An integrated machine learning predictive scheme for longitudinal laboratory data to evaluate the factors determining renal function changes in patients with different chronic kidney disease stages

Ming-Hsien Tsai[1,2], Mao-Jhen Jhou[3], Tzu-Chi Liu[3],
Yu-Wei Fang[1,2] and Chi-Jie Lu[3,4,5]*

[1]Division of Nephrology, Department of Medicine, Shin Kong Wu Ho-Su Memorial Hospital, Taipei,
Taiwan, [2]Department of Medicine, School of Medicine, Fu Jen Catholic University, New Taipei City,
Taiwan, [3]Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City,
Taiwan, [4]Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, Taiwan,
[5]Department of Information Management, Fu Jen Catholic University, New Taipei City, Taiwan

Background and objectives: Chronic kidney disease (CKD) is a global health concern. This study aims to identify key factors associated with renal function changes using the proposed machine learning and important variable selection (ML&IVS) scheme on longitudinal laboratory data. The goal is to predict changes in the estimated glomerular filtration rate (eGFR) in a cohort of patients with CKD stages 3–5.

Design: A retrospective cohort study.

Setting and participants: A total of 710 outpatients who presented with stable nondialysis-dependent CKD stages 3–5 at the Shin-Kong Wu Ho-Su Memorial Hospital Medical Center from 2016 to 2021.

Methods: This study analyzed trimonthly laboratory data including 47 indicators. The proposed scheme used stochastic gradient boosting, multivariate adaptive regression splines, random forest, eXtreme gradient boosting, and light gradient boosting machine algorithms to evaluate the important factors for predicting the results of the fourth eGFR examination, especially in patients with CKD stage 3 and those with CKD stages 4–5, with or without diabetes mellitus (DM).

Main outcome measurement: Subsequent eGFR level after three consecutive laboratory data assessments.

Results: Our ML&IVS scheme demonstrated superior predictive capabilities and identified significant factors contributing to renal function changes in various CKD groups. The latest levels of eGFR, blood urea nitrogen (BUN), proteinuria, sodium, and systolic blood pressure as well as mean levels of eGFR, BUN, proteinuria, and triglyceride were the top 10 significantly important factors for predicting the subsequent eGFR level in patients with CKD stages 3–5. In individuals with DM, the latest levels of BUN and proteinuria, mean levels of phosphate and proteinuria, and variations in diastolic blood pressure levels emerged as important factors for predicting the decline of renal function. In individuals without DM, all phosphate

patterns and latest albumin levels were found to be key factors in the advanced CKD group. Moreover, proteinuria was identified as an important factor in the CKD stage 3 group without DM and CKD stages 4−5 group with DM.

Conclusion: The proposed scheme highlighted factors associated with renal function changes in different CKD conditions, offering valuable insights to physicians for raising awareness about renal function changes.

KEYWORDS

chronic kidney disease, diabetes, machine learning, estimated glomerular filtration, feature selection

1. Introduction

Chronic kidney disease (CKD), characterized by decreased glomerular filtration rate, is a global health concern with a high prevalence rate (10–15%); moreover, this disease is highly associated with morbidity and mortality, leading to financial and medical burdens (1–3). The gradual loss of kidney function in patients with CKD may lead to end-stage kidney disease (ESKD) that requires kidney replacement therapy. According to the latest annual report of the United States Renal Data System, the average annual increase in the ESKD prevalence worldwide from 2003 to 2016 ranged from 0.1 to 109 per million population (4), thus placing a greater burden on the health insurance system of many countries. The ESKD prevalence in Taiwan is high (4), with a substantial increase observed between 2010 and 2018 (5).

Timely intervention for delaying the progression of CKD to ESKD may not only improve the quality of life of patients but also reduce the associated morbidity and mortality (6, 7). As the exacerbation of renal function in patients with CKD is usually silent, it is clinically important to develop an accurate prediction model for the risk of CKD progression. Such a model is expected to facilitate physicians in making personalized treatment decisions, thereby improving the overall prognosis. Various statistical models have been developed to predict the risk of ESKD based on variables such as age, sex, blood pressure, comorbidities, laboratory data, and most commonly, the estimated glomerular filtration rate (eGFR) and proteinuria level (8). Among them, the most popular statistical model is the four-variable kidney failure risk equation (KFRE) based on age, sex, eGFR, and urine albumin-to-creatine ratio (9). Further, an eight-variable equation based on KFRE (which further included serum albumin, bicarbonate, calcium, and phosphate levels) was proposed to provide a more accurate prediction (10). The traditional statistical methods are often based on predefined hypotheses and assumptions. Researchers formulate hypotheses and test them using statistical methods. Moreover, these methods often focus on drawing inferences and conclusions about a population based on a sample. These methods aim to provide insights into causal relationships and generalizability (11). However, the traditional statistical methods have several limitations in effectively dealing with the challenges posed by big data and complex data structures.

Machine learning (ML) methods excel in analyzing unstructured data and complex patterns, whereas traditional statistical methods often require human intervention and expertise in model selection, hypothesis formulation, and result interpretation (11). In the big data era, several applications of ML, which is a subset of artificial intelligence (AI), have emerged in health informatics (12), allowing computers to perform a specific task without direct instruction (13). In contrast to theory-driven formula that requires a predefined hypothesis based on prior knowledge, ML models typically follow a data-driven approach that allows the model to learn from experience alone (14). Therefore, compared with the traditional statistical methods, ML models may demonstrate better performance in predicting a determined outcome, as they have no strict assumptions when modeling (15–21). The utilization of ML algorithms in CKD is a promising research topic that aims at assisting healthcare professionals in diagnosing and managing patients with CKD via computer-aided decision support systems for the early identification of critical events such as ESKD or eGFR doubling (22–25).

To date, only a few studies have used ML methods in CKD populations for identifying metabolomic signatures of pediatric CKD etiology (26), developing a lifestyle scoring system for CKD prevention (27), using retinal images and clinical metadata to predict eGFR and CKD stage (28), and predicting incident ESKD in patients with diabetes mellitus (DM) and CKD (29). However, previous ML prediction models have primarily integrated baseline laboratory data and clinical information, and many studies have focused on predicting ESKD rather than eGFR changes (25). Furthermore, studies on CKD-related risk factor screening or analysis have predominantly relied on a single model without considering hybrid approaches, especially when employing ML methods (26, 27, 29, 30). According to a previous study, the mean annual eGFR decline in healthy individuals was estimated to be 0.97 ± 0.02 mL/min/1.73 m^2 (31). However, even for individuals with the same underlying comorbidities or extent of kidney function impairment, the eGFR decline could be highly variable. Early identification and management of patients with CKD based on longitudinal biochemical data are essential.

Therefore, this study aimed to identify significant factors that influence the prediction of eGFR changes using the ML and important variable selection (ML&IVS) scheme in a CKD cohort with longitudinal laboratory data. In the context of longitudinal data analysis, it is crucial to prioritize the examination of the rate and variation of biochemical data rather than solely relying on baseline data. This study employed five effective

ML algorithms, namely, stochastic gradient boosting (SGB), multivariate adaptive regression splines (MARS), random forest (RF), eXtreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM). Using these algorithms, we developed an integrated multistage ML-based predictive scheme for all four subgroups according to eGFR and presence of DM to predict eGFR changes and subsequently evaluate and integrate relatively important risk factors.

2. Materials and methods

2.1. Dataset

This retrospective study included 710 patients with nondialysis-dependent CKD who were recruited from outpatient nephrology clinics of the Shin-Kong Wu Ho-Su Memorial Hospital Medical Center for a prospective cohort study from 2016 to 2021. The inclusion criteria were as follows: patients aged $\geq$20 years, those who sustained ($\geq$3 months) a decrease in eGFR of $\leq$60 mL/min/1.73 m^2 based on the four-variable modification of diet in renal disease study equation (32), and those who were regularly followed up in our CKD multidisciplinary care program (33). Patients who did not visit the nephrological outpatient department for $\geq$4 months and had incomplete data were excluded. In our CKD multidisciplinary care program, patients were regularly followed up in the nephrological outpatient department every 3 months.

This study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the Institutional Review Board of the Shin-Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan (IRB no. 20200901R). Informed consent was waived because our study was based on a medical chart review. Patient information was anonymized and de-identified before analysis.

2.2. Definition of longitudinal variables

This study aimed to predict eGFR changes and the corresponding relationship between risk factors in the fourth examination of each patient; thus, the results of the first three examinations of each patient were utilized as independent variables. As the results of the first three examinations were collected from each patient, the independent variables used in this study could be regarded as longitudinal predictor variables.

The definitions of the longitudinal variables used in this study are presented in Table 1. As shown, $V_{i,t}$ represents the t_{th} time examination result of i_{th} variable [e.g., $V_{1,2}$ is the systolic blood pressure (SBP) result of the second examination]. A total of 17 variables were utilized in this study. Moreover, as all patients had previous records of their first three examinations, the variables used in this study could be further defined as Equation (1).

$$V_{i,t}, \ \forall i, \ t \in \mathbb{N} \tag{1}$$

where $1 \leq i \leq 17$; $1 \leq t \leq 3$.

Then, this study utilized three approaches for generating more predictor variables that can be used to construct the ML

TABLE 1 Definition of longitudinal variables.

	Variables	Description
$V_{1,t}$	SBP	SBP in the t_{th} examination
$V_{2,t}$	DBP	DBP in the t_{th} examination
$V_{3,t}$	Height	Height in the t_{th} examination
$V_{4,t}$	Weight	Weight in the t_{th} examination
$V_{5,t}$	HB	HB in the t_{th} examination
$V_{6,t}$	BUN	BUN in the t_{th} examination
$V_{7,t}$	UA	UA in the t_{th} examination
$V_{8,t}$	Na	Na in the t_{th} examination
$V_{9,t}$	K	K in the t_{th} examination
$V_{10,t}$	iCa	iCa in the t_{th} examination
$V_{11,t}$	P	P in the t_{th} examination
$V_{12,t}$	Albumin	Albumin in the t_{th} examination
$V_{13,t}$	TG	TG in the t_{th} examination
$V_{14,t}$	LDL	LDL in the t_{th} examination
$V_{15,t}$	AC Sugar	AC sugar in the t_{th} examination
$V_{16,t}$	UP	UP in the t_{th} examination
$V_{17,t}$	eGFR	eGFR value in the t_{th} examination
Y	eGFR in the fourth examination	eGFR value in the fourth examination

AC, before meals; BUN, blood urea nitrogen; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; Hb, hemoglobin; iCa, ionized calcium; LDL, low-density lipoprotein; P, phosphate; SBP, systolic blood pressure; TG, triglycerides; UA, uric acid; UP, urine protein.

predictive models. These three approaches, namely, "closest," "mean," and "standard deviation," could provide various data from the independent variables. The "closest" approach generated the predictor variable using the result of the latest examination, which is the third examination result ($V_{i,3}$) in this study. The predictor variable generated using the "closest" approach ($V_i C$) was defined as Equation (2). For example, $V_1 C$ is the SBP result of the third examination ($V_{1,3}$) and can be written as SBP(C).

The "mean" approach generated the predictor variable by calculating the mean of the results of the first three examinations ($V_{i,1}$, $V_{i,2}$, $V_{i,3}$). The predictor variable generated using the "mean" approach ($V_i M$) was defined as Equation (3). For demonstration, $V_1 M$ was constructed by calculating the mean of the results of the first ($V_{1,1}$), second ($V_{1,2}$), and third ($V_{1,3}$) SBP examinations. $V_1 M$ can also be written as SBP(M).

Similar to the concept of the "mean" approach, the "standard deviation" approach generated the predictor variable ($V_i S$) by calculating the standard deviation of the results of the first three examinations ($V_{i,1}$, $V_{i,2}$, $V_{i,3}$), as shown in Equation (4). For example, $V_1 S$ is the standard deviation of the results of the first three SBP examinations ($V_{1,1}$, $V_{1,2}$, $V_{1,3}$) and can be written as SBP(S).

$$V_i C = V_{i,3} \tag{2}$$

$$V_i M = Mean(V_{i,1}, V_{i,2}, V_{i,3}) \tag{3}$$

$$V_i S = Std(V_{i,1}, V_{i,2}, V_{i,3}) \tag{4}$$

The above mentioned 3 approaches were utilized for all 17 independent variables to generate the overall predictor variables used in this study. In addition, as height and weight do not change drastically between examinations, this study only used the "closest" approach for these variables. Thus, 47 predictor variables were generated and utilized for constructing the ML predictive models. The descriptive statistical data of 47 predictor variables and the target variable (eGFR of the fourth examination) are presented in Table 2.

2.3. Proposed scheme

This study proposed an integrated multistage predictive scheme known as ML&IVS based on five ML methods, including SGB, MARS, RF, XGBoost, and LightGBM, for predicting eGFR changes and subsequently identifying important risk factors. The five ML algorithms have been successfully used in various medical/healthcare applications (16, 19, 21, 23, 24, 34–37). The overall process of ML&IVS scheme is shown in Figure 1.

In this process, first, 4-year health examination data of patients with CKD were collected; the dataset is described in Section "2.1. Dataset." Second, 47 longitudinal risk variables were defined, and the patients were categorized into five subgroups: Group 1 (patients with CKD stage 3 and DM), Group 2 (patients with CKD stage 3 without DM), Group 3 (patients with CKD stages 4–5 and DM), Group 4 (patients with CKD stages 4 and 5 without DM), and Group 5 (overall patients). The definitions of variables are presented in Section "2.2. Definition of longitudinal variables."

Third, RF, SGB, MARS, XGBoost, and LightGBM were used to construct predictive models for all five groups. As multiple linear regression (MLR) is a commonly used statistical method in medical/healthcare applications, it was used in this study as a bench method. RF is an ensemble learning decision tree algorithm that combines bootstrap resampling and bagging (38). The RF principle entails random generation of many different and unpruned classification and regression tree (CART) decision trees, in which the decrease in Gini impurity is regarded as the splitting criterion, and all generated trees are combined into a forest. Then, all trees in the forest are averaged or voted to generate output probabilities and construct the final robust model. SGB is a tree-based gradient boosting learning algorithm that combines both bagging and boosting techniques to minimize the loss function and thereby solve the overfitting problem of the traditional decision trees (39, 40). SGB comprises many stochastic weak learners of decision trees that are sequentially generated through multiple iterations, in which each tree focuses on correcting or explaining the errors of the tree generated in the previous iteration. In other words, the residual of the previous iteration tree is used as the input for the newly generated tree. This iterative process is repeated until a convergence condition or stopping criterion is reached for the maximum number of iterations. Finally, the cumulative results of several trees are used to construct the final robust model. MARS is a nonlinear spline regression method and nonparametric form of the regression analysis algorithm (41). MARS uses multiple piecewise linear segments (splines) with different gradients. It considers each sample as a knot and divides it into several sections for the successive linear regression analysis of data within each section. For

TABLE 2 Descriptive statistical data of 47 predictor variables and the target variable.

Predictor variables	Abb. Name	V_iC (Mean ± SD)	V_iM (Mean ± SD)	V_iS (Mean ± SD)
V_1	SBP (mmHg)	133.46 ± 16.1	134.93 ± 14.25	10.43 ± 7.65
V_2	DBP (mmHg)	71.26 ± 9.98	71.88 ± 8.64	7.18 ± 5.36
V_3	Height (cm)	159.99 ± 8.99	–	–
V_4	Weight (kg)	67.63 ± 13.29	–	–
V_5	Hb (g/dL)	11.88 ± 2.02	11.91 ± 1.93	0.58 ± 0.47
V_6	BUN (mg/dL)	36.12 ± 17.9	36.01 ± 15.41	6.18 ± 5.62
V_7	UA (mg/dL)	5.91 ± 1.86	6.18 ± 1.46	1.23 ± 1.07
V_8	Na (MEq/L)	140 ± 3.13	139.83 ± 2.63	1.66 ± 1.38
V_9	K (MEq/L)	4.51 ± 0.54	4.52 ± 0.47	0.3 ± 0.21
V_{10}	iCa (mg/dL)	4.67 ± 0.27	4.67 ± 0.22	0.15 ± 0.11
V_{11}	P (mg/dL)	3.78 ± 0.72	3.79 ± 0.61	0.36 ± 0.26
V_{12}	Albumin (g/dL)	4.08 ± 0.4	4.07 ± 0.38	0.16 ± 0.12
V_{13}	TG (mg/dL)	156.87 ± 105.69	162.04 ± 101.39	42.34 ± 66.47
V_{14}	LDL (mg/dL)	85.99 ± 30.04	88.1 ± 24.2	15.7 ± 14.71
V_{15}	AC sugar (mg/dL)	113.34 ± 37.74	114.63 ± 32.29	16.92 ± 22.7
V_{16}	UPCR (mg/mg)	1,192.38 ± 1,936.41	1,236.66 ± 1,980.14	415.84 ± 732.27
V_{17}	eGFR (mL/min/1.73 m^2)	31.93 ± 12.94	31.85 ± 11.65	3.49 ± 3.05
Target variable (Y)		N		Mean ± SD
eGFR in the fourth examination		710		31.59 ± 13.29

V_iC, closest value; V_iM, mean value; V_iS, standard deviation; AC, before meals; BUN, blood urea nitrogen; DBP: diastolic blood pressure; eGFR, estimated glomerular filtration rate; Hb, hemoglobin; iCa, ionized calcium; LDL, low-density lipoprotein; P, phosphate; SBP, systolic blood pressure; TG, triglycerides; UA, uric acid; UPCR, urine protein-to-creatinine ratio.
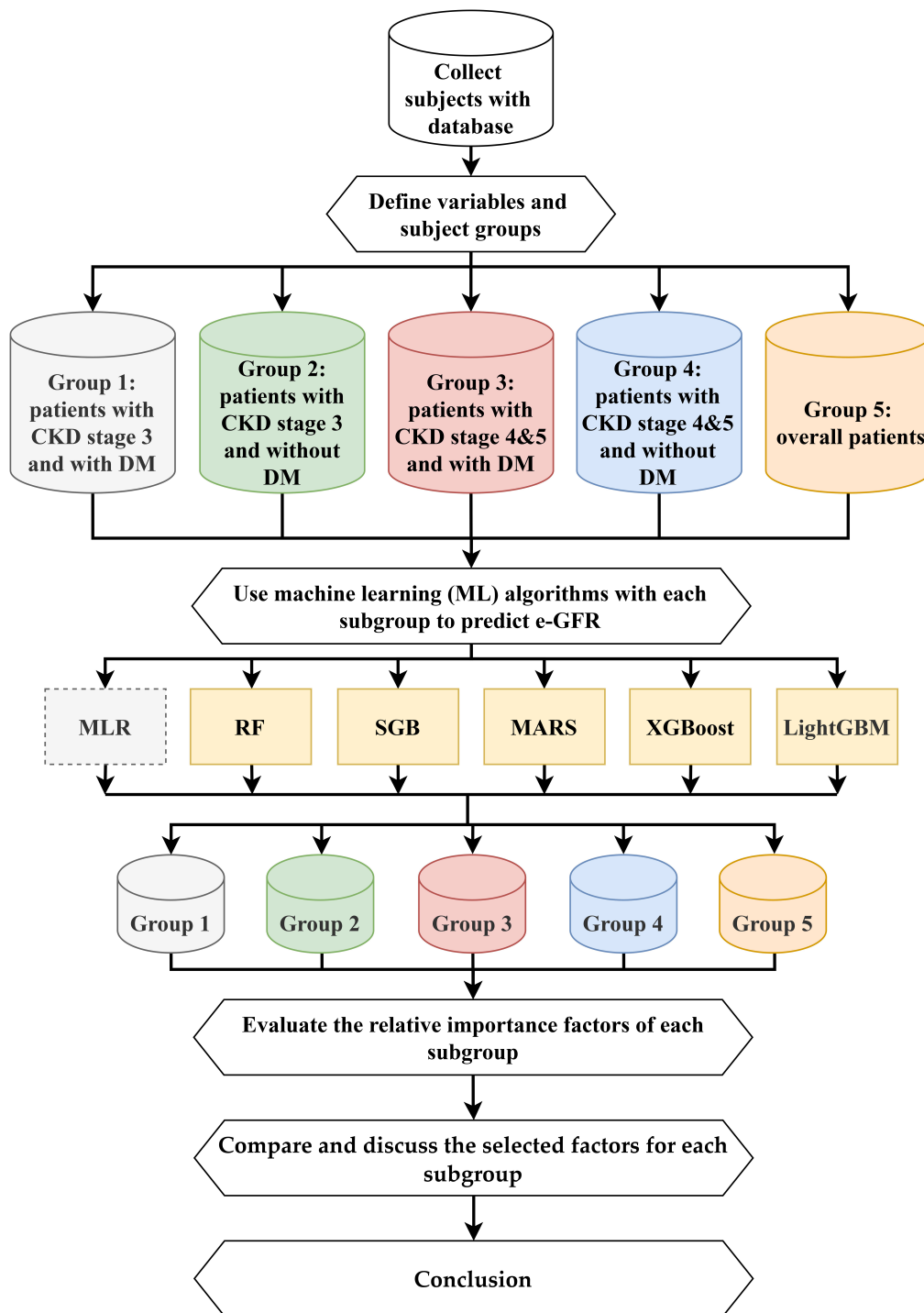
FIGURE 1
Proposed machine learning predictive and important variable selection scheme (ML&IVS).

determining knots, a forward algorithm is used to select all possible basic functions and their corresponding knots, and a backward algorithm is used to eliminate all basic functions to generate the best combinations of existing knots.

XGBoost is a gradient boosting technology based on an SGB-optimized extension (42). It trains many weak models sequentially to ensemble them using the gradient boosting method of outputs, which can achieve a promising prediction performance. In XGBoost, Taylor binomial expansion is used to approximate the objective function and arbitrary differentiable loss functions, thereby accelerating the convergence process of model construction. Then, XGBoost employs a regularized boosting technique to penalize the complexity of the model and correct overfitting, thus increasing the accuracy of the model. LightGBM is a decision tree-based distributed gradient boosting framework that uses the decision tree technique with improved histograms.

To improve the accuracy of tree models, LightGBM uses gradient-based one-sided sampling and a unique feature bundling algorithm to fit a loss function with negative gradients and learn the residual approximation of decision trees in iterations (43).

Fourth, after constructing effective ML models for all five groups, the relative importance values of each risk factor can be obtained via ML algorithms. In this study, the most important risk factor had an importance value of 100, whereas the least important risk factor had an importance value of 0. Values can be repeated, i.e., two or more variables can have similar variable importance. Because different ML algorithms have different calculus architectures and selection features, the five ML algorithms in this study may generate different variable importance values for a single risk factor. Within the same group, a single, robust, and complete variable importance value can be generated for each risk factor to facilitate subsequent comparison of variable rankings and identification of important risk factors. A single pooled value of variable importance was generated based on the average value of variable importance derived from the five ML models.

Fifth, the important variables among groups were compared, and their differences were discussed. Sixth, the conclusions of this study were presented based on the abovementioned findings.

For constructing each model, nested cross-validation (Nest-CV) was utilized. Nest-CV is a variation of the CV family. Under the structure of Nest-CV, two loops are required (inner and outer loops). The inner loop is used for hyperparameter tuning (which is equal to k-fold CV), whereas the outer loop is used for model evaluation, with the optimal hyperparameter set in the inner loop. Nest-CV is commonly utilized, and several studies have demonstrated that Nest-CV can overcome the problem of overfitting effectively (44–47).

During model construction under Nest-CV structure, the outer loop first randomly splits the dataset into several folds (10 folds in this study). Then, for each iteration, one fold of the data from the outer loop is used for testing, whereas the remaining folds are used for training. Next, the training data from the inner loop are used for hyperparameter tuning with 10-fold CV. In the 10-fold CV approach, the training dataset was further randomly divided into 10 folds, wherein 9 folds were used to construct the model with a different set of hyperparameters, and 1 fold was used for validation. All 10 folds were used for validation at least once, and all possible combinations of the hyperparameters were investigated using a grid search. The hyperparameter set during validation with the lowest root mean square error (RMSE) was regarded as the most optimized set. After determining the set and training the model using the set, the test data from the outer loop were used for evaluation, which involves the completion of one iteration. Entire Nest-CV was completed when each fold of the data from the outer loop was used for testing at least once. After constructing the ML models, the variable importance values can be extracted from these models. As 10-fold Nest-CV is utilized, the extracted variable importance ranking from each ML model will have 10 scores for each variable. Therefore, to obtain the final corresponding variable importance values, the approach of averaging the importance scores for each variable was employed.

As the target variable of this study (changes in eGFR) for model construction is a numerical variable, the metrics used for model performance comparison included symmetric mean absolute percentage error (SMAPE), mean absolute percentage

error (MAPE), relative absolute error (RAE), root relative squared error (RRSE), and RMSE (Table 3). In addition, "R" software version 3.6.2 and "RStudio" version 1.1.453 were used for model construction. The related "R" packages utilized for constructing RF, SGB, MARS, XGBoost, and LightGBM included "randomForest" package version 4.7-1.1 (48), "gbm" package version 2.1.8 (49), "earth" package version 5.3.1 (50), "XGBoost" package version 1.5.0.2 (51), and "lightgbm" package version 3.3.2, respectively (52). The "caret" package version 6.0-92 (53) was used in all algorithms to estimate the optimal hyperparameters for constructing the best prediction model.

3. Empirical study

The dataset was divided into five subgroups, i.e., Group 1 (patients with CKD stage 3 and DM), Group 2 (patients with CKD stage 3 without DM), Group 3 (patients with CKD stages 4–5 and DM), Group 4 (patients with CKD stages 4 and 5 without DM), and Group 5 (overall patients). All groups followed the same modeling process for predicting eGFR in the fourth examination. The performance of the methods used in all five groups is presented in Table 4.

As shown in Table 4, the error of the metrics and their corresponding standard deviation (SD) are presented; percentage values after n in the first column indicate the corresponding proportion of the group. In Group 1, ML methods showed better performance than the MLR method. Among these ML methods, SGB showed the best performance with the following values: SMAPE, 0.097; MAPE, 0.098; RAE, 0.672; and RMSE, 4.922. RF had similar RRSE as SGB but lower SD than SGB.

Similarly, in Group 2, ML methods outperformed the MLR method. Furthermore, SGB showed the best performance in this subgroup with the following values: SMAPE, 0.102; MAPE, 0.104; RAE, 0.682; RRSE, 0.713; and RMSE, 5.259. In Group 3, RF showed the best performance with the following values: SMAPE, 0.135; MAPE, 0.147; RAE, 0.383; RRSE, 0.432; and RMSE, 3.009.

In Group 4, RF showed the best performance with the following values: SMAPE, 0.159; MAPE, 0.178; RAE, 0.406; RRSE, 0.464; and RMSE, 3.333. Finally, in Group 5, the CKD stage 3–5 dataset was considered the full dataset without subgrouping. ML methods showed better performance than the MLR method. Among them,

TABLE 3 Equations of performance metrics.

Metrics	Description	Calculation
SMAPE	Symmetric mean absolute percentage error	$SMAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{\|y_i - \hat{y}_i\|}{(\|y_i\| + \|\hat{y}_i\|)/2} \times 100$
MAPE	Mean absolute percentage error	$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{y_i - \hat{y}_i}{y_i} \right\| \times 100$
RAE	Relative absolute error	$RAE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i)^2}}$
RRSE	Root relative squared error	$RRSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}$
RMSE	Root mean squared error	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$

where $\hat{y}_i$ and y_i represent predicted and actual values, respectively; n, number of instances.

TABLE 4 Performance of the MLR and five ML methods in all five groups.

Subgroup, total $N = 710$	Methods	SMAPE	MAPE	RAE	RRSE	RMSE
Group 1 (patients with CKD stage 3 and DM) $n = 200$ (28.17%)	MLR	0.116 (0.01)	0.118 (0.01)	0.799 (0.08)	0.821 (0.09)	5.886 (0.71)
	RF	0.098 (0.01)	0.099 (0.01)	0.680 (0.06)	**0.688 (0.05)**	4.926 (0.37)
	SGB	**0.097 (0.01)**	**0.098 (0.01)**	**0.672 (0.07)**	0.688 (0.06)	**4.922 (0.35)**
	MARS	0.099 (0.01)	0.100 (0.01)	0.687 (0.09)	0.715 (0.07)	5.121 (0.38)
	XGBoost	0.099 (0.01)	0.100 (0.01)	0.691 (0.07)	0.708 (0.05)	5.066 (0.33)
	LightGBM	0.099 (0.01)	0.100 (0.01)	0.687 (0.06)	0.702 (0.05)	5.028 (0.39)
Group 2 (patients with CKD stage 3 without DM) $n = 200$ (28.17%)	MLR	0.126 (0.02)	0.126 (0.02)	0.816 (0.13)	0.831 (0.11)	6.132 (0.52)
	RF	0.103 (0.01)	0.105 (0.01)	0.686 (0.08)	0.717 (0.08)	5.294 (0.35)
	SGB	**0.102 (0.01)**	**0.104 (0.01)**	**0.682 (0.09)**	**0.713 (0.09)**	**5.259 (0.45)**
	MARS	0.109 (0.01)	0.112 (0.01)	0.724 (0.10)	0.768 (0.09)	5.676 (0.40)
	XGBoost	0.106 (0.01)	0.107 (0.01)	0.707 (0.09)	0.741 (0.09)	5.478 (0.44)
	LightGBM	0.110 (0.01)	0.112 (0.01)	0.730 (0.07)	0.740 (0.08)	5.463 (0.28)
Group 3 (patients with CKD stages 4–5 and DM) $n = 185$ (26.06%)	MLR	0.217 (0.03)	0.238 (0.05)	0.625 (0.05)	0.694 (0.03)	4.838 (0.35)
	RF	**0.135 (0.03)**	**0.147 (0.04)**	**0.383 (0.07)**	**0.432 (0.08)**	**3.009 (0.62)**
	SGB	0.141 (0.03)	0.155 (0.04)	0.397 (0.06)	0.438 (0.08)	3.056 (0.60)
	MARS	0.145 (0.02)	0.152 (0.03)	0.433 (0.09)	0.507 (0.11)	3.519 (0.70)
	XGBoost	0.147 (0.02)	0.161 (0.03)	0.411 (0.05)	0.457 (0.07)	3.191 (0.55)
	LightGBM	0.146 (0.03)	0.163 (0.04)	0.407 (0.06)	0.448 (0.08)	3.124 (0.62)
Group 4 (patients with CKD stages 4–5 without DM) $n = 125$ (17.61%)	MLR	0.280 (0.04)	0.296 (0.05)	0.634 (0.10)	0.707 (0.09)	5.068 (0.43)
	RF	**0.159 (0.02)**	**0.178 (0.03)**	**0.406 (0.05)**	**0.464 (0.07)**	**3.333 (0.51)**
	SGB	0.173 (0.03)	0.200 (0.03)	0.427 (0.07)	0.484 (0.08)	3.472 (0.57)
	MARS	0.197 (0.04)	0.200 (0.03)	0.482 (0.08)	0.557 (0.08)	3.997 (0.57)
	XGBoost	0.171 (0.03)	0.193 (0.04)	0.435 (0.07)	0.483 (0.07)	3.483 (0.59)
	LightGBM	0.175 (0.03)	0.206 (0.05)	0.436 (0.07)	0.486 (0.08)	3.495 (0.59)
Group 5 (overall patients) $n = 710$ (100%)	MLR	0.149 (0.01)	0.156 (0.01)	0.369 (0.04)	0.414 (0.04)	5.507 (0.28)
	RF	**0.138 (0.01)**	0.150 (0.01)	0.354 (0.04)	0.403 (0.04)	5.354 (0.33)
	SGB	0.145 (0.01)	0.160 (0.01)	0.362 (0.04)	0.406 (0.04)	5.397 (0.29)
	MARS	0.139 (0.01)	**0.147 (0.01)**	**0.350 (0.04)**	**0.397 (0.03)**	**5.281 (0.27)**
	XGBoost	0.144 (0.01)	0.157 (0.01)	0.361 (0.04)	0.405 (0.03)	5.394 (0.26)
	LightGBM	0.143 (0.01)	0.157 (0.01)	0.366 (0.04)	0.411 (0.04)	5.463 (0.28)

Bold values indicates the best performance in the subgroup.

MARS showed the best performance with the following values: MAPE, 0.147; RAE, 0.350; RRSE, 0.397; and RMSE, 5.281.

Overall, the performance of the five ML methods used in each subgroup had low SD, indicating that the ML usage in this study was reasonable and robust. Furthermore, different ML methods have different mechanisms for generating data regarding various risk factors, and each method visualizes data with a distinct perspective. In other words, ML methods can generate valuable information for supporting decision making with various perspectives; thus, the information generated from all ML methods can be considered. To accurately rank the risk factors using ML methods in different subgroups, this study revealed the top 10 risk factors in each subgroup (Figure 2).

As shown in Figure 2, in Group 1, risk factors were ranked by the five ML methods using the importance scores. Then, we calculated the average importance score of each risk factor to determine the top 10 important risk factors. Thus, in Group 1, the rank 1 risk factor was $V_{17}C$, rank 2 was $V_{17}M$, and rank 3 was $V_{17}S$. Similarly, in Groups 2–5, the top 10 important risk factors were calculated. Further, to easily compare the important risk factors between subgroups, the results from Figure 2 were further displayed in Table 5 and the distribution plots of the top 10 important features of each subgroups are shown in Supplementary Figures 1–5.

Table 5 presents the differences in the top 10 important risk factors among all 5 groups. eGFR(C) and eGFR (M) were the first (rank 1) and second (rank 2) important risk factors, respectively, in all groups. Interestingly, BUN(C) was the third important risk factor in all groups, except for Group 1. In general, the ranking of risk factors in each subgroup was different. For example, UA(C) only appeared in Group 1 and K(S) only appeared in Group 2.
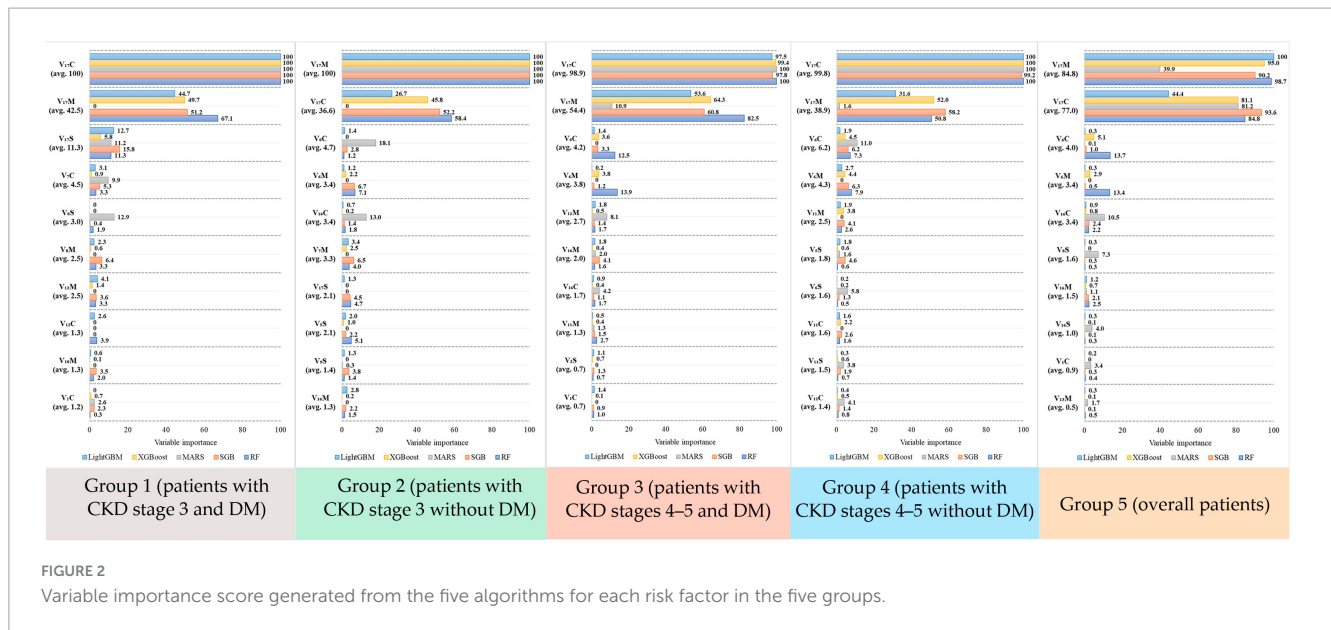
FIGURE 2
Variable importance score generated from the five algorithms for each risk factor in the five groups.

TABLE 5 Ranking of the top 10 important variables of the five groups.

Subgroup	Group 1 (patients with CKD stage 3 and DM)	Group 2 (patients with CKD stage 3 without DM)	Group 3 (patients with CKD stages 4–5 and DM)	Group 4 (patients with CKD stages 4–5 without DM)	Group 5 (overall patients)
Rank	Variable	Variable	Variable	Variable	Variable
1	eGFR(C)	eGFR(M)	eGFR(C)	eGFR(C)	eGFR(M)
2	eGFR(M)	eGFR(C)	eGFR(M)	eGFR(M)	eGFR(C)
3	eGFR(S)	BUN(C)	BUN(C)	BUN(C)	BUN(C)
4	UA(C)	BUN(M)	BUN(M)	BUN(M)	BUN(M)
5	Na(S)	UPCR(C)	Albumin(M)	P(M)	UPCR(C)
6	Na(M)	UA(M)	UPCR(M)	Hb(S)	Na(C)
7	Albumin(M)	eGFR(S)	UPCR(C)	BUN(S)	UPCR(M)
8	Albumin(C)	Hb(S)	P(M)	P(C)	UPCR(S)
9	iCa(M)	K(S)	DBP(S)	P(S)	SBP(C)
10	SBP(C)	UPCR(M)	SBP(C)	Albumin(C)	TG(M)

eGFR(C): $V_{17}C$; eGFR(M): $V_{17}M$; eGFR(S): $V_{17}S$; BUN(C): V_6C; BUN(M): V_6M; BUN(S): V_6S; UA(C): V_7C; UA(M): V_7M; Na(C): V_8C; Na(M): V_8M; Na(S): V_8S; UP(C): $V_{16}C$; UP(M): $V_{16}M$; UP(S): $V_{16}S$; Albumin(C): $V_{12}C$; Albumin(M): $V_{12}M$; P(C): $V_{11}C$; P(M): $V_{11}M$; P(S): $V_{11}S$; HB(S): V_5S; iCa(M): $V_{10}M$; K(S): V_9S; SBP(C): V_1C; SBP(S): V_1S; DBP(S): V_2S; TG(M): V_{13}.

4. Discussion

This study utilized longitudinal electronic health records to identify the risk factors associated with the prediction of eGFR changes in different CKD groups. This analysis considered the current values, mean values, and variation of biochemical data. Our proposed ML&IVS scheme yielded valuable results that can aid clinicians in effectively managing CKD progression and providing preventive measures in the CKD multidisciplinary care program. In addition to identifying important factors in CKD progression, our ML scheme, with some necessary modifications, can be seamlessly integrated into an electronic system, enabling the early identification of CKD progression. To the best of our knowledge, this study represents the first attempt to utilize ML methods to predict short-term eGFR changes in patients with CKD. The innovative application of ML in this context holds great potential for advancing the field and improving patient care.

ML has demonstrated promising performance in the nephrological field, including kidney function prediction via ultrasonography [54], acute kidney injury prediction in critical care [55, 56], specific pattern identification on renal pathology slides [57, 58], optimal dialysis prescription [59, 60], calculation of further eGFRs [61], mortality risk prediction in patients undergoing dialysis [62], and ESKD prediction based on clinical data [63–65]. In this study, five ML methods were adopted to obtain the 10 most important factors for predicting eGFR changes in different CKD groups. The latest eGFR and mean eGFR were the two most important factors among all subgroups. Moreover, the SD of eGFR was important in all subgroups. Under the assumption of the stochastic process, the latest value is always important to predict the next value. The mean value represents long-term

trends, whereas the SD indicates the severity of fluctuation in the short term.

In this study, we employed the widely used Pearson correlation method to identify the top 10 factors with the highest correlation coefficients within each group (Supplementary Table 1). However, these high correlation coefficients appeared to exhibit less variability between groups. This raised concerns about their ability to provide meaningful distinction between the groups. Nevertheless, promising results were obtained through our proposed ML&IVS scheme. We revealed that the crucial factors identified via ML&IVS scheme displayed greater differences between groups than those identified solely through the Pearson correlation method. This finding indicated the effectiveness of our scheme in providing more informative and discriminative features that can potentially improve the understanding and characterization of different groups in this study.

Numerous studies have attempted to identify the risk factors associated with clinical CKD progression (66), and most of them have adopted traditional statistical methods under a theory-driven assumption. In this study, we used ML methods to determine important laboratory factors for short-term eGFR prediction. In all patients with CKD stages 3–5, the latest and mean phosphate levels, SBP variation, latest albumin levels, hemoglobin variation, and mean uric acid levels were important factors in addition to the indicators of current kidney function (eGFR and BUN). Such findings were consistent with those of previous studies reporting that serum phosphate (67), blood pressure (68), lower serum albumin (69), hemoglobin (70), and uric acid (71) levels were associated with CKD progression.

DM is the leading cause of CKD, and approximately 20–30% of patients with type 2 DM have moderate-to-severe CKD (eGFR of <60 mL/min/1.73 m^2). Such kidney damage is associated with the accumulation of uremic toxins, inflammatory factors, and oxidative stress (72). Therefore, patients with DM have different risk profiles of CKD progression compared with those without DM (73); moreover, proteinuria is usually observed in patients with DM (74) and is well-known to be robustly associated with CKD progression (75). However, in our study, proteinuria was an important factor only in patients with moderate CKD without DM and those with advanced CKD with DM. Such discrepancy may be due to our prediction of short-term eGFR changes. The serum albumin level might be an alternative to proteinuria severity for predicting CKD stage 3 with DM.

The role of uric acid in CKD progression remains controversial. Our results revealed that serum uric acid is an important factor for predicting CKD stage 3 with/without DM. High uric acid levels may cause glomerular injury, tubulointerstitial fibrosis, atherosclerosis, and vascular injuries (76). Some observational studies have revealed that serum uric acid level is associated with renal function impairment (77–79). Moreover, a recent study reported that even uric acid levels under therapeutic criteria may increase the risk of CKD stage transition (74). However, three randomized controlled trials have reported that lowering uric acid levels is not beneficial in preserving renal function (80–82). This discrepancy is attributed to hyperuricemia resulting from renal function impairment. Therefore, asymptotic hyperuricemia does not require medical treatment but needs lifestyle modification.

Phosphorus is an important mineral for maintaining cell structure and energy. It is mainly found intracellularly (70%). Approximately 29% of phosphorus resides in the bones and <1% circulates in the serum (83). In CKD, kidneys cannot excrete phosphorus, resulting in hyperphosphatemia and consequent renal osteodystrophy (84), which was more significant in advanced CKD. A meta-analysis reported an independent association between serum phosphorus level and kidney failure in patients with nondialysis-dependent CKD (67), revealing that higher phosphorus levels might lead to a steeper decline in renal function. Moreover, fibroblast growth factor 23 (FGF23) was secreted by osteocytes in bones, and its levels increased at an early stage (starting at an eGFR of <90 mL/min/1.73 m^2). FGF23 participates in serum phosphate hemostasis by decreasing phosphorus absorption from the alimentary tract (85). Therefore, hyperphosphatemia may increase FGF23 levels, leading to anemia, cardiovascular disease, and eventually death (86, 87), resulting in a poor renal progression.

This study has some limitations. The dataset used in this study was from a single medical center, which may limit the generalizability of our findings. Therefore, federated learning, which refers to collaborative ML without centralized training data, is necessary to include more data from multiple centers in future studies (88). In addition, the prescribed medications were not collected. Some medicines may have affected the renal function progression. Finally, this cohort consisted of approximately 700 patients, which might have affected the model performance. Next, this study has some advantages. This study used longitudinal data to predict eGFR changes, which can provide a causal inference. The proposed ML&IVS scheme, which integrated the results of risk factor identification and information from five well-known ML methods, could provide more robust and useful information to support our results.

5. Conclusion

When longitudinal data are used to predict eGFR changes in patients with CKD with or without DM, the proposed ML&IVS scheme can effectively integrate the most significant risk factors from each model, resulting in more robust and comprehensive identification of important risk factors for predicting eGFR changes. It is crucial to increase awareness regarding eGFR changes, particularly in government health-promotion initiatives within the multidisciplinary CKD care program. This study provides valuable insights for initiating further discussions and follow-up research in this field. These findings contribute to our understanding of the factors influencing eGFR changes and can guide future investigations on the early identification and management of CKD progression.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: data are available on request to the first author due to privacy/ethical restrictions. Requests to access these datasets should be directed to M-HT, chaomsmyth.tw@gmail.com.

Author contributions

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1155426/full#supplementary-material

References

1. Centers for Disease Control and Prevention [CDC]. Prevalence of chronic kidney disease and associated risk factors–United States, 1999–2004. *MMWR Morb Mortal Wkly Rep.* (2007) 56:161–5.

2. Hwang S, Tsai J, Chen H. Epidemiology, impact and preventive care of chronic kidney disease in Taiwan. *Nephrology.* (2010) 15:3–9. doi: 10.1111/j.1440-1797.2010.01304.x

3. Tsai M, Hsu C, Lin M, Yen M, Chen H, Chiu Y, et al. Incidence, prevalence, and duration of chronic kidney disease in Taiwan: results from a community-based screening program of 106,094 individuals. *Nephron.* (2018) 140:175–84. doi: 10.1159/000491708

4. Saran R, Robinson B, Abbott K, Agodoa L, Bhave N, Bragg-Gresham J, et al. US renal data system 2017 annual data report: epidemiology of kidney disease in the United States. *Am J Kidney Dis.* (2018) 71:A7. doi: 10.1053/j.ajkd.2018.01.002

5. Lai T, Hsu C, Lin M, Wu V, Chen Y. Trends in the incidence and prevalence of end-stage kidney disease requiring dialysis in Taiwan: 2010–2018. *J Formos Med Assoc.* (2022) 121:S5–11. doi: 10.1016/j.jfma.2021.12.013

6. National Kidney Foundation [Nkf]. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Am J Kidney Dis.* (2002) 39:S1–266.

7. Stevens P, Levin A. Kidney disease: improving global outcomes chronic kidney disease guideline development work group members. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. *Ann Intern Med.* (2013) 158:825–30. doi: 10.7326/0003-4819-158-11-201306040-00007

8. Ramspek C, de Jong Y, Dekker F, van Diepen M. Towards the best kidney failure prediction tool: a systematic review and selection aid. *Nephrol Dial Transplant.* (2020) 35:1527–38. doi: 10.1093/ndt/gfz018

9. Tangri N, Stevens L, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. *JAMA.* (2011) 305:1553–9. doi: 10.1001/jama.2011.451

10. Tangri N, Grams M, Levey A, Coresh J, Appel L, Astor B, et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis. *JAMA.* (2016) 315:164–74. doi: 10.1001/jama.2015.18202

11. Rajula H, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis,

drug development, and treatment. *Medicina (Kaunas).* (2020) 56:455. doi: 10.3390/medicina56090455

12. El Aboudi N, Benhlima L. Big data management for healthcare systems: architecture, requirements, and implementation. *Adv Bioinform.* (2018) 2018:4059018. doi: 10.1155/2018/4059018

13. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Fut Healthc J.* (2019) 6:94–8. doi: 10.7861/futurehosp.6-2-94

14. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. *Minim Invasive Ther Allied Technol.* (2019) 28:73–81. doi: 10.1080/13645706.2019.1575882

15. Liu Y, Chen P, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA.* (2019) 322:1806–16. doi: 10.1001/jama.2019.16489

16. Chang C, Yeh J, Chen Y, Jhou M, Lu C. Clinical predictors of prolonged hospital stay in patients with myasthenia gravis: a study using machine learning algorithms. *J Clin Med.* (2021) 10:4393. doi: 10.3390/jcm10194393

17. Saberi-Karimian M, Khorasanchi Z, Ghazizadeh H, Tayefi M, Saffar S, Ferns G, et al. Potential value and impact of data mining and machine learning in clinical diagnostics. *Crit Rev Clin Lab Sci.* (2021) 58:275–96. doi: 10.1080/10408363.2020.1857681

18. Wu Y, Hu H, Cai J, Chen R, Zuo X, Cheng H, et al. Machine learning for predicting the 3-year risk of incident diabetes in Chinese adults. *Front Public Health.* (2021) 9:626331. doi: 10.3389/fpubh.2021.626331

19. Huang Y, Cheng Y, Jhou M, Chen M, Lu C. Important risk factors in patients with nonvalvular atrial fibrillation taking dabigatran using integrated machine learning scheme-a post hoc analysis. *J Pers Med.* (2022) 12:756. doi: 10.3390/jpm12050756

20. Sun C, Tang Y, Liu T, Lu C. An integrated machine learning scheme for predicting mammographic anomalies in high-risk individuals using questionnaire-based predictors. *Int J Environ Res Public Health.* (2022) 19:9756. doi: 10.3390/ijerph19159756

21. Liao P, Chen M, Jhou M, Chen T, Yang C, Lu C. Integrating health data-driven machine learning algorithms to evaluate risk factors of early stage hypertension at different levels of HDL and LDL cholesterol. *Diagnostics.* (2022) 12:1965. doi: 10.3390/diagnostics12081965

22. Shih C, Lu C, Chen G, Chang C. Risk prediction for early chronic kidney disease: results from an adult health examination program of 19,270 individuals. *Int J Environ Res Public Health.* (2020) 17:4973. doi: 10.3390/ijerph17144973

23. Chiu Y, Jhou M, Lee T, Lu C, Chen M. Health data-driven machine learning algorithms applied to risk indicators assessment for chronic kidney disease. *Risk Manag Healthc Pol.* (2021) 14:4401–12. doi: 10.2147/RMHP.S319405

24. Jhou M, Chen M, Lee T, Yang C, Chiu Y, Lu C-J. A hybrid risk factor evaluation scheme for metabolic syndrome and stage 3 chronic kidney disease based on multiple machine learning techniques. *Healthcare.* (2022) 10:2496. doi: 10.3390/healthcare10122496

25. Lei N, Zhang X, Wei M, Lao B, Xu X, Zhang M, et al. Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis. *BMC Med Inform Decis Mak.* (2022) 22:205. doi: 10.1186/s12911-022-01951-1

26. Cai X, Cheng Y, Ge S, Xu G. Clinical characteristics and risk factor analysis of Pneumocystis jirovecii pneumonia in patients with CKD: a machine learning-based approach. *Eur J Clin Microbiol Infect Dis.* (2023) 42:323–38. doi: 10.1007/s10096-023-04555-3

27. Luo W, Gong L, Chen X, Gao R, Peng B, Wang Y, et al. Lifestyle and chronic kidney disease: a machine learning modeling study. *Front Nutr.* (2022) 9:918576. doi: 10.3389/fnut.2022.918576

28. Zhang K, Liu X, Xu J, Yuan J, Cai W, Chen T, et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nat Biomed Eng.* (2021) 5:533–45. doi: 10.1038/s41551-021-00745-6

29. Zou Y, Zhao L, Zhang J, Wang Y, Wu Y, Ren H, et al. Development and internal validation of machine learning algorithms for end-stage renal disease risk prediction model of people with type 2 diabetes mellitus and diabetic kidney disease. *Ren Fail.* (2022) 44:562–70. doi: 10.1080/0886022X.2022.2056053

30. Zhang X, Fang Y, Zou Z, Hong P, Zhuo Y, Xu Y, et al. Risk factors for progression of CKD with and without diabetes. *J Diabetes Res.* (2022) 2022:9613062. doi: 10.1155/2022/9613062

31. Cohen E, Nardi Y, Krause I, Goldberg E, Milo G, Garty M, et al. A longitudinal assessment of the natural rate of decline in renal function with age. *J Nephrol.* (2014) 27:635–41. doi: 10.1007/s40620-014-0077-9

32. Levey A, Coresh J, Greene T, Marsh J, Stevens L, Kusek J, et al. Expressing the modification of diet in renal disease study equation for estimating glomerular filtration rate with standardized serum creatinine values. *Clin Chem.* (2007) 53:766–72. doi: 10.1373/clinchem.2006.077180

33. Chen P, Lai T, Chen P, Lai C, Yang S, Wu V, et al. Multidisciplinary care program for advanced chronic kidney disease: reduces renal replacement and medical costs. *Am J Med.* (2015) 128:68–76. doi: 10.1016/j.amjmed.2014.07.042

34. Tseng C, Lu C, Chang C, Chen G, Cheewakriangkrai C. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artif Intell Med.* (2017) 78:47–54. doi: 10.1016/j.artmed.2017.06.003

35. Wu T, Chen H, Jhou M, Chen Y, Chang T, Lu C. Evaluating the effect of topical atropine use for myopia control on intraocular pressure by using machine learning. *J Clin Med.* (2020) 10:111. doi: 10.3390/jcm10010111

36. Chang C, Huang T, Shueng P, Chen S, Chen C, Lu C, et al. Developing a stacked ensemble-based classification scheme to predict second primary cancers in head and neck cancer survivors. *Int J Environ Res Public Health.* (2021) 18:12499. doi: 10.3390/ijerph182312499

37. Huang L, Chen F, Jhou M, Kuo C, Wu C, Lu C, et al. Comparing multiple linear regression and machine learning in predicting diabetic urine albumin-creatinine ratio in a 4-year follow-up study. *J Clin Med.* (2022) 11:3661. doi: 10.3390/jcm11133661

38. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

39. Friedman J. Greedy function approximation: a gradient boosting machine. *Ann Stat.* (2001) 29:1189–232.

40. Friedman J. Stochastic gradient boosting. *Comput Stat Data Anal.* (2002) 38:367–78. doi: 10.1016/S0167-9473(01)00065-2

41. Friedman J. Multivariate adaptive regression splines. *Ann Stat.* (1991) 19:1–67. doi: 10.1214/aos/1176347963

42. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* San Francisco, CA: Association of Computing Machinery (ACM) (2016). p. 785–94.

43. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. *Proceedings of the advances in neural information processing systems.* Long Beach, CA: Neural Information Processing Systems (NIPS) (2017). p. 3147–55.

44. Parvandeh S, Yeh H, Paulus M, McKinney B. Consensus features nested cross-validation. *Bioinformatics.* (2020) 36:3093–8.

45. Tsamardinos I, Rakhshani A, Lagani V. Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. *Int J Artificial Intell Tools.* (2015) 24:1540023.

46. Wei L, Owen D, Rosen B, Guo X, Cuneo K, Lawrence T, et al. A deep survival interpretable radiomics model of hepatocellular carcinoma patients. *Phys Med.* (2021) 82:295–305.

47. Vabalas A, Gowen E, Poliakoff E, Casson A. Machine learning algorithm validation with a limited sample size. *PLoS One.* (2019) 14:e0224365. doi: 10.1371/journal.pone.0224365

48. Breiman L, Cutler A. *RandomForest: breiman and cutler's random forests for classification and regression. R Package version, 4.7-1.1.* (2022). Available online at: https://CRAN.R-project.org/package=randomForest (accessed May 25, 2022).

49. Greenwell B, Boehmke B, Cunningham J. *Gbm: generalized boosted regression models. R Package version, 2.1.8.* (2020). Available online at: https://CRAN.R-project.org/package=gbm (accessed May 25, 2022).

50. Milborrow S. *Derived from Mda: MARS by T. Hastie and R. Tibshirani. Earth: multivariate adaptive regression splines. R Package version, 5.3.1.* (2021). Available online at: http://CRAN.R-project.org/package=earth (accessed May 25, 2022).

51. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. *Xgboost: extreme gradient boosting. R Package version, 1.5.0.2.* (2022). Available online at: https://CRAN.R-project.org/package=xgboost (accessed May 25, 2022).

52. Microsoft. *LightGBM: light gradient boosting machine. R Package version, 3.3.2.* (2022). Available online at: https://github.com/microsoft/LightGBM (accessed May 25, 2022).

53. Kuhn M. *Caret: classification and regression training. R Package version, 6.0-92.* (2022). Available online at: https://CRAN.R-project.org/package=caret (accessed May 25, 2022).

54. Kuo C, Chang C, Liu K, Lin W, Chiang H, Chung C, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. *NPJ Digit Med.* (2019) 2:29. doi: 10.1038/s41746-019-0104-2

55. Dong J, Feng T, Thapa-Chhetry B, Cho B, Shum T, Inwald D, et al. Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care. *Crit Care.* (2021) 25:288. doi: 10.1186/s13054-021-03724-0

56. Chaudhary K, Vaid A, Duffy Á, Paranjpe I, Jaladanki S, Paranjpe M, et al. Utilization of deep learning for subphenotype identification in sepsis-associated acute kidney injury. *Clin J Am Soc Nephrol.* (2020) 15:1557–65. doi: 10.2215/CJN.09330819

57. Magherini R, Mussi E, Volpe Y, Furferi R, Buonamici F, Servi M. Machine learning for renal pathologies: an updated survey. *Sensors.* (2022) 22:4989. doi: 10.3390/s22134989

58. Huo Y, Deng R, Liu Q, Fogo A, Yang H. AI applications in renal pathology. *Kidney Int.* (2021) 99:1309–20. doi: 10.1016/j.kint.2021.01.015

59. Barbieri C, Cattinelli I, Neri L, Mari F, Ramos R, Brancaccio D, et al. Development of an artificial intelligence model to guide the management of blood pressure, fluid volume, and dialysis dose in end-stage kidney disease patients: proof of concept and first clinical assessment. *Kidney Dis.* (2019) 5:28–33. doi: 10.1159/000493479

60. Kim H, Heo S, Kim J, Kim A, Nam C, Kim B. Dialysis adequacy predictions using a machine learning method. *Sci Rep.* (2021) 11:15417. doi: 10.1038/s41598-021-94964-1

61. Zhao J, Gu S, McDermaid A. Predicting outcomes of chronic kidney disease from EMR data based on random forest regression. *Math Biosci.* (2019) 310:24–30. doi: 10.1016/j.mbs.2019.02.001

62. Akbilgic O, Obi Y, Potukuchi P, Karabayir I, Nguyen D, Soohoo M, et al. Machine learning to identify dialysis patients at high death risk. *Kidney Int Rep.* (2019) 4:1219–29. doi: 10.1016/j.ekir.2019.06.009

63. Chen T, Li X, Li Y, Xia E, Qin Y, Liang S, et al. Prediction and risk stratification of kidney outcomes in IgA nephropathy. *Am J Kidney Dis.* (2019) 74:300–9. doi: 10.1053/j.ajkd.2019.02.016

64. Segal Z, Kalifa D, Radinsky K, Ehrenberg B, Elad G, Maor G, et al. Machine learning algorithm for early detection of end-stage renal disease. *BMC Nephrol.* (2020) 21:518. doi: 10.1186/s12882-020-02093-0

65. Bai Q, Su C, Tang W, Li Y. Machine learning to predict end stage kidney disease in chronic kidney disease. *Sci Rep.* (2022) 12:8377. doi: 10.1038/s41598-022-12316-z

66. Janmaat C, van Diepen M, van Hagen C, Rotmans J, Dekker F, Dekkers O. Decline of kidney function during the pre-dialysis period in chronic kidney disease patients: a systematic review and meta-analysis. *Clin Epidemiol.* (2018) 10:613–22. doi: 10.2147/CLEP.S153367

67. Da J, Xie X, Wolf M, Disthabanchong S, Wang J, Zha Y, et al. Serum phosphorus and progression of CKD and mortality: a meta-analysis of cohort studies. *Am J Kidney Dis.* (2015) 66:258–65. doi: 10.1053/j.ajkd.2015.01.009

68. Lee J, Park J, Joo Y, Lee C, Yun H, Yoo T, et al. Association of blood pressure with the progression of CKD: findings from KNOW-CKD study. *Am J Kidney Dis.* (2021) 78:236–45. doi: 10.1053/j.ajkd.2020.12.013

69. Lang J, Katz R, Ix J, Gutierrez O, Peralta C, Parikh C, et al. Association of serum albumin levels with kidney function decline and incident chronic kidney disease in elders. *Nephrol Dial Transplant.* (2018) 33:986–92. doi: 10.1093/ndt/gfx229

70. Pan W, Han Y, Hu H, He Y. Association between hemoglobin and chronic kidney disease progression: a secondary analysis of a prospective cohort study in Japanese patients. *BMC Nephrol.* (2022) 23:295. doi: 10.1186/s12882-022-02920-6

71. Tsai C, Lin S, Kuo C, Huang C. Serum uric acid and progression of kidney disease: a longitudinal analysis and mini-review. *PLoS One.* (2017) 12:e0170393. doi: 10.1371/journal.pone.0170393

72. Betônico C, Titan S, Correa-Giannella M, Nery M, Queiroz M. Management of diabetes mellitus in individuals with chronic kidney disease: therapeutic perspectives and glycemic control. *Clinics.* (2016) 71:47–53. doi: 10.6061/clinics/2016(01)08

73. Tsai M, Lin M, Hsu C, Yen A, Chen T, Chiu S, et al. Factors associated with renal function state transitions: a population-based community survey in Taiwan. *Front Public Health.* (2022) 10:930798. doi: 10.3389/fpubh.2022.930798

74. Alicic R, Rooney M, Tuttle K. Diabetic kidney disease: challenges, progress, and possibilities. *Clin J Am Soc Nephrol.* (2017) 12:2032–45. doi: 10.2215/CJN.11491116

75. Pasternak M, Liu P, Quinn R, Elliott M, Harrison T, Hemmelgarn B, et al. Association of albuminuria and regression of chronic kidney disease in adults with newly diagnosed moderate to severe chronic kidney disease. *JAMA Netw Open.* (2022) 5:e2225821. doi: 10.1001/jamanetworkopen.2022.25821

76. Johnson R, Nakagawa T, Jalal D, Sánchez-Lozada L, Kang D, Ritz E. Uric acid and chronic kidney disease: which is chasing which? *Nephrol Dial Transplant.* (2013) 28:2221–8. doi: 10.1093/ndt/gft029

77. Li L, Yang C, Zhao Y, Zeng X, Liu F, Fu P. Is hyperuricemia an independent risk factor for new-onset chronic kidney disease?: a systematic review and meta-analysis based on observational cohort studies. *BMC Nephrol.* (2014) 15:122. doi: 10.1186/1471-2369-15-122

78. Johnson R, Bakris G, Borghi C, Chonchol M, Feldman D, Lanaspa M, et al. Hyperuricemia, acute and chronic kidney disease, hypertension, and cardiovascular disease: report of a scientific workshop organized by the National Kidney Foundation. *Am J Kidney Dis.* (2018) 71:851–65. doi: 10.1053/j.ajkd.2017.12.009

79. Srivastava A, Kaze A, McMullan C, Isakova T, Waikar S. Uric acid and the risks of kidney failure and death in individuals with CKD. *Am J Kidney Dis.* (2018) 71:362–70. doi: 10.1053/j.ajkd.2017.08.017

80. Kimura K, Hosoya T, Uchida S, Inaba M, Makino H, Maruyama S, et al. Febuxostat therapy for patients with stage 3 CKD and asymptomatic hyperuricemia: a randomized trial. *Am J Kidney Dis.* (2018) 72:798–810. doi: 10.1053/j.ajkd.2018.06.028

81. Badve S, Pascoe E, Tiku A, Boudville N, Brown F, Cass A, et al. Effects of allopurinol on the progression of chronic kidney disease. *N Engl J Med.* (2020) 382:2504–13. doi: 10.1056/NEJMoa1915833

82. Doria A, Galecki A, Spino C, Pop-Busui R, Cherney D, Lingvay I, et al. Serum urate lowering with allopurinol and kidney function in type 1 diabetes. *N Engl J Med.* (2020) 382:2493–503. doi: 10.1056/NEJMoa1916624

83. Giachelli C. The emerging role of phosphate in vascular calcification. *Kidney Int.* (2009) 75:890–7. doi: 10.1038/ki.2008.644

84. Waziri B, Duarte R, Naicker S. Chronic kidney disease-mineral and bone disorder (CKD-MBD): current perspectives. *Int J Nephrol Renovasc Dis.* (2019) 12:263–76. doi: 10.2147/IJNRD.S191156

85. Perwad F, Zhang M, Tenenhouse H, Portale A. Fibroblast growth factor 23 impairs phosphorus and vitamin D metabolism in vivo and suppresses 25-hydroxyvitamin D-1alpha-hydroxylase expression in vitro. *Am J Physiol Renal Physiol.* (2007) 293:F1577–83. doi: 10.1152/ajprenal.00463.2006

86. Scialla J, Xie H, Rahman M, Anderson A, Isakova T, Ojo A, et al. Fibroblast growth factor-23 and cardiovascular events in CKD. *J Am Soc Nephrol.* (2014) 25:349–60. doi: 10.1681/ASN.2013050465

87. Tsai M, Leu J, Fang Y, Liou H. High fibroblast growth factor 23 levels associated with low hemoglobin levels in patients with chronic kidney disease stages 3 and 4. *Medicine.* (2016) 95:e3049. doi: 10.1097/MD.0000000000003049

88. Rahman A, Hossain M, Muhammad G, Kundu D, Debnath T, Rahman M, et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. *Cluster Comput.* (2022):1–41. doi: 10.1007/s10586-022-03658-4 [Online ahead of print].