Bertenthal D, Bacchetti P, Garg AX, Kaufman JS, Walter LC, Mehta KM, Steinman MA, Allon M, McClellan WM, Landefeld CS. Age affects outcomes in chronic kidney disease. J Am Soc Nephrol. 2007;18(10):2758–65.
16. Kirsztajn GM, Suassuna JH, Bastos MG. Dividing stage 3 of chronic kidney disease (CKD): 3A and 3B. Kidney Int. 2009;76(4):462–3.
17. Glassock RJ, El Nahas M, Winearls CG. Chronic kidney disease in Taiwan. Lancet. 2008;372:1949–50.
18. Delanaye P, Cavalier E. Staging chronic kidney disease and estimating glomerular filtration rate: an opinion paper about the new international recommendations. Clin Chem Lab Med. 2013;51(10):1911–7.
19. Yang W, Lu J, Weng J, Jia W, Ji L, Xiao J, Metabolic G. Prevalence of diabetes among men and women in China. N Engl J Med. 2010;362(12):1090–101.
20. Xu Y, Wang L, He J, Bi Y, Li M, Wang T, Wang L. Prevalence and control of diabetes in Chinese adults. JAMA. 2013;310(9):948–59.
21. Hsu CY, Vittinghoff E, Lin F, Shlipak MG. The incidence of end-stage renal disease is increasing faster than the prevalence of chronic renal insufficiency. Ann Intern Med. 2004;141(2):95–101.
22. Wen CP, Cheng TY, Tsai MK, Chang YC, Chan HT, Tsai SP, Chiang PH, Hsu CC, Sung PK, Hsu YH, Wen SF. All-cause mortality attributable to chronic kidney disease: a prospective cohort study based on 462 293 adults in Taiwan. Lancet. 2008;371(9631):2173–82.
23. Wang J, Wang F, Saran R, He Z, Zhao MH, Li Y, Zhang L, Bragg-Gresham J. Mortality risk of chronic kidney disease: a comparison between the adult populations in urban China and the United States. PLoS ONE. 2018;13(3):193–734.
24. Singh K, Betensky RA, Wright A, Curhan GC, Bates DW, Waikar SS. A concept-wide association study of clinical notes to discover new predictors of kidney failure. Clin J Am Soc Nephrol. 2016;11(12):2150–8.
25. Vemulapalli V, Qu J, Garren JM, Rodrigues LO, Kiebish MA, Sarangarajan R, Narain NR, Akmaev VR. Non-obvious correlations to disease management unraveled by Bayesian artificial intelligence analyses of CMS data. Artif Intell Med. 2016;74:1–8.
26. Yu ZG. Artificial Intelligence and Medical [J]. J Med Univer. 2018;39(8):1.
27. Orchard P, Agakova A, Pinnock H, Burton CD, Sarran C, Agakov F, McKinstry B. Improving prediction of risk of hospital admission in chronic obstructive pulmonary disease: application of machine learning to telemonitoring data. J Med Int Res. 2018;20(9):263.
28. Kim JS, Kim YJ, Ryoo SM, Sohn CH, Seo DW, Ahn S, Lim KS, Kim WY. One–Year progression and risk factors for the development of chronic kidney disease in septic shock patients with acute kidney injury: a single-centre retrospective cohort study. J Clin Med. 2018;7:12.
29. Whaley-Connell A, Sowers JR. Obesity and kidney disease: from population to basic science and the search for new therapeutic targets. Kidney Int. 2017;92:313–23.
30. Ritz E, Wanner C. Lipid changes and statins in chronic renal insufficiency. J Am Soc Nephrol. 2006;17(12 Suppl 3):S226–S23030.
31. Wen J, Chen Y, Huang Y, Lu Y, Liu X, Zhou H, Yuan H. Association of the TG/HDL-C and Non-HDL-C/HDL-C ratios with chronic kidney disease in an adult chinese population. Kidney Blood Press Res. 2017;42(6):1141–54.
32. Sakoh T, Nakayama M, Tanaka S, Yoshitomi R, Ura Y, Nishimoto H, Fukui A, Shikuwa Y, Tsuruya K, Kitazono T. Association of serum total bilirubin with renal outcome in Japanese patients with stages 3–5 chronic kidney disease. Metabolism. 2015;64(9):1096–102.
33. Wang J, Wang B, Liang M, Wang G, Li J, Zhang Y, Huo Y, Cui Y, Xu X, Qin X. Independent and combined effect of bilirubin and smoking on the progression of chronic kidney disease. Clin Epidemiol. 2018;10:121–32.
34. Tanaka M, Fukui M, Okada H, Senmaru T, Asano M, Akabame S, Yamazaki M, Tomiyasu K, Oda Y, Hasegawa G, Toda H, Nakamura N. Low serum bilirubin concentration is a predictor of chronic kidney disease. Atherosclerosis. 2014;234(2):421–5.
35. Maruta Y, Hasegawa T, Yamakoshi E, Nishiwaki H, Koiwa F, Imai E, Hishida A. Association between serum Na–Cl level and renal function decline in chronic kidney disease: results from the chronic kidney disease Japan cohort (CKD-JAC) study. Clin Exp Nephrol. 2019;23(2):215–22.
36. Winnicki E, McCulloch CE, Mitsnefes MM, Furth SL, Warady BA, Ku E. 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. 2018;172(2):174–80.