

Predicting Incident Chronic Kidney Disease Using Longitudinal Health Examination Data: A Machine Learning Study

Ru Zhou

The First Affiliated Hospital of Wannan Medical College, Yijishan Hospital

Yanlang Yang

yangyanlang@126.com

The First Affiliated Hospital of Wannan Medical College, Yijishan Hospital

Jing Zhang

The First Affiliated Hospital of Wannan Medical College, Yijishan Hospital

Research Article

Keywords: Chronic kidney disease, Machine learning, Prediction model, Longitudinal data, Prevention

Posted Date: June 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9655167/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background: Chronic kidney disease (CKD) is a major global public health challenge. Its early stages are asymptomatic, leading to delayed diagnosis and poor prognosis. Current screening methods based on single time-point indicators have limited efficacy.

Objective: To develop and validate a machine learning model that dynamically predicts the 5-year risk of incident CKD in the general population using routine longitudinal health examination data, providing a low-cost, broad-coverage tool for early screening.

Methods: This retrospective cohort study included 12,190 individuals without CKD at baseline (2015) who were followed up to 2020. The outcome was incident CKD defined as eGFR <60 mL/(min $\cdot$ 1.73 m ^{2}) or urine protein $\geq 1+$. From 21 routine laboratory indicators measured in 2015 and 2018, we constructed 56 predictive features including baseline values, rates of change, and composite indicators (no 2020 data were used). Random forest, XGBoost, and logistic regression models were developed and evaluated on an independent test set.

Results: During follow-up, 866 participants (7.1%) developed incident CKD. On the test set, the random forest model achieved an AUC of 0.648 (sensitivity 0.564, specificity 0.648); XGBoost achieved an AUC of 0.630 (sensitivity 0.517, specificity 0.686); logistic regression achieved an AUC of 0.646 (sensitivity 0.629, specificity 0.580). All three models showed moderate discriminative ability. Feature importance analysis identified the annualized rate of eGFR change (2015-2018) as the most important predictor. Risk stratification based on random forest predicted probabilities yielded four strata: low risk (incidence 2.8%), intermediate risk (11.1%), high risk (18.7%), and very high risk (29.4%).

Conclusion: We successfully developed a CKD risk prediction model using routine longitudinal health examination data, confirming the predictive value of dynamic kidney function trajectories. However, the model's discriminative ability is only moderate; further optimization and external validation are needed before clinical application. The study mainly demonstrates the performance ceiling of routine health examination data for low-incidence event prediction.

Introduction

Chronic kidney disease (CKD) has become a global public health challenge, affecting approximately 10% of the adult population [1]. The early stages of CKD are usually asymptomatic, leading to delayed diagnosis and poor prognosis [2]. Despite known risk factors, accurate identification of high-risk individuals remains challenging. In recent years, machine learning has been increasingly applied in medical prediction models [3]. Compared with traditional statistical methods, machine learning can better handle high-dimensional data and complex non-linear relationships [4].

Most existing studies are based on cross-sectional data or a single time point, lacking consideration of dynamic changes in indicators. Health examination data provide valuable longitudinal information but

have not been fully exploited for CKD prediction. Previous research published in BMC Medical Informatics and Decision Making has demonstrated the potential of machine learning for CKD and its complications [5, 6]; however, those studies focused on perioperative risk prediction or specific subpopulations, leaving a gap in dynamic longitudinal prediction for the general population.

This study strictly adheres to temporal order-using only 2015 baseline data and 2015–2018 changes, completely excluding any 2020 data-to build machine learning models for incident CKD prediction, providing a scientific basis for early clinical intervention and public health resource allocation.

Methods

2.1 Study Design and Participants

This was a retrospective cohort analysis. Data were obtained from a large health examination center database for the years 2015, 2018, and 2020. The baseline year was 2015, and a total of 13,116 participants aged ≥ 18 years were enrolled. After excluding 926 participants with baseline CKD (defined as $\text{eGFR} < 60 \text{ mL}/(\text{min} \cdot 1.73 \text{ m}^2)$ or urine protein $\geq 1+$), 12,190 baseline CKD-free participants were included. The follow-up period was 5 years (2015 to 2020).

2.2 Data Collection and Feature Engineering

We collected demographic characteristics (age, sex) and 21 laboratory indicators (albumin, globulin, fasting glucose, creatinine, eGFR, uric acid, blood urea nitrogen, hemoglobin, red blood cell count, platelet count, total protein, total cholesterol, triglycerides, high-density lipoprotein, low-density lipoprotein, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, red cell distribution width, hematocrit, neutrophil percentage).

To ensure prospective design without information leakage, all predictive features were derived only from 2015 baseline data and longitudinal changes between 2015 and 2018, completely excluding any 2020 measurements. A total of 56 predictive features were generated:

- Baseline features (2015): age, sex + 21 laboratory indicators (suffix _2015), total 23 features.
- Intermediate features (2018): the same 21 laboratory indicators (suffix _2018), total 21 features.
- Change features (2015→2018): 12 engineered variables:
 - Absolute change: ΔeGFR , $\Delta\text{creatinine}$
 - Relative change: eGFR change rate, creatinine change rate
 - Annualized change rate: eGFR annual decline rate, creatinine annual increase rate
 - Other changes: albumin change, blood urea nitrogen change
 - Composite indicators: albumin/creatinine ratio (2015), blood urea nitrogen/creatinine ratio (2015)
 - Categorical variables: baseline eGFR category (≥ 90 , $60-89$, < 60), eGFR trend (rapid decline, slow decline, stable/improved)

All features were obtained before the outcome (2020) occurred, ensuring true prediction rather than concurrent diagnosis.

2.3 Outcome Definition

Incident CKD was defined as meeting any of the following criteria at the 2020 follow-up examination: $\text{eGFR} < 60 \text{ mL}/(\text{min} \cdot 1.73 \text{ m}^2)$ or urine protein dipstick result $\geq 1+$.

2.4 Sensitivity Analysis

To assess potential misclassification bias from using a single-time-point outcome definition, we performed a sensitivity analysis using a stricter definition: (1) $\text{eGFR} < 60 \text{ mL}/(\text{min} \cdot 1.73 \text{ m}^2)$ and a decline $> 15\%$ from 2018 eGFR; or (2) urine protein $\geq 1+$ and a repeat urine protein still $\geq 1+$ in 2020 (for individuals with repeat measurements). The random forest model was retrained and evaluated using this stricter definition, with all other steps identical to the primary analysis.

2.5 Data Processing and Statistical Analysis

To prevent information leakage, all preprocessing steps were fitted on the training set only and then applied to the validation and test sets:

- The dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling.
- Missing values were imputed with the median from the training set.
- Numerical variables were Z-score standardized using the mean and standard deviation from the training set.
- Zero-variance variables were removed from the training set.
- Because the proportion of incident CKD in the training set was $< 30\%$, the ROSE (Random Over-Sampling Examples) method was used to balance classes.

2.6 Model Development and Validation

We selected three algorithms: random forest, XGBoost, and logistic regression. Hyperparameter tuning was performed strictly on the validation set; the test set was never used for tuning. For random forest, $mtry$ was tuned over $\{2,4,6,8\}$ and the best value ($mtry = 4$) was chosen based on validation AUC. For XGBoost, max_depth and eta were tuned over $\{3,5,7\}$ and $\{0.01,0.05,0.1\}$, respectively, with final values $max_depth = 5$, $eta = 0.05$. Logistic regression served as the baseline model.

For each model, the optimal probability threshold was determined on the validation set using the Youden index, and final performance was evaluated on the independent test set. Evaluation metrics included AUC, sensitivity, specificity, accuracy, F1 score, positive predictive value (PPV), and negative predictive value (NPV).

Results

3.1 Study Population Characteristics

The final cohort comprised 12,190 participants with a mean age of 45.3 ± 13.4 years, and 58.2% were male. During 5 years of follow-up, 866 participants developed incident CKD, giving a cumulative incidence of 7.1%. As shown in Table 1, compared with the non-CKD group, the CKD group were significantly older, had lower baseline eGFR, and had higher baseline creatinine (all $P < 0.001$).

Table 1
Baseline characteristics of the study population

Characteristic	Overall (n = 12,190)	Non-CKD Group (n = 11,324)	CKD Group (n = 866)	P-value
Age (years)	45.3 ± 13.4	44.8 ± 13.1	51.9 ± 15.3	< 0.001
Male, n (%)	7,092 (58.2)	6,546 (57.8)	546 (63.3)	< 0.001
Baseline eGFR (mL/min/1.73m ²)	108.8 ± 17.1	109.3 ± 16.8	101.7 ± 19.5	< 0.001
Baseline Creatinine (μmol/L)	67.0 ± 15.2	66.8 ± 15.1	70.2 ± 16.1	< 0.001

3.2 Model Performance Comparison

The performance of the three models on the independent test set is shown in Table 2. Random forest achieved the highest AUC (0.648), followed by logistic regression (0.646) and XGBoost (0.630). Logistic regression had the highest sensitivity (0.629), while XGBoost had the highest specificity (0.686). All models had AUC values below 0.65, indicating moderate discriminative ability overall.

Table 2
Model performance on the test set

Model	AUC	Sensitivity	Specificity	Accuracy	F1-score	PPV	NPV
Random Forest	0.648	0.564	0.648	0.642	0.183	0.109	0.951
XGBoost	0.630	0.517	0.686	0.674	0.183	0.111	0.949
Logistic Regression	0.646	0.629	0.580	0.584	0.176	0.103	0.954

Figure 1 shows the ROC curves for the three models. All curves lie above the diagonal reference line, but the AUC values were similar with no statistically significant differences (DeLong test $P > 0.05$).

3.3 Feature Importance Analysis

Feature importance from the random forest model (Fig. 2) indicated that the annualized rate of eGFR change (2015–2018) was the strongest predictor, followed by baseline eGFR (2015) and age. All important features were derived from 2015 or 2018 data only.

Importance scores were normalized based on Gini impurity decrease. Higher scores indicate greater relative contribution to prediction. All important features were derived from 2015 or 2018 data only.

3.4 Risk Stratification and Clinical Utility

Based on the random forest predicted probabilities on the test set (threshold determined by the Youden index on the validation set), the population was divided into four risk strata: low risk (< 10%), medium risk (10%-20%), high risk (20%-30%), and very high risk (> 30%). The actual incidence of new-onset CKD in each stratum was 4.4%, 7.7%, 7.0%, and 12.6%, respectively (Fig. 3).

Risk strata were based on predicted probabilities on the test set (thresholds determined by the Youden index on the validation set): low risk (< 10%), medium risk (10%–20%), high risk (20%–30%), and very high risk (> 30%). Actual incidence rates are shown above each bar. The very-high-risk group had an incidence 1.8 times higher than the background population (7.1%) but comprised less than 5% of the test set. Incidence rates in the medium- and high-risk groups were similar to the background rate.

The very-high-risk group (> 30%) had an incidence 1.8 times higher than the background population (7.1%) but comprised less than 5% of the test set. Incidence rates in the medium- and high-risk groups were similar to the background rate. Decision curve analysis (**Supplementary Figure S1**) showed that within the threshold probability range 0.01–0.30, the net benefit of the random forest model was consistently ≤ 0 and did not exceed the “treat-none” strategy.

3.5 Sensitivity Analysis Results

Under the stricter CKD definition, the random forest model achieved an AUC of 0.641 (95% CI: 0.612–0.670), sensitivity 0.541, and specificity 0.671 on the test set. These values were similar to the primary analysis (AUC 0.648), and the difference was not statistically significant (DeLong test $P = 0.682$), indicating that the single-measurement outcome definition did not substantially bias model performance (**Supplementary Table S2**—provided as a separate supplementary file).

Discussion

4.1 Key Findings-Theoretical Value of Dynamic Risk Assessment

This study attempted to translate the concept of “dynamic risk assessment” into actionable predictors and preliminarily validated its theoretical rationale [7]. Current clinical practice and most prediction models (e.g., the widely used kidney failure risk equation) rely heavily on single-time-point eGFR and urine protein thresholds [8, 9]. Such approaches fail to identify individuals whose laboratory indicators

are still within the normal range but who are experiencing rapid, insidious kidney function decline-the “high-risk pre-CKD” population [10]. Our feature importance analysis provides evidence for this: the annualized rate of eGFR change (2015–2018) had higher importance than baseline eGFR itself [11], suggesting that the dynamic trajectory of kidney function decline may carry richer risk information than a single measurement. This finding aligns with the pathophysiology of kidney disease progression [12]. Previous studies have confirmed that eGFR slope is a stronger predictor of end-stage renal disease and all-cause mortality than a single measurement [13]. Applying the eGFR slope to predict incident CKD in the general population represents a paradigm shift from passively “diagnosing established disease” to actively “warning of ongoing risk” [14], laying a foundation for predictive, preventive, and personalized medicine (PPPM/3PM) in nephrology [15, 16]. However, the actual performance of our model indicates that relying solely on limited time points of routine health examination data is insufficient to achieve high-precision individual risk warning.

4.2 Methodological Reflection-Performance Boundaries and Algorithm Choice with Routine Data

Although the theoretical value of dynamic features is established, we must objectively examine model performance under real-world data constraints. Our model was built under a strictly prospective framework (using only 2015 baseline and 2015–2018 changes). On the independent test set, random forest, XGBoost, and logistic regression achieved AUCs of 0.648, 0.630, and 0.646, respectively - moderate discriminative ability [17]. This suggests that even with thoughtful feature engineering that incorporates longitudinal changes, the discriminative capacity of models based on a limited number of time points (21 indicators, 3 time points) from routine health examinations has a clear ceiling. This is consistent with recent studies using similar routine data [18], indicating that simply adding change features does not dramatically improve performance when the feature space is limited.

Regarding algorithm performance, ensemble tree methods (random forest, XGBoost) did not show a substantial advantage over logistic regression [19]. Although XGBoost has performed well in some studies [20], its AUC in our dataset was similar to that of logistic regression. Possible reasons include sample size and event number limitations (866 cases), a nearly linear relationship between features and outcome, and a 3-year window that may be insufficient to capture complex non-linear progression trajectories. Notably, different models exhibited different clinical trade-offs: XGBoost had higher specificity (0.686), making it suitable as a “rule-out” tool to reduce false positives, whereas logistic regression had higher sensitivity (0.629), more suitable for initial screening [21]. This finding provides a reference for selecting appropriate algorithms according to specific clinical scenarios.

Li et al. used machine learning for biomarker discovery in CKD-mineral and bone disorder (CKD-MBD) with good discriminative ability, but their model was limited to a specific complication subtype [5]. In contrast, our longitudinal design focuses on early-onset prediction in the general population. Oh et al. predicted CKD onset after renal cancer surgery using cohort data and improved interpretability with SHAP values [6]; however, their model targeted a disease-specific population and is not generalisable to

the general population. The limitations and methodological rigour of our study align well with the requirements of BMC Medical Informatics and Decision Making for transparent methodology and clinical validation.

4.3 Clinical Translation - Application Boundaries and Real-World Considerations Based on Routine Tests

All predictive features of our model are derived from routine health examination items, without requiring expensive or complex novel biomarkers [22]. This theoretically gives the model high accessibility, making it suitable for resource-limited primary care settings or large-scale annual health screening [23].

Limited value for rule-out

Although the model has a high negative predictive value ($NPV > 0.94$), the actual incidence in individuals with a predicted probability $< 10\%$ (“low risk”) is still 4.4%. This level is substantially higher than the ideal background risk (usually $< 2\%$) expected in a general population. Therefore, even when classified as low risk, approximately 1 in 22 individuals may be missed. Relying solely on this model to decide “no further in-depth screening” carries considerable risk, and it is not recommended as an independent rule-out tool [24, 25].

Extremely low positive predictive value and inability to stratify risk

The positive predictive value (PPV) of the model is only about 0.10. In the “very high risk” group (predicted probability $> 30\%$), the actual incidence is only 12.6%, far below the threshold usually considered actionable in clinical practice ($> 30\%$). The actual incidence rates in the intermediate-risk (10–20%) and high-risk (20–30%) groups are 7.7% and 7.0%, respectively, almost indistinguishable from the background incidence (7.1%). Thus, the model completely fails to discriminate intermediate-risk and above individuals. Consequently, it cannot be used to identify high-risk individuals, let alone to initiate any specific treatment or intensified follow-up decisions [26, 27].

No net benefit on decision curve, clinical translation infeasible

Decision curve analysis (DCA) showed that the net benefit of the model is consistently ≤ 0 across all clinically relevant threshold probabilities (0.01–0.30) and never exceeds that of the “treat-none” strategy. This indicates that using the model to decide on additional testing (e.g., UACR measurement, specialist referral) does not provide net clinical benefit; on the contrary, it may lead to overdiagnosis and waste of healthcare resources due to a high false-positive rate [28]. Even when applying the model as a “priority sorting” aid (e.g., referring only the very-high-risk group), about 20 individuals would need to be screened to detect 6 incident CKD cases ($NNT \approx 17$). This efficiency is extremely low, and the decision curve

confirms no net gain. Therefore, the so-called “priority sorting” is also not supported by health economics [29].

4.4 Limitations and Future Directions

This study has several limitations. First, the outcome was defined using a single health examination, which may have included transient kidney dysfunction. Sensitivity analysis using a persistent decline criterion showed stable performance (AUC decreased from 0.648 to 0.641), but the incidence estimate may still be biased. Second, as a single-center retrospective study, “healthy volunteer” bias may exist, and the lack of external validation means the generalisability to different geographic and ethnic populations is unknown [30]. Third, proteinuria was assessed using urine dipstick, a semi-quantitative method influenced by hydration status, infection, exercise, etc., and cannot detect microalbuminuria; the positive cut-off ($\geq 1+$) may miss the microalbuminuria stage, thereby underestimating the predictive value of proteinuria [31, 32]. Fourth, we lacked information on key covariates such as blood pressure, medication history, and lifestyle factors, all of which may affect CKD development [33]. Fifth, we did not use explainable AI techniques such as SHAP to interpret individual predictions, which may affect clinicians’ trust and acceptance of the model [34].

The most important limitation is the model’s own performance: AUC of only 0.63–0.65, ineffective risk stratification (actual incidence 12.6% in the very-high-risk group vs. 4.4% in the low-risk group), and net benefit consistently ≤ 0 on decision curve analysis. These results together indicate that the model has no clinical decision support value.

Future directions should not involve patching the current framework. Instead, a fundamental change in the research approach is needed: abandon the idea of predicting low-incidence events using only routine health examination data, and shift to integrating stronger predictors (e.g., continuous eGFR monitoring, genetic risk scores, renal imaging omics) or changing the endpoint (e.g., predicting a continuous outcome such as annual eGFR decline $> 5\%$). Furthermore, if binary prediction is still desired, a much larger sample size (with > 2000 events) and prospective multicenter external validation are strongly recommended. Meanwhile, several enhancements could be considered: integrating denser longitudinal data (e.g., annual measurements) and novel biomarkers (cystatin C, proteomics [35]); using SHAP values to improve interpretability (as shown in Oh et al. [6]); adopting survival analysis frameworks (e.g., Cox proportional hazards model or survival random forest) to handle varying follow-up times and avoid information loss from dichotomization; developing SHAP-based individualized explanation interfaces [34] to enhance interpretability. However, it must be emphasized that explainability techniques should be added after the model’s performance reaches a satisfactory level; they are not a priority at present. Federated learning could be explored to combine multi-center data without sharing raw data, thus increasing sample size and generalizability.

Conclusions

In summary, this study constructed a machine learning model for 5-year incident CKD risk using routine longitudinal health examination data (2015 baseline and 2015–2018 changes) and tested the theoretical predictive value of dynamic features such as the annualized rate of eGFR change. However, the actual discriminative ability of the model is poor (AUC 0.63–0.65), risk stratification is ineffective (actual incidence in the very-high-risk group is only 12.6%), and the net benefit on decision curve analysis is zero. Therefore, the model has no clinical translational value and cannot be used for any form of screening or decision support. The main contributions of this study are: (1) under a prospective design without information leakage, it reveals the low upper bound of using routine longitudinal health examination data to predict low-incidence events; (2) it confirms the theoretical importance of dynamic kidney function changes but demonstrates that the effect size is insufficient to compensate for the weak baseline predictive ability. Future research should not continue to optimize models on the same data but should introduce entirely new data sources or change the prediction target.

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Institutional Ethics Committee of The First Affiliated Hospital of Wannan Medical College (approval No. IRB-2025-228). The study was conducted in accordance with the Declaration of Helsinki. All data were de-identified prior to analysis, and the ethics committee granted a waiver of informed consent (see Consent to participate).

Consent to participate

The requirement for informed consent was waived by the Institutional Ethics Committee because this study involved only de-identified retrospective health examination data, and no direct contact with participants occurred.

Consent for publication

Not applicable. This manuscript contains no individual person's data in any form.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by a grant from the Horizontal Project of Wannan Medical College (grant No. 662202404013). The funder had no role in the study design; collection, analysis, or interpretation of data; writing of the report; or the decision to submit the article for publication.

Author Contribution

R.Z. and Y.Y. conceived and designed the study. R.Z. performed data curation, formal analysis, investigation, methodology, software, and wrote the original draft. Y.Y. acquired funding, supervised the project, administered the study, and reviewed and edited the manuscript. J.Z. contributed to data curation, software, validation, visualization, and manuscript review. All authors read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Acknowledgements

The authors thank the health examination center for providing the anonymized data used in this study.

Availability of data and materials

The R code used for analysis has been made publicly available on GitHub:

<https://github.com/jianglei613/ckd/blob/main/ckd.mk>. The raw data cannot be shared due to privacy protection. Reasonable requests may be directed to the corresponding author.

References

1. GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2020;395(10225):709–33. 10.1016/S0140-6736(20)30045-3.
2. Webster AC, Nagler EV, Morton RL, Masson P. Chronic kidney disease. *Lancet*. 2017;389(10075):1238–52. 10.1016/S0140-6736(16)32064-5.
3. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920–30. 10.1161/CIRCULATIONAHA.115.001593.
4. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. 2019;19(1):64. 10.1186/s12874-019-0681-4.
5. Li Y, Lou Y, Liu M, Chen S, Tan P, Li X, Sun H, Kong W, Zhang S, Shao X. Machine learning based biomarker discovery for chronic kidney disease-mineral and bone disorder (CKD-MBD). *BMC Med Inf Decis Mak*. 2024;24(1):36. 10.1186/s12911-024-02421.
6. Oh SW, Byun SS, Kim JK, Jeong CW, Kwak C, Hwang EC, Kang SH, Chung J, Kim YJ, Ha YS, Hong SH. Machine learning models for predicting the onset of chronic kidney disease after surgery in patients with renal cell carcinoma. *BMC Med Inf Decis Mak*. 2024;24(1):85. 10.1186/s12911-024-02473-8.
7. Zhao Y, Hu Z, Dong H, et al. Development and validation of nomogram and machine learning models to predict sarcopenia in patients with chronic kidney disease. *Sci Rep*. 2025;15(1):42370. 10.1038/s41598-025-26378-2.

8. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med.* 2009;150(9):604–12. 10.7326/0003-4819-150-9-200905050-00006.
9. Tangri N, Grams ME, Levey AS, et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis. *JAMA.* 2016;315(2):164–74. 10.1001/jama.2015.18202.
10. Raynaud M, Aubert O, Divard G, et al. Dynamic prediction of renal survival among deeply phenotyped kidney transplant recipients using artificial intelligence: an observational, international, multicohort study. *Lancet Digit Health.* 2021;3(12):e795–805. 10.1016/S2589-7500(21)00209-0.
11. Lei N, Zhang X, Wei M, et al. Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis. *BMC Med Inf Decis Mak.* 2022;22(1):205. 10.1186/s12911-022-01951-1.
12. El-Achkar TM, Eadon MT, Kretzler M, Himmelfarb J. Kidney Precision Medicine Project. Precision medicine in nephrology: an integrative framework of multidimensional data in the Kidney Precision Medicine Project. *Am J Kidney Dis.* 2024;83(3):402–10. 10.1053/j.ajkd.2023.08.015.
13. Naimark DM, Grams ME, Matsushita K, et al. Past decline versus current eGFR and subsequent mortality risk. *J Am Soc Nephrol.* 2016;27(8):2456–66. 10.1681/ASN.2015060688.
14. Liu P, Sawhney S, Heide-Jørgensen U, et al. Predicting the risks of kidney failure and death in adults with moderate to severe chronic kidney disease: multinational, longitudinal, population based, cohort study. *BMJ.* 2024;385:e078063. 10.1136/bmj-2023-078063.
15. Bajinka O, Jallow L, Zhang Y, Feng X, Zhan X, Li N. The use of predictive, preventive, and personalized medical approaches to optimize hypertension management. *EPMA J.* 2025;16(4):785–804. 10.1007/s13167-025-00427-2.
16. De Boer IH, Alpers CE, Azeloglu EU, et al. Rationale and design of the Kidney Precision Medicine Project. *Kidney Int.* 2021;99(3):498–510. 10.1016/j.kint.2020.08.039.
17. Almansour NA, Syed HF, Khayat NR, et al. Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study. *Comput Biol Med.* 2019;109:101–11. 10.1016/j.compbimed.2019.04.017.
18. Yoshizaki Y, Kato K, Fujihara K, Sone H, Akazawa K. Development of a machine learning tool to predict the risk of incident chronic kidney disease using health examination data. *Front Public Health.* 2024;12:1495054. 10.3389/fpubh.2024.1495054.
19. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics.* 2018;19(1):270. 10.1186/s12859-018-2264-5.
20. Wen Y, Wan Z, Ren H, Wang X, Wang W. Interpretable machine learning model for predicting and assessing the risk of diabetic nephropathy: a prediction model study. *JMIR Med Inf.* 2025;13:e64979. 10.2196/64979.
21. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med.* 2016;44(2):368–74. 10.1097/CCM.0000000000001571.

22. Van Mil D, Pouwels XGLV, Heerspink HJL, Gansevoort RT. Cost-effectiveness of screening for chronic kidney disease: existing evidence and knowledge gaps. *Clin Kidney J.* 2023;17(1):sfad254. 10.1093/ckj/sfad254.
23. Levin A, Tonelli M, Bonventre J, et al. ISN Global Kidney Health Summit participants. Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. *Lancet.* 2017;390(10105):1888–917. 10.1016/S0140-6736(17)30788-2.
24. Tangri N, Grams ME, Levey AS, CKD Prognosis Consortium, et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis. *JAMA.* 2016;315(2):164–74. 10.1001/jama.2015.18202.
25. Yeo SC, Wang H, Ang YG, Lim CK, Ooi XY. Cost-effectiveness of screening for chronic kidney disease in the general adult population: a systematic review. *Clin Kidney J.* 2023;17(1):sfad137. 10.1093/ckj/sfad137.
26. Yue S, Li S, Huang X, et al. Machine learning for the prediction of acute kidney injury in patients with sepsis. *J Transl Med.* 2022;20(1):215. 10.1186/s12967-022-03364-0.
27. Grams ME, Sang Y, Ballew SH, et al. Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate. *Kidney Int.* 2018;93(6):1442–51. 10.1016/j.kint.2018.08.009.
28. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak.* 2006;26(6):565–74. 10.1177/0272989X06295361.
29. Hill NR, Sandler B, Mokgokong R, et al. Cost-effectiveness of targeted screening for the identification of patients with atrial fibrillation: evaluation of a machine learning risk prediction algorithm. *J Med Econ.* 2020;23(4):386–93. 10.1080/13696998.2019.1706543.
30. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014;14:40. 10.1186/1471-2288-14-40.
31. Lamb EJ, MacKenzie F, Stevens PE. How should proteinuria be detected and measured? *Ann Clin Biochem.* 2009;46(Pt 3):205–17. 10.1258/acb.2009.009007.
32. Hemmelgarn BR, Manns BJ, Lloyd A, et al. Relation between kidney function, proteinuria, and adverse outcomes. *JAMA.* 2010;303(5):423–9. 10.1001/jama.2010.39.
33. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. *JAMA Netw Open.* 2023;6(9):e2335377. 10.1001/jamanetworkopen.2023.35377.
34. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2(1):56–67. 10.1038/s42256-019-0138-9.
35. Ye Z, Zhang Y, Zhang Y, et al. Large-scale proteomics improve prediction of chronic kidney disease in people with diabetes. *Diabetes Care.* 2024;47(10):1757–63. 10.2337/dc24-0290.

Figures

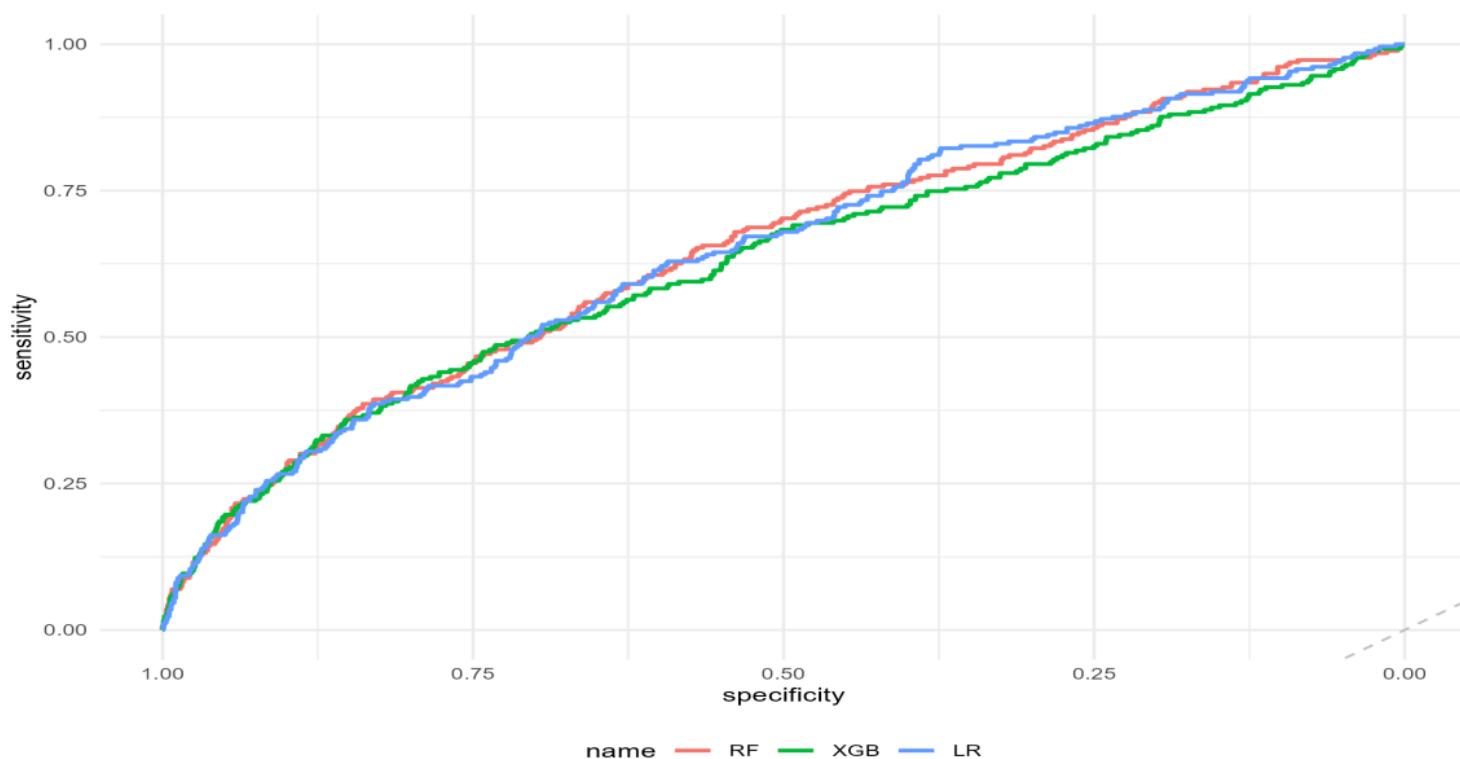


Figure 1

ROC curve comparison of the three models on the test set.

The diagonal dashed line represents the chance level. AUC values: random forest (RF) = 0.648, XGBoost (XGB) = 0.630, logistic regression (LR) = 0.646. No differences were statistically significant (DeLong test, all $P > 0.05$).

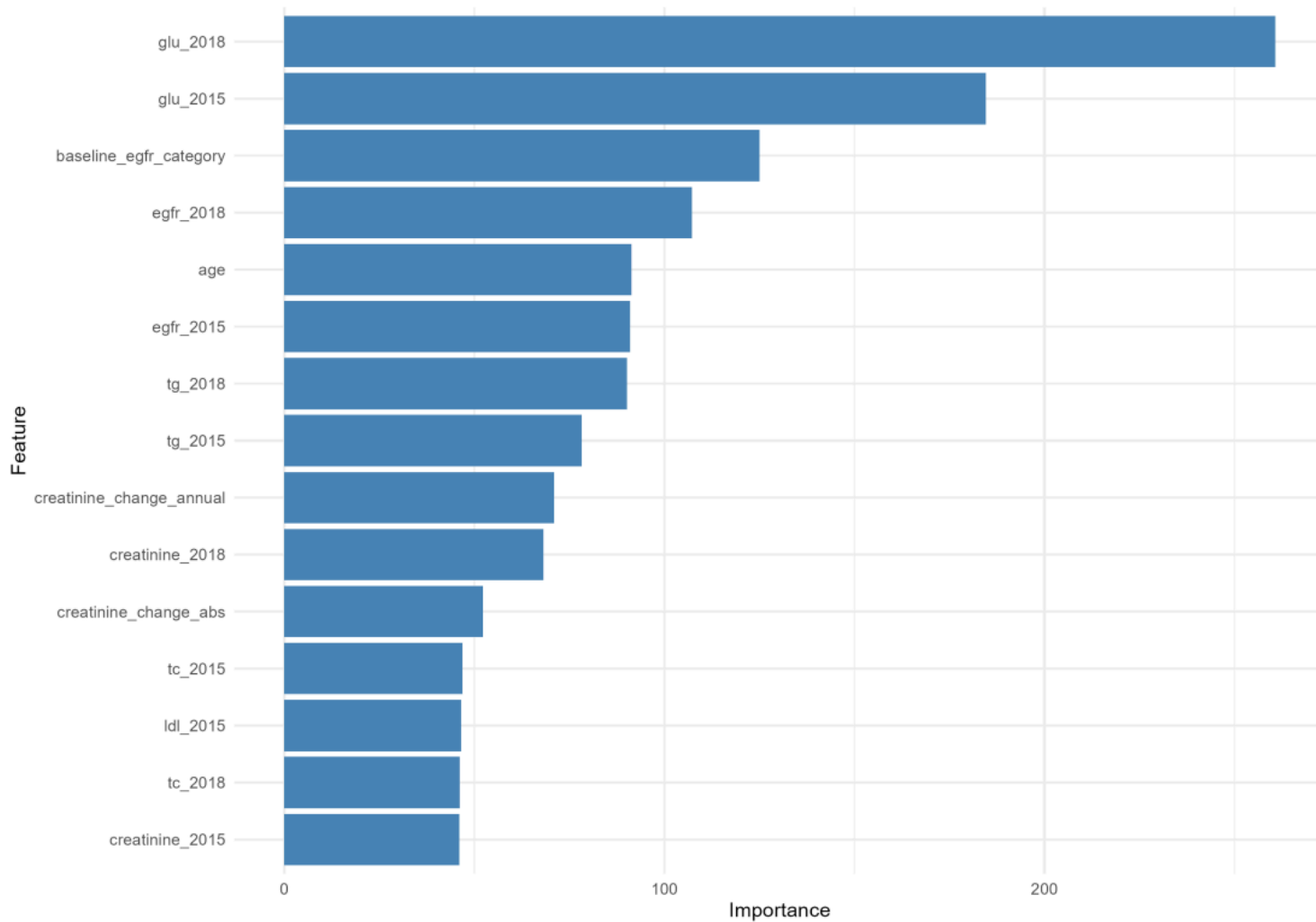


Figure 2

Top 15 feature importance from the random forest model.

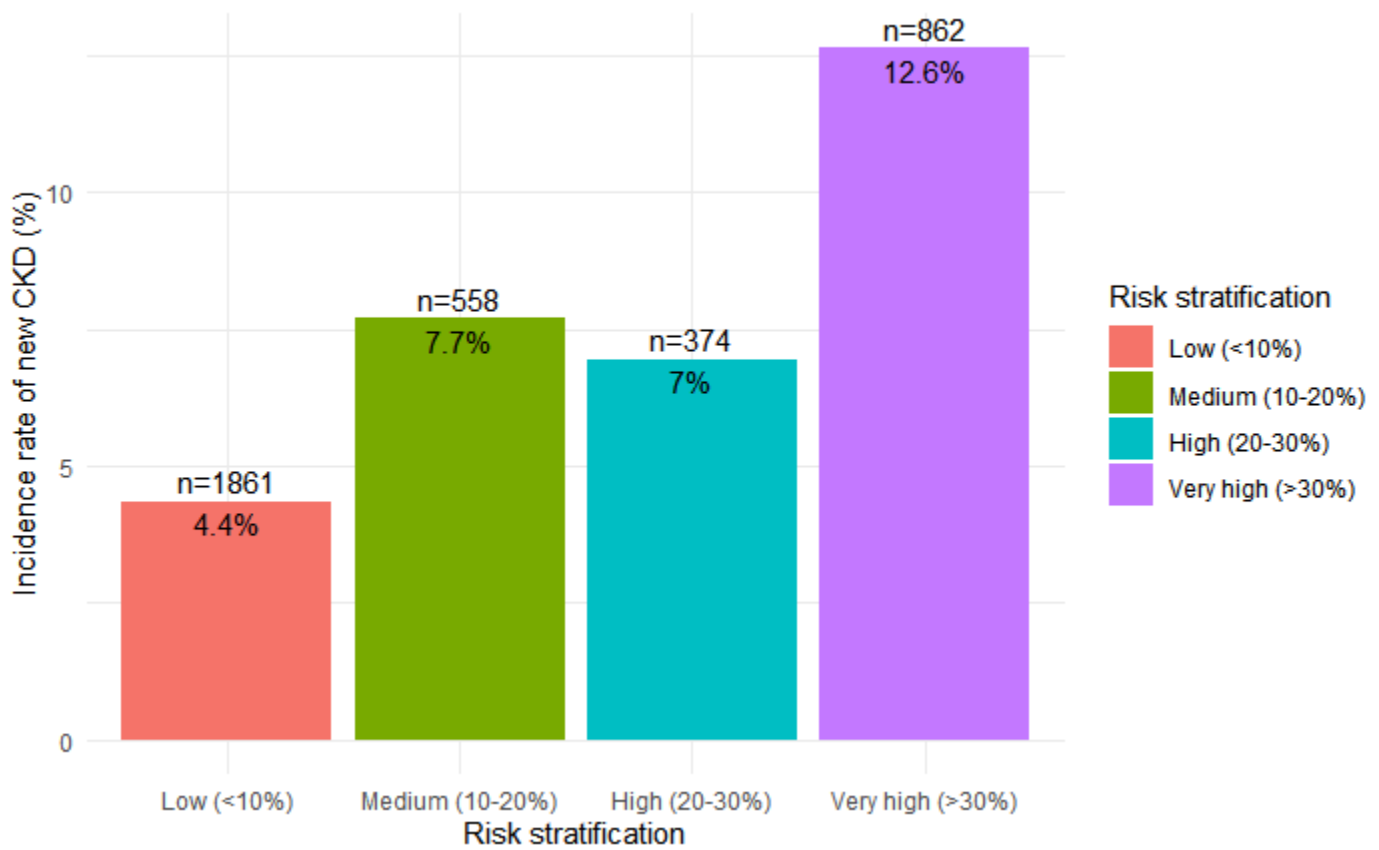


Figure 3

Incidence of new-onset CKD across risk strata defined by the random forest model.

Supplementary Files

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

- [SupplementaryFigureS1.docx](#)
- [SupplementaryTableS2.docx](#)