



OPEN

Machine learning to predict end stage kidney disease in chronic kidney disease

Qiong Bai¹, Chunyan Su¹, Wen Tang¹✉ & Yike Li²✉

The purpose of this study was to assess the feasibility of machine learning (ML) in predicting the risk of end-stage kidney disease (ESKD) from patients with chronic kidney disease (CKD). Data were obtained from a longitudinal CKD cohort. Predictor variables included patients' baseline characteristics and routine blood test results. The outcome of interest was the presence or absence of ESKD by the end of 5 years. Missing data were imputed using multiple imputation. Five ML algorithms, including logistic regression, naïve Bayes, random forest, decision tree, and K-nearest neighbors were trained and tested using fivefold cross-validation. The performance of each model was compared to that of the Kidney Failure Risk Equation (KFRE). The dataset contained 748 CKD patients recruited between April 2006 and March 2008, with the follow-up time of 6.3 ± 2.3 years. ESKD was observed in 70 patients (9.4%). Three ML models, including the logistic regression, naïve Bayes and random forest, showed equivalent predictability and greater sensitivity compared to the KFRE. The KFRE had the highest accuracy, specificity, and precision. This study showed the feasibility of ML in evaluating the prognosis of CKD based on easily accessible features. Three ML models with adequate performance and sensitivity scores suggest a potential use for patient screenings. Future studies include external validation and improving the models with additional predictor variables.

Chronic kidney disease (CKD) is a significant healthcare burden that affects billions of individuals worldwide^{1,2} and makes a profound impact on global morbidity and mortality³⁻⁵. In the United States, approximately 11% of the population or 37 million people suffer from CKD that results in an annual Medicare cost of \$84 billion⁶. The prevalence of this disease is estimated at 10.8% in China, affecting about 119.5 million people⁷.

Gradual loss of the kidney function can lead to end stage kidney disease (ESKD) in CKD patients, precipitating the need for kidney replacement therapy (KRT). Timely intervention in those CKD patients who have a high risk of ESKD may not only improve these patients' quality of life by delaying the disease progression, but also reduce the morbidity, mortality and healthcare costs resulting from KRT^{8,9}. Because the disease progression is typically silent¹⁰, a reliable prediction model for risk of ESKD at the early stage of CKD can be clinically essential. Such a model is expected to facilitate physicians in making personalized treatment decisions for high-risk patients, thereby improving the overall prognosis and reducing the economic burden of this disease.

A few statistical models were developed to predict the likelihood of ESKD based on certain variables, including age, gender, lab results, and most commonly, the estimated glomerular filtration rate (eGFR) and albuminuria^{11,12}. Although some of these models demonstrated adequate predictability in patients of a specific race, typically Caucasians¹³⁻¹⁵, literature on their generalizability in other ethnic groups, such as Chinese, remains scarce^{13,16}. In addition, models based on non-urine variables, such as patients' baseline characteristics and routine blood tests, have reportedly yield sufficient performance^{17,18}. Therefore, it may be feasible to predict ESKD without urine tests, leading to a simplified model with equivalent reliability.

With the advent of the big data era, new methods became available in developing a predictive model that used to rely on traditional statistics. Machine learning (ML) is a subset of artificial intelligence (AI) that allows the computer to perform a specific task without explicit instructions. When used in predictive modeling, ML algorithm can be trained to capture the underlying patterns of the sample data and make predictions about the new data based on the acquired information¹⁹. Compared to traditional statistics, ML represents more sophisticated math functions and usually results in better performance in predicting an outcome that is determined

¹Department of Nephrology, Peking University Third Hospital, 49 North Garden Rd, Haidian District, Beijing 100191, People's Republic of China. ²Department of Otolaryngology-Head and Neck Surgery, Bill Wilkerson Center, Vanderbilt University Medical Center, Nashville, TN, USA. ✉email: tangwen@126.com; yike.li.1@vumc.org

by a large set of variables with non-linear, complex interactions²⁰. ML has recently been applied in numerous studies and demonstrated high level of performance that surpassed traditional statistics and even humans^{20–23}.

This article presents a proof-of-concept study with the major goal to establish ML models for predicting the risk of ESKD on a Chinese CKD dataset. The ML models were trained and tested based on easily obtainable variables, including the baseline characteristics and routine blood tests. Results obtained from this study suggest not only the feasibility of ML models in performing this clinically critical task, but also the potential in facilitating personalized medicine.

Materials and methods

Study population. The data used for this retrospective work were obtained from a longitudinal cohort previously enrolled in an observational study^{24,25}. The major inclusion criteria for the cohort were adult CKD patients (≥ 18 years old) with stable kidney functions for at least three months prior to recruitment. Patients were excluded if they had one or more of the following situations: (1) history of KRT in any form, including hemodialysis, peritoneal dialysis or kidney transplantation; (2) any other existing condition deemed physically unstable, including life expectancy < 6 months, acute heart failure, and advanced liver disease; (3) any pre-existing malignancy. All patients were recruited from the CKD management clinic of Peking University Third Hospital between April 2006 and March 2008. Written informed consent was obtained from all patients. They were treated according to routine clinical practice determined by the experienced nephrologists and observed until December 31st, 2015. Detailed information regarding patient recruitment and management protocol has been described in a previous publication²⁴.

Data acquisition. Patient characteristics included age, gender, education level, marriage status, and insurance status. Medical history comprised history of smoking, history of alcohol consumption, presence of each comorbid condition—diabetes, cardiovascular disease and hypertension. Clinical parameters contained body mass index (BMI), systolic pressure and diastolic pressure. Blood tests consisted of serum creatinine, uric acid, blood urea nitrogen, white blood cell count, hemoglobin, platelets count, alanine aminotransferase (ALT), aspartate aminotransferase (AST), total protein, albumin, alkaline phosphatase (ALP), high-density lipoprotein, low-density lipoprotein, triglycerides, total cholesterol, calcium, phosphorus, potassium, sodium, chloride, and bicarbonate. The estimated glomerular filtration rate and type of primary kidney disease were also used as predictors.

All baseline variables were obtained at the time of subject enrollment. The primary study end point was kidney failure which necessitated the use of any KRT. Subjects with the outcome of kidney failure were labeled as ESKD+, and the rest ESKD-. Patients who died before reaching the study end point or lost to follow up were discarded. Patients who developed ESKD after five years were labeled as ESKD-.

Data preprocessing. All categorical variables, such as insurance status, education, and primary disease, were encoded using the one-hot approach. Any variable was removed from model development if the missing values were greater than 50%. Missing data were handled using multiple imputation with five times of repetition, leading to five slightly different imputed datasets where each of the missing values was randomly sampled from their predictive distribution based on the observed data. On each imputed set, all models were trained and tested using a fivefold cross validation method. To minimize selection bias, subject assignment to train/test folds was kept consistent across all imputed sets. Data were split in a stratified fashion to ensure the same distribution of the outcome classes (ESKD+ vs. ESKD-) in each subset as the entire set.

Model development. The model was trained to perform a binary classification task with the goal of generating the probability of ESKD+ based on the given features. Five ML algorithms were employed in this study, including logistic regression, naïve Bayes, random forest, decision tree, and K-nearest neighbors. Grid search was performed to obtain the best hyperparameter combination for each algorithm.

Assessment of model performance. The performance of a classifiers was measured using accuracy, precision, recall, specificity, F1 score and area under the curve (AUC), as recommended by guidelines for results reporting of clinical prediction models²⁶. All classifiers developed in this study were further compared with the Kidney Failure Risk Equation (KFRE), which estimates the 5-year risk of ESKD based on patient's age, gender, and eGFR¹². The KFRE is currently the most widely used model in predicting CKD progression to ESKD. The reported outcome of a model represented the average performance of 5 test folds over all imputed sets.

Statistical analysis. Basic descriptive statistics were applied as deemed appropriate. Results are expressed as frequencies and percentages for categorical variables; the mean \pm standard deviation for continuous, normally distributed variables; and the median (interquartile range) for continuous variables that were not normally distributed. Patient characteristics were compared between the original dataset and the imputed sets using one-way analysis of variance (ANOVA). The AUC of each model was measured using the predicted probability. The optimal threshold of a classifier was determined based on the receiver operating characteristic (ROC) curve at the point with minimal distance to the upper left corner. For each ML model, this threshold was obtained during the training process and applied unchangeably to the test set. For the KFRE, the threshold was set at a default value of 0.5. Model development, performance evaluation and data analyses were all performed using Python²⁷. The alpha level was set at 0.05.

Ethical approval. This research was conducted ethically in accordance with the World Medical Association Declaration of Helsinki. The study protocol has been approved by the Peking University Third Hospital Medical Science Research Ethics Committee on human research (No. M2020132).

Results

Cohort characteristics. The dataset contained a total of 748 subjects with the follow-up duration of 6.3 ± 2.3 years. The baseline characteristics are summarized in Table 1. Most patients were in stage 2 (24.5%) or 3 (47.1%) CKD at baseline. ESKD was observed in 70 patients (9.4%), all of whom subsequently received KRT, including hemodialysis in 49 patients, peritoneal dialysis in 17 and kidney transplantation in 4.

Model performance. Details of the five imputed sets are provided in the supplemental materials. There was no significant difference between the imputed sets and the original dataset in each variable where missing data were replaced by imputed values. The hyperparameter settings for each classifier are displayed in Table 2. The best overall performance, as measured by the AUC score, was achieved by the random forest algorithm (0.81, see Table 3). Nonetheless, this score and its 95% confidence interval had overlap with those of the other three models, including the logistic regression, naïve Bayes, and the KFRE (Fig. 1). Interestingly, the KFRE model that was based on 3 simple variables, demonstrated not only a comparable AUC score but also the highest accuracy, specificity, and precision. At the default threshold, however, the KFRE was one of the least sensitive models (47%).

Discussion

With extensive utilization of electronic health record and recent progress in ML research, AI is expanding its impact on healthcare and has gradually changed the way clinicians pursue for problem-solving²⁸. Instead of adopting a theory-driven strategy that requires a preformed hypothesis from prior knowledge, training an ML model typically follows a data-driven approach that allows the model to learn from experience alone. Specifically, the model improves its performance iteratively on a training set by comparing the predictions to the ground truths and adjusting model parameters so as to minimize the distance between the predictions and the truths. In nephrology, ML has demonstrated promising performances in predicting acute kidney injury or time to allograft loss from clinical features^{29,30}, recognizing specific patterns in pathology slides^{31,32}, choosing an optimal dialysis prescription³³, or mining text in the electronic health record to find specific cases^{34,35}. Additionally, a few recent studies were performed to predict the progression of CKD using ML methods. These models were developed to estimate the risk of short-term mortality following dialysis³⁶, calculate the future eGFR values³⁷, or assess the 24-h urinary protein levels¹⁸. To our best knowledge, there hasn't been any attempt to apply ML methods to predict the occurrence of ESKD in CKD patients.

In the present study, a prediction model for ESKD in CKD patients was explored using ML techniques. Most classifiers demonstrated adequate performance based on easily accessible patient information that is convenient for clinical translation. In general, three ML models, including the logistic regression, naïve Bayes and random forest, showed non-inferior performance to the KFRE in this study. These findings imply ML as a feasible approach for predicting disease progression in CKD, which could potentially guide physicians in establishing personalized treatment plans for this condition at an early stage. These ML models with higher sensitivity scores may also be practically favored in patient screening over the KFRE.

To our best understanding, this study was also the first to validate the KFRE in CKD patients of Mainland China. The KFRE was initially developed and validated using North American patients with CKD stage 3–5¹². There were seven KFRE models that consisted of different combinations of predictor variables. The most commonly used KFRE included a 4-variable model (age, gender, eGFR and urine ACR) or an 8-variable model (age, gender, eGFR, urine ACR, serum calcium, phosphorous, bicarbonate, and albumin). Besides, there was a 3-variable model (age, gender, and eGFR) that required no urine ACR and still showed comparable performance to the other models in the original article. Despite its favorable performance in prediction for ESKD in patients of Western countries^{14,15,38,39}, the generalizability of KFRE in Asian population remained arguable following the suboptimal results revealed by some recent papers^{13,40,41}. In the current study, the KFRE was validated in a Chinese cohort with CKD stage 1–5 and showed an AUC of 0.80. This result indicated the KFRE was adequately applicable to the Chinese CKD patients and even earlier disease stages. In particular, the high specificity score (0.95) may favor the use of this equation in ruling in patients who require close monitoring of disease progression. On the other hand, a low sensitivity (0.47) at the default threshold may suggest it may be less desirable than the other models for ruling out patients.

Urine test is a critical diagnostic approach for CKD. The level of albuminuria (i.e. ACR) has also been regarded as a major predictor for disease progression and therefore used by most prognostic models. However, quantitative testing for albuminuria is not always available in China especially in rural areas, which precludes clinicians from using most urine-based models for screening patients. In this regard, several simplified models were developed to predict CKD progression without the need of albuminuria. These models were based on patient characteristics (e.g. age, gender, BMI, comorbidity) and/or blood work (e.g. creatinine/eGFR, BUN), and still able to achieve an AUC of 0.87–0.89^{12,18} or a sensitivity of 0.88³⁷. Such performance was largely consistent with the findings of this study and comparable or even superior to some models incorporating urine tests^{16,42}. Altogether, it suggested a reliable prediction for CKD progression may be obtained from routine clinical variables without urine measures. These models are expected to provide a more convenient screening tool for CKD patients in developing regions.

Missing data are such a common problem in ML research that they can potentially lead to a biased model and undermine the validity of study outcomes. Traditional methods to handle missing data include complete case analysis, missing indicator, single value imputation, sensitivity analyses, and model-based methods (e.g. mixed models or generalized estimating equations)^{43–45}. In most scenarios, complete case analysis and single

| Variables | Original data |
|--|----------------------|
| Age (years) | 57.8 ± 17.6 |
| Gender (male/female) | 419/329 |
| SBP (mmHg) | 129.5 ± 17.8 |
| DBP (mmHg) | 77.7 ± 11.1 |
| BMI (kg/m ²) | 24.8 ± 3.7 |
| Primary disease | |
| Primary GN | 292 (39.0%) |
| Diabetes | 224 (29.9%) |
| Hypertension | 97 (13.0%) |
| CIN | 64 (8.6%) |
| Others | 18 (2.4%) |
| Unknown | 53 (7.1%) |
| Creatinine (μmol/L) | 130.0 (100.0, 163.0) |
| Urea (mmol/L) | 7.9 (5.6, 10.4) |
| ALT (U/L) | 17.0 (12.0, 24.0) |
| AST (U/L) | 18.0 (15.0, 22.0) |
| ALP (U/L) | 60.0 (50.0, 75.0) |
| Total protein (g/L) | 71.6 ± 8.4 |
| Albumin (g/L) | 42.2 ± 5.6 |
| Urine acid (μmol/L) | 374.0 (301.0, 459.0) |
| Calcium (mmol/L) | 2.2 ± 0.1 |
| Phosphorous (mmol/L) | 1.2 ± 0.2 |
| Ca × P (mg ² /dL ²) | 33.5 ± 5.6 |
| Blood leukocyte (10 ⁹ /L) | 7.1 ± 2.4 |
| Hemoglobin (g/L) | 131.0 ± 20.3 |
| Platelet (10 ⁹ /L) | 209.8 ± 57.1 |
| eGFR (ml/min/1.73m ²) | 46.1 (32.6, 67.7) |
| CKD stage | |
| Stage 1 | 58 (7.8%) |
| Stage 2 | 183 (24.5%) |
| Stage 3 | 352 (47.1%) |
| Stage 4 | 119 (15.9%) |
| Stage 5 | 36 (4.8%) |
| Total cholesterol | 5.1 (4.3, 5.9) |
| Triglyceride | 1.8 (1.3, 2.6) |
| HDL-c | 1.3 (1.1, 1.6) |
| LDL-c | 3.0 (2.4, 3.7) |
| Fasting glucose (mmol/L) | 5.4 (4.9, 6.2) |
| Potassium (mmol/L) | 4.3 ± 0.5 |
| Sodium (mmol/L) | 140.2 ± 2.8 |
| Chlorine (mmol/L) | 106.9 ± 3.7 |
| Bicarbonate (mmol/L) | 25.9 ± 3.6 |
| Medical history | |
| Hypertension | 558 (74.6%) |
| Diabetes mellitus | 415 (55.5%) |
| Cardiovascular or cerebrovascular disease | 177 (23.7%) |
| Smoking | 91 (12.6%) |

Table 1. Baseline patient characteristics. *SBP* systolic blood pressure, *DBP* diastolic blood pressure, *GN* glomerulonephritis, *CIN* chronic interstitial nephritis, *BMI* body mass index, *eGFR* estimated glomerular filtration rate, *ALT* alanine aminotransferase, *AST* aspartate transaminase, *ALP* alkaline phosphatase, *CKD* chronic kidney disease, *HDL-c* high density lipoprotein cholesterol, *LDL-c* low density lipoprotein cholesterol, *Ca × P* calcium-phosphorus product.

value imputation are favored by researchers primarily due to the ease of implementation^{45–47}. However, these methods may be associated with significant drawbacks. For example, by excluding samples with missing data

| Algorithms | Hyperparameters |
|---------------------|--|
| Logistic regression | penalty = 'l2', class_weight = 'balanced', max_iter = 100000, C = 10, solver = 'liblinear' |
| Naive Bayes | type = 'multinomial', alpha = 150 |
| Decision tree | criterion = 'gini', splitter = 'best', max_depth = 16, max_features = 15, min_samples_leaf = 5, min_samples_split = 0.0001 |
| Random forest | class_weight = 'balanced', criterion = 'gini', max_depth = 9, max_features = 17, min_samples_leaf = 6, min_samples_split = 30, n_estimators = 32 |
| K-nearest neighbors | weights = 'distance', metric = 'minkowski', n_neighbors = 16, leaf_size = 10 |

Table 2. Hyperparameters of the algorithms.

| | Accuracy | Sensitivity | Specificity | Precision | F1 Score | AUC |
|---------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Logistic regression | 0.75 (0.72, 0.79) | 0.79 (0.73, 0.85) | 0.75 (0.71, 0.79) | 0.26 (0.24, 0.29) | 0.38 (0.36, 0.41) | 0.79 (0.77, 0.82) |
| Naïve Bayes | 0.86 (0.85, 0.87) | 0.72 (0.68, 0.75) | 0.87 (0.86, 0.89) | 0.37 (0.35, 0.40) | 0.49 (0.46, 0.51) | 0.80 (0.77, 0.82) |
| Random forest | 0.82 (0.80, 0.85) | 0.76 (0.71, 0.81) | 0.83 (0.80, 0.86) | 0.34 (0.30, 0.39) | 0.46 (0.43, 0.49) | 0.81 (0.78, 0.83) |
| K nearest neighbor | 0.84 (0.81, 0.86) | 0.60 (0.57, 0.64) | 0.86 (0.83, 0.89) | 0.35 (0.30, 0.40) | 0.43 (0.40, 0.46) | 0.73 (0.71, 0.75) |
| Decision tree | 0.84 (0.82, 0.86) | 0.44 (0.39, 0.49) | 0.89 (0.86, 0.91) | 0.33 (0.26, 0.40) | 0.35 (0.32, 0.39) | 0.66 (0.63, 0.68) |
| KFRE | 0.90 (0.90, 0.91) | 0.47 (0.42, 0.52) | 0.95 (0.94, 0.96) | 0.50 (0.45, 0.55) | 0.48 (0.43, 0.52) | 0.80 (0.78, 0.83) |

Table 3. The performance of all algorithms. All outcomes are expressed as mean and (95% confidence interval). *KFRE* kidney failure risk equation, *AUC* area under the curve.

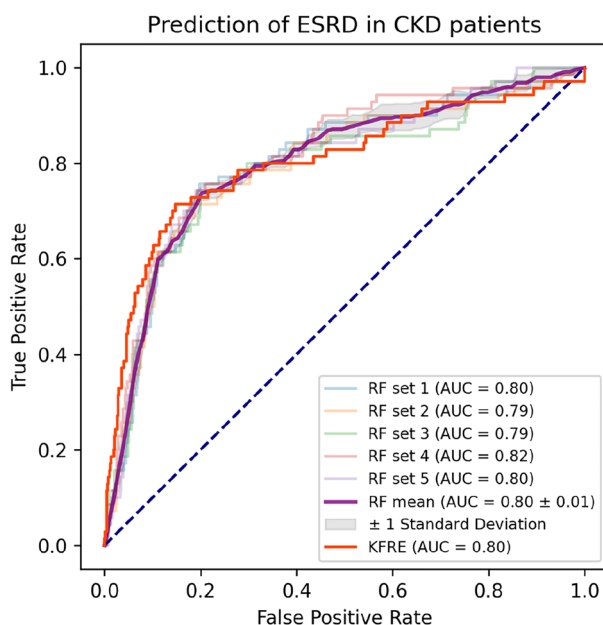


Figure 1. ROC curves of the random forest algorithm and the KFRE model.

from analyses, complete case analysis can result in reduction of model power, overestimation of benefit and underestimation of harm^{43,46}; Single value imputation replaces the missing data by a single value—typically the mean or mode of the complete cases, thereby increasing the homogeneity of data and overestimating the precision^{43,48}. In this regard, multiple imputation solves these problems by generating several different plausible imputed datasets, which account for the uncertainty about the missing data and provide unbiased estimates of the true effect^{49,50}. It is deemed effective regardless of the pattern of missingness^{43,51}. Multiple imputation is now widely recognized as the standard method to deal with missing data in many areas of research^{43,45}. In the current study, a 5-set multiple imputation method was employed to obtain reasonable variability of the imputed data. The performance of each model was analyzed on each imputed set and pooled for the final result. These procedures ensured that the model bias resulting from missing data was minimized. In the future, multiple imputation is expected to become a routine method for missing data handling in ML research, as the extra amount of

computation associated with multiple imputation over those traditional methods can simply be fulfilled by the high level of computational power required by ML.

Although ML has been shown to outperform traditional statistics in a variety of tasks by virtue of the model complexity, some studies demonstrated no gain or even declination of performance compared to traditional regression methods^{52,53}. In this study, the simple logistic regression model also yielded a comparable or even superior predictability for ESKD to other ML algorithms. The most likely explanation is that the current dataset only had a small sample size and limited numbers of predictor variables, and the ESKD+ cases were relatively rare. The lack of big data and imbalanced class distribution may have negative impact on the performance of complex ML algorithms, as they are typically data hungry⁵⁴. On the other hand, this finding could imply simple interactions among the predictor variables. In other words, the risk of ESKD may be largely influenced by only a limited number of factors in an uncomplicated fashion, which is consistent with some previous findings^{12,18,55}. The fact that the 3-variable KFRE, which is also a regression model, yielded equivalent outcomes to the best ML models in this study may further support this implication. It is therefore indicated that traditional regression models may continue to play a key role in disease risk prediction, especially when a small sample size, limited predictor variables, or an imbalanced dataset is encountered. The fact that some of the complex ML models are subject to the risk of overfitting and the lack of interpretability further favors the use of simple regression models, which can be translated to explainable equations.

Several limitations should be noted. First, this cohort consisted of less than 1000 subjects and ESKD only occurred in a small portion of them, both of which might have affected model performance as discussed earlier. Second, although this study aimed to assess the feasibility of a prediction model for ESKD without any urine variables, this was partially due to the lack of quantitative urine tests at our institute when this cohort was established. As spot urine tests become increasingly popular, urine features such as ACR will be as accessible and convenient as other lab tests. They are expected to play a critical role in more predictive models. Third, the KFRE was previously established on stages 3–5 CKD patients while the current cohort contained stages 1–5. This discrepancy may have affected the KFRE performance. Forth, the generalizability of this model has not been tested on any external data due to the lack of such resource in this early feasibility study. Therefore, additional efforts are required to improve and validate this model before any clinical translation. Finally, although a simple model without urine variables is feasible and convenient, model predictability may benefit from a greater variety of clinical features, such as urine tests, imaging, or biopsy. Future works should include training ML models with additional features using a large dataset, and validating them on external patients.

In conclusion, this study showed the feasibility of ML in evaluating the prognosis of CKD based on easily accessible features. Logistic regression, naïve Bayes and random forest demonstrated comparable predictability to the KFRE in this study. These ML models also had greater sensitivity scores that were potentially advantageous for patient screenings. Future studies include performing external validation and improving the model with additional predictor variables.

Received: 19 July 2021; Accepted: 9 May 2022

Published online: 19 May 2022

References

- Zhang, L. *et al.* Trends in chronic kidney disease in China. *N. Engl. J. Med.* **375**, 905–906. <https://doi.org/10.1056/NEJMc1602469> (2016).
- Bello, A. K. *et al.* Effective CKD care in European countries: Challenges and opportunities for health policy. *Am. J. Kidney Dis.* **65**, 15–25. <https://doi.org/10.1053/j.ajkd.2014.07.033> (2015).
- Subbiah, A. K., Chhabra, Y. K. & Mahajan, S. Cardiovascular disease in patients with chronic kidney disease: A neglected subgroup. *Heart Asia* **8**, 56–61. <https://doi.org/10.1136/heartasia-2016-010809> (2016).
- Pecoits-Filho, R. *et al.* Interactions between kidney disease and diabetes: Dangerous liaisons. *Diabetol. Metab. Syndr.* **8**, 50. <https://doi.org/10.1186/s13098-016-0159-z> (2016).
- Weiner, D. E. *et al.* Chronic kidney disease as a risk factor for cardiovascular disease and all-cause mortality: A pooled analysis of community-based studies. *J. Am. Soc. Nephrol.* **15**, 1307–1315. <https://doi.org/10.1097/01.asn.0000123691.46138.e2> (2004).
- Saran, R., Robinson, B., Abbott, K. C. *et al.* US Renal Data System 2019 Annual Data Report: Epidemiology of kidney disease in the United States. *Am. J. Kidney Dis.* **75**, A6–A7. <https://doi.org/10.1053/j.ajkd.2019.09.003> (2020).
- Zhang, L. *et al.* Prevalence of chronic kidney disease in China: A cross-sectional survey. *Lancet (London, England)* **379**, 815–822. [https://doi.org/10.1016/S0140-6736\(12\)60033-6](https://doi.org/10.1016/S0140-6736(12)60033-6) (2012).
- Johns, T. S., Yee, J., Smith-Jules, T., Campbell, R. C. & Bauer, C. Interdisciplinary care clinics in chronic kidney disease. *BMC Nephrol.* **16**, 161. <https://doi.org/10.1186/s12882-015-0158-6> (2015).
- Lin, E., Chertow, G. M., Yan, B., Malcolm, E. & Goldhaber-Fiebert, J. D. Cost-effectiveness of multidisciplinary care in mild to moderate chronic kidney disease in the United States: A modeling study. *PLoS Med.* **15**, e1002532. <https://doi.org/10.1371/journal.pmed.1002532> (2018).
- Zhong, J., Yang, H. C. & Fogo, A. B. A perspective on chronic kidney disease progression. *Am. J. Physiol. Ren. Physiol.* **312**, F375–F384. <https://doi.org/10.1152/ajprenal.00266.2016> (2017).
- Tangri, N. *et al.* A dynamic predictive model for progression of CKD. *Am. J. Kidney Dis.* **69**, 514–520. <https://doi.org/10.1053/j.ajkd.2016.07.030> (2017).
- Tangri, N. *et al.* A predictive model for progression of chronic kidney disease to kidney failure. *JAMA* **305**, 1553–1559. <https://doi.org/10.1001/jama.2011.451> (2011).
- Tangri, N. *et al.* Multinational assessment of accuracy of equations for predicting risk of kidney failure: A meta-analysis. *JAMA* **315**, 164–174. <https://doi.org/10.1001/jama.2015.18202> (2016).
- Major, R. W. *et al.* The Kidney Failure Risk Equation for prediction of end stage renal disease in UK primary care: An external validation and clinical impact projection cohort study. *PLoS Med.* **16**, e1002955. <https://doi.org/10.1371/journal.pmed.1002955> (2019).
- Peeters, M. J. *et al.* Validation of the kidney failure risk equation in European CKD patients. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. Eur. Ren. Assoc.* **28**, 1773–1779. <https://doi.org/10.1093/ndt/gft063> (2013).

16. Echouffo-Tcheugui, J. B. & Kengne, A. P. Risk models to predict chronic kidney disease and its progression: A systematic review. *PLoS Med.* **9**, e1001344. <https://doi.org/10.1371/journal.pmed.1001344> (2012).
17. Chang, H. L. *et al.* A predictive model for progression of CKD. *Medicine (Baltimore)* **98**, e16186. <https://doi.org/10.1097/MD.00000000000016186> (2019).
18. Xiao, J. *et al.* Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J. Transl. Med.* **17**, 119. <https://doi.org/10.1186/s12967-019-1860-0> (2019).
19. Song, H., Triguero, I. & Özcan, E. A review on the self and dual interactions between machine learning and optimisation. *Prog. Artif. Intell.* **8**, 143–165. <https://doi.org/10.1007/s13748-019-00185-z> (2019).
20. Mortazavi, B. J. *et al.* Analysis of machine learning techniques for heart failure readmissions. *Circ. Cardiovasc. Qual. Outcomes* **9**, 629–640. <https://doi.org/10.1161/CIRCOUTCOMES.116.003039> (2016).
21. Deo, R. C. Machine learning in medicine. *Circulation* **132**, 1920–1930. <https://doi.org/10.1161/CIRCULATIONAHA.115.001593> (2015).
22. Weng, S. F., Reips, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* **12**, e0174944. <https://doi.org/10.1371/journal.pone.0174944> (2017).
23. Meiring, C. *et al.* Optimal intensive care outcome prediction over time using machine learning. *PLoS ONE* **13**, e0206862. <https://doi.org/10.1371/journal.pone.0206862> (2018).
24. Lai, X. *et al.* Outcomes of stage 1–5 chronic kidney disease in Mainland China. *Ren. Fail.* **36**, 520–525. <https://doi.org/10.3109/0886022X.2013.875859> (2014).
25. Bai, Q., Su, C. Y., Zhang, A. H., Wang, T. & Tang, W. Loss of the normal gradient in arterial compliance and outcomes of chronic kidney disease patients. *Cardiorenal Med.* **9**, 297–307. <https://doi.org/10.1159/000500479> (2019).
26. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931. <https://doi.org/10.1093/eurheartj/ehu207> (2014).
27. Python: A dynamic, open source programming language. <https://www.python.org/> (2019).
28. Yu, K., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2**, 719–731. <https://doi.org/10.1038/s41551-018-0305-z> (2018).
29. Mohamadlou, H. *et al.* Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Can. J. Kidney Health Dis.* **5**, 2054358118776326. <https://doi.org/10.1177/2054358118776326> (2018).
30. Lee, H. C. *et al.* Derivation and validation of machine learning approaches to predict acute kidney injury after cardiac surgery. *J. Clin. Med.* <https://doi.org/10.3390/jcm7100322> (2018).
31. Hermesen, M. *et al.* Deep learning-based histopathologic assessment of kidney tissue. *J. Am. Soc. Nephrol.* **30**, 1968–1979. <https://doi.org/10.1681/ASN.2019020144> (2019).
32. Boor, P. Artificial intelligence in nephropathology. *Nat. Rev. Nephrol.* **16**, 4–6. <https://doi.org/10.1038/s41581-019-0220-x> (2020).
33. Barbieri, C. *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. (Basel)* **5**, 28–33. <https://doi.org/10.1159/000493479> (2019).
34. Garcelon, N., Burgun, A., Salomon, R. & Neuraz, A. Electronic health records for the diagnosis of rare diseases. *Kidney Int.* **97**, 676–686. <https://doi.org/10.1016/j.kint.2019.11.037> (2020).
35. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18. <https://doi.org/10.1038/s41746-018-0029-1> (2018).
36. Akbilgic, O. *et al.* Machine learning to identify dialysis patients at high death risk. *Kidney Int. Rep.* **4**, 1219–1229. <https://doi.org/10.1016/j.ekir.2019.06.009> (2019).
37. Zhao, J., Gu, S. & McDermaid, A. Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression. *Math. Biosci.* **310**, 24–30. <https://doi.org/10.1016/j.mbs.2019.02.001> (2019).
38. Tangri, N., Ferguson, T. & Komenda, P. Pro: Risk scores for chronic kidney disease progression are robust, powerful and ready for implementation. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. Eur. Ren. Assoc.* **32**, 748–751. <https://doi.org/10.1093/ndt/gfx067> (2017).
39. Marks, A. *et al.* Looking to the future: Predicting renal replacement outcomes in a large community cohort with chronic kidney disease. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. Eur. Ren. Assoc.* **30**, 1507–1517. <https://doi.org/10.1093/ndt/gfv089> (2015).
40. Wang, Y. *et al.* Validation of the kidney failure risk equation for end-stage kidney disease in Southeast Asia. *BMC Nephrol.* **20**, 451. <https://doi.org/10.1186/s12882-019-1643-0> (2019).
41. Yamanouchi, M. *et al.* Value of adding the renal pathological score to the kidney failure risk equation in advanced diabetic nephropathy. *PLoS ONE* **13**, e0190930. <https://doi.org/10.1371/journal.pone.0190930> (2018).
42. Lin, C. C. *et al.* Development and validation of a risk prediction model for end-stage renal disease in patients with type 2 diabetes. *Sci. Rep.* **7**, 10177. <https://doi.org/10.1038/s41598-017-09243-9> (2017).
43. Pedersen, A. B. *et al.* Missing data and multiple imputation in clinical epidemiological research. *Clin. Epidemiol.* **9**, 157–166. <https://doi.org/10.2147/CLEP.S129785> (2017).
44. Sterne, J. A. *et al.* Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ (Clin. Res. Ed.)* **338**, b2393. <https://doi.org/10.1136/bmj.b2393> (2009).
45. Jakobsen, J. C., Gluud, C., Wetterslev, J. & Winkel, P. When and how should multiple imputation be used for handling missing data in randomised clinical trials: A practical guide with flowcharts. *BMC Med. Res. Methodol.* **17**, 162. <https://doi.org/10.1186/s12874-017-0442-1> (2017).
46. Karadaghy, O. A., Shew, M., New, J. & Bur, A. M. Development and assessment of a machine learning model to help predict survival among patients with oral squamous cell carcinoma. *JAMA Otolaryngol. Head Neck Surg.* **145**, 1115–1120. <https://doi.org/10.1001/jamaoto.2019.0981> (2019).
47. Bur, A. M. *et al.* Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. *Oral Oncol.* **92**, 20–25. <https://doi.org/10.1016/j.oraloncology.2019.03.011> (2019).
48. Zhang, Z. Missing data imputation: Focusing on single imputation. *Ann. Transl. Med.* **4**, 9. <https://doi.org/10.3978/j.issn.2305-5839.2015.12.38> (2016).
49. Graham, J. W. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.* **60**, 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530> (2009).
50. Klebanoff, M. A. & Cole, S. R. Use of multiple imputation in the epidemiologic literature. *Am. J. Epidemiol.* **168**, 355–357. <https://doi.org/10.1093/aje/kwn071> [American Journal of Epidemiology] (2008).
51. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: What is it and how does it work?. *Int. J. Methods Psychiatr. Res.* **20**, 40–49. <https://doi.org/10.1002/mpr.329> (2011).
52. Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T. & Schneeweiss, S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw. Open* **3**, e1918962. <https://doi.org/10.1001/jamanetworkopen.2019.18962> (2020).
53. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004> (2019).

54. Nusinovici, S. *et al.* Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* **122**, 56–69. <https://doi.org/10.1016/j.jclinepi.2020.03.002> (2020).
55. Nusinovici, S. *et al.* Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* <https://doi.org/10.1016/j.jclinepi.2020.03.002> (2020).

Author contributions

Q.B. was involved in the data collection, data analysis, and drafting the manuscript. C.S. was involved in data collection. W.T. conceptualized the idea, interpreted the results and wrote part of the draft. Y.L. conceptualized the idea, analyzed the data, performed all coding, evaluated all machine learning models, drafted and edited the manuscript.

Funding

This work was supported by PKU-Baidu Fund (2020BD030 to Wen Tang), and by fund from China International Medical Foundation (Z-2017-24-2037 to Wen Tang). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-12316-z>.

Correspondence and requests for materials should be addressed to W.T. or Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022