Prediction of acute and chronic kidney diseases during the post-covid-19 pandemic with machine learning models: utilizing national electronic health records in the US

Yue Zhang,[a] *Nasrollah Ghahramani,*[a,b] *Runjia Li,*[c] *Vernon M. Chinchilli,*[a] *and Djibril M. Ba*[a,*]

[a]Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA, USA
[b]Department of Medicine, Penn State College of Medicine, Hershey, PA, USA
[c]Department of Biostatistics, University of Pittsburgh School of Public Health, Pittsburgh, PA, USA

eBioMedicine
2025;115: 105726
Published Online xxx
https://doi.org/10.1016/j.ebiom.2025.105726

Summary

Background COVID-19 has been linked to acute kidney injury (AKI) and chronic kidney disease (CKD), but machine learning (ML) models predicting these risks post-pandemic have been absent. We aimed to use large electronic health records (EHR) and ML algorithms to predict the incidence of AKI and CKD during the post-pandemic period, assess the necessity of including COVID-19 infection history as a predictor, and develop a practical webpage application for clinical use.

Methods National EHR data from TriNetX, emulating a prospective cohort of 104,565 patients from 07/01/2022 to 03/31/2024, were used. A total of 69 baseline variables were included, with demographics, comorbidities, lab test results, vital signs, medication histories, hospitalization visits, and COVID-19-related variables. Prediction windows of 1 month and 1 year were defined to assess AKI and CKD incidence. Eight machine learning models, primarily including extreme gradient boosting (XGBoost), neural network, and random forest (RF), were applied. Cross-validation and model tuning were conducted during the training process. Model performance was evaluated using six metrics, including the area under the receiver-operating-characteristic curve (AUROC). A combination of model-driven, data-driven, and clinical-driven methods was employed to identify the final models. An application with the final models was built using the R Shiny framework.

Findings The final models, incorporating 9 variables—primarily including eGFR, inpatient visit number, and number of COVID-19 infections—were selected. XGBoost demonstrated the best performance for predicting the incidence of AKI in 1 month (AUROC = 0.803), AKI in 1 year (AUROC = 0.799), and CKD in 1 year (AUROC = 0.894). Random Forest (RF) was selected for predicting the incidence of CKD in 1 month (AUROC = 0.896). A comparison of AUROC with and without COVID-19 infection confirmed its importance as a critical predictor in the model. The final models were translated into a convenient tool to facilitate their use in clinical settings.

Interpretation Our study demonstrates the applicability of using large national EHR data in developing high-performance machine learning models to predict AKI and CKD risks in the post-COVID-19 period. Incorporating the number of COVID-19 infections in the past year showed improved prediction performance and should be considered in future models for kidney disease prediction. A user-friendly application was created to support clinicians in risk assessment and surveillance.

Funding Artificial Intelligence and Biomedical Informatics Pilot Funding, Penn State College of Medicine.

Keywords: COVID-19; Kidney diseases; Machine learning; Real world data; Electronic health records

Introduction

Chronic Kidney Disease (CKD) involves a gradual loss of kidney function and is common among US adults, with 15% of US adults or 37 million people estimated to have CKD in 2021.[1] Despite its high prevalence, up to 90% of individuals with CKD may be unaware of their condition.[2] CKD often remains asymptomatic in its early stages, leading to delayed diagnosis until advanced

*Corresponding author. Department of Public Health Sciences, Penn State College of Medicine, 90 Hope Drive, Hershey, 17033, PA, USA.
E-mail address: djibrilba@pennstatehealth.psu.edu (D.M. Ba).

Research in context

Evidence before this study

Epidemiological research links COVID-19 infection to increased kidney disease incidence. While machine learning (ML) algorithms have recently been applied to predict kidney diseases using clinical data, their use for this purpose in the post-COVID-19 era remains unexplored, especially with large national electronic health records (EHR). We searched PubMed on March 1, 2025, with no date or language restrictions for publications, using Medical Subject Headings (Mesh) "COVID-19" [Mesh] AND "Kidney Diseases" [Mesh] AND "Machine Learning" [Mesh], which yield nine results. Of these, eight studies focused on COVID-19-infected individuals from 2020 to 2021, with small clinical datasets, while one study with data for a short period (December 2022 to January 2023). Specifically, five studies predicted mortality, three focused on dialysis or kidney transplantation, and one aimed to predict acute kidney disease over a short term using data from a single clinical centre. Therefore, no studies have employed machine learning in predicting chronic kidney disease (CKD) and acute kidney injury (AKI) in the general population during the post-pandemic period using national EHR data.

Added value of this study

Using large national EHR data from 07/01/2022 to 03/31/2024 in the United States, we emulated a prospective cohort study to accurately identify baseline covariates (predictors) for predicting kidney disease. We developed high-performance models to predict CKD and AKI incidence within one month and one year following both inpatient and outpatient visits during the post-COVID-19 period. Our findings highlight that age, sex, BMI, diastolic and systolic blood pressure, eGFR, BUN, and a one-year history of COVID-19 infections and inpatient admissions are essential predictors for model accuracy. Additionally, we created a user-friendly clinical application to aid clinicians in assessing individual patient risk and enhancing timely surveillance.

Implications of all the available evidence

Our study has three important implications for the field: (1) Clinical Implications: Our findings highlight the necessity of including COVID-19 infection history in future ML models for predicting kidney diseases. (2) Methodological Implications: This study serves as an example of how to emulate a prospective cohort study using EHR data to better define predictors. It can aid ML experts in obtaining accurate input data and assist epidemiologists and clinicians in developing advanced ML models. (3) ML Application Implications: We provide a solution for integrating ML models into real-world clinical settings. The web-based application we developed utilises only nine variables, yet achieves strong predictive performance, enhancing accessibility and practicality across various clinical environments.

stages, such as end-stage kidney disease (ESRD), which require interventions like dialysis or transplantation.[1] Acute kidney injury, characterised by a sudden decline in kidney function, affects approximately 20% of hospitalised patients in the US.[3,4] Delayed diagnosis of AKI is associated with an increased risk of in-hospital mortality, prolonged hospital stays, and a greater likelihood of progression to CKD or ESRD.[4,5] Early diagnosis is crucial to improving outcomes and preventing AKI and CKD.[6] Therefore, it is essential to develop strategies to predict the risk of incident CKD and AKI, facilitate early diagnosis, and enable timely intervention to manage disease progression effectively.

With the rapid advancement of machine learning (ML) algorithms, these techniques have been increasingly applied to predict the risk of various diseases, including kidney diseases.[7] However, current ML studies have notable limitations: (1) During the post-pandemic periods, COVID-19 has emerged as a significant risk factor for kidney diseases,[8–12] but most ML algorithms do not account for this factor, as their training data predates 2020; (2) ML prediction models utilizing electronic health records (EHRs) often lack the detailed cohort construction and variable identification as done using traditional epidemiological study designs such as cohort studies[13,14]; (3) Studies primarily conducted by experts in computer science tend to emphasise ML algorithms while underestimating the importance of clinically meaningful variables and real-world application[13,15,16]; (4) Many ML models are trained on limited sample sizes from longitudinal cohorts or clinical trials, which may reduce generalizability[17–19]; (5) While several studies have demonstrated the effectiveness of ML in predicting kidney diseases, few have translated these findings into practical solutions for clinical application.

In this study, we aimed to utilise national EHR data and ML algorithms, with emulation of a prospective cohort study, to predict the risk of AKI and CKD in both the short term and long term during the post-COVID-19 pandemic. Additionally, we aimed to assess the necessity of including COVID-19 as a predictor, and develop a web-based application to assist clinicians in applying the ML models we developed in clinical settings.

Methods

In our study, we focused on two processes: learning and prediction. In the learning phase, we utilised electronic health records data to build the cohort and addressed missing data and class imbalance issues. We then developed full training models, conducted feature (variable) selection, and performed reduced model selection, ultimately deriving ML models for four distinct outcomes. In the prediction phase, we developed a

prediction application that incorporated the four ML models. This application is designed to assist physicians in predicting individual risks of kidney diseases and providing timely surveillance (Fig. 1). This study followed TRIPOD + AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis with Artificial Intelligence) guideline.[20]

Ethics

Our study performed secondary data analysis of fully de-identified electronic health record data, which does not meet the definition of human subject research as defined in 45 CFR 46.102(e) and/or (l), and the Institutional Review Board at the researchers' institution does not consider this to be human subject research. Therefore, further human subjects' approval was not necessary for the present study.

Study population

National electronic health records from the TriNetX Research Network were used for this study, covering the period from 01/01/2020 to 03/31/2024 in the United States. TriNetX is a federated, multi-institutional health research network that provides access to electronic medical records—including diagnoses, procedures, medications, laboratory test results, and vital signs—from approximately 250 million patients across 120 healthcare organizations.[21]

We initially requested 1.5 million random patients aged over 18 from TriNetX, and received 1,312,610 patients as our total sample size. After excluding patients with (1) missing demographic information (sex, race, or year of birth), (2) locations outside the United States, (3) encounter histories only before 07/01/2022, and (4) history of kidney diseases before the index dates, we identified 311,364 eligible patients with partially missing lab test results. To ensure the accuracy of creatinine values, a critical variable related to kidney disease, we further excluded patients with missing creatinine lab test results, resulting in 104,565 eligible patients for machine learning model training and testing (Fig. 2a).

Study design and data assessment

In our study, we emulated a prospective cohort design with a recruitment start date of 07/01/2022 and a follow-up end date of 03/31/2024, which defined the cohort follow-up period (Fig. 2b). The first hospital visit (either inpatient or outpatient) after the study start date was defined as the index date for each patient. We defined two prediction windows starting the index date: 1 month and 1 year. During each prediction window, we assessed

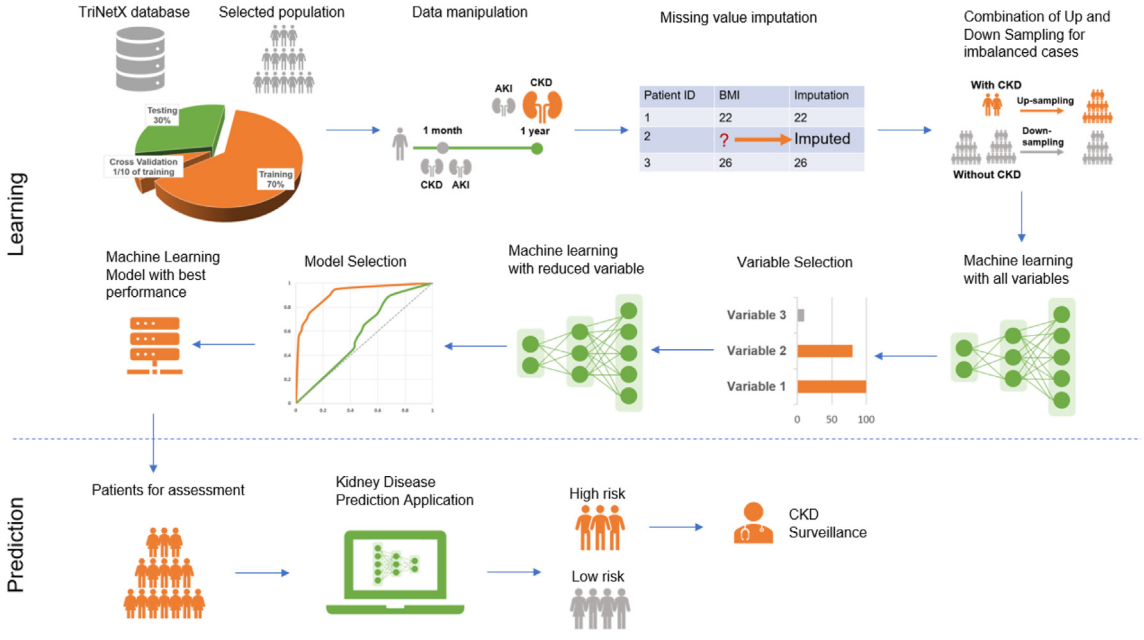

Fig. 1: Overall machine learning and prediction strategy flow (Example: Incidence of chronic kidney disease in 1 year). The figure illustrates the overall process: (1) Data acquired from TriNetX datasets; (2) Data manipulated to derive four outcomes: incidence of AKI at 1 month and 1 year, and CKD at 1 month and 1 year; (3) Missing data imputed; (4) Imbalance addressed through up-sampling and down-sampling; (5) Full model training and 10-fold cross-validation; (6)Variable selection based on variable importance score; (7) Training and cross-validation of the reduced model; (8) Final models selected with the best performance; (9) Kidney disease prediction application built to enhance model utility; (10) Individual patient risk assessed by inputting corresponding values, with high or low risk predictions; (11) Physicians can provide timely surveillance for high-risk patients. Note: CKD: Chronic Kidney Diseases; AKI: Acute Kidney Injury.

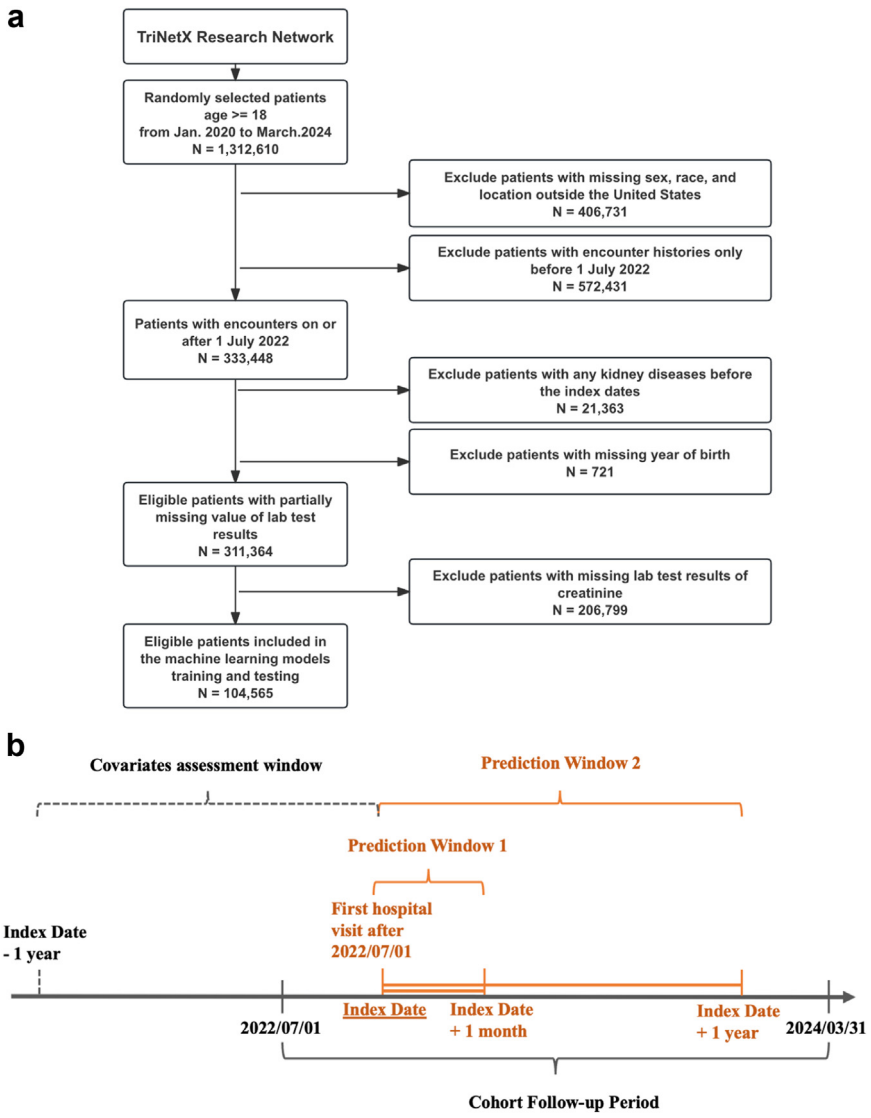

Fig. 2: a. Patient selection flow for eligible patients included in the machine learning models from the TriNetX Research Network. b, Timeline of cohort construction, index date definition, prediction windows, and covariate assessment period.

two separate kidney outcomes: (1) AKI, identified by International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) code N17, and (2) chronic kidney disease (CKD), identified by ICD-10-CM code N18.

At the index date, we collected demographic variables including age, sex (male, female), race (White, Black or African American, Asian, others), and location (Midwest, Northeast, South, West), which are recorded by TriNetX. The baseline covariates assessment window was defined as 1 year prior to the index date, during which we collected the following: (1) Comorbidities: 29 comorbidities, defined by ICD-10-CM codes (full list in Supplementary Note 1); (2) Medications: Histories of 13 medication uses, defined by RxNorm codes[22] (full list in Supplementary Note 2); (3) COVID-19 related covariates: Including the number of COVID-19 infections (re-infections), Long COVID diagnoses, and number of COVID-19 vaccination (identified codes in Supplementary Note 1); (4) Hospitalization related variables: Including number of inpatient visits, outpatient visits, and ICU visits (identified codes in Supplementary Note 1).

Lab test results and vital signs were assessed by Logical Observation Identifiers Names and Codes (LOINC), within 13 months before the index date to capture data from annual health check-ups. We applied Winsorization methods with 1% and 99% limits to address extreme values. For patients with only one

record of a lab test or vital sign, that single value was used. For patients with multiple records, the mean value was calculated to mitigate the impact of any extreme values.[23] We initially collected 14 lab test results and 4 vital signs, excluding those with more than 35% missing data. For our study population, we excluded individuals with missing creatinine values. To ensure the accuracy of estimated glomerular filtration rate (eGFR) and creatinine clearance, we calculated eGFR using the CKD-EPI 2021 equations[24] and creatinine clearance using the Cockcroft-Gault formula,[25] rather than extracting these values from datasets. Body mass index (BMI) was calculated based on height and weight measurements, and these two variables were replaced by BMI in the machine learning training. The complete list of lab tests and vital signs, along with LOINC, units, and missing data proportions, is provided in Supplementary Note 3.

Data processing

The data from the derivation cohort were randomly divided, with 70% utilised for training and 30% for testing. Additionally, one-tenth of the training dataset was randomly selected for cross-validation. As a result, the training and validation datasets maintained the same proportion of positive samples.

Within the training and testing sets, missing data imputation was performed. There are several methods for multiple missing data imputation, and each method may perform differently depending on the dataset. To select the most suitable imputation method for our dataset, we simulated a pseudo-dataset and evaluated the performance of each imputation method using the root mean square error (RMSE).[26,27] The pseudo-dataset was created with two steps: (1) We selected 31,524 patients from the entire dataset who had no missing values across all variables; (2) We then randomly assigned missing values to 5.51% of the Blood Urea Nitrogen (BUN) and 30.81% of the Creatinine Clearance (CrCl), mirroring the true missing proportions in the entire dataset. We compared three different imputation methods (with five settings), including (1) Predictive mean matching (PMM)[28,29]; (2) Classification And Regression Tree (CART)[30]; (3) XGBoost, with three settings: default settings, specified metrics, or cross-validation.[31] XGBoost with default settings showed the best performance, with the lowest RMSE in both the low missing proportion group (BUN) and the high missing proportion group (CrCl) (Supplementary Figure S1). Consequently, XGBoost with default settings was selected for further missing data imputation in this study.

The training datasets for the four kidney outcomes (incidence of AKI in 1 month, AKI in 1 year, CKD in 1 month, and CKD in 1 year) were imbalanced between the case and non-case groups. To avoid poor performance for the minority class (kidney outcomes) compared to the dominant class (non-cases group), we balanced the number of cases and non-cases using both up-sampling and down-sampling methods for training datasets. For up-sampling, we applied synthetic minority over-sampling technique (SMOTE).[32] For down-sampling, we randomly selected patients from the non-cases group. We did not apply the up and down sampling to testing datasets.

Model development and testing

A total of 69 variables from the training datasets were used to develop prediction models for the incidence of AKI and CKD in 1 month and 1 year, respectively. Eight machine learning models were employed: adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), neural network (NN), support vector machine (SVM), decision tree (DT), gradient boosting machine (GBM), logistic regression (LR), and random forest (RF). A parameter grid was constructed to train each model using 10-fold cross-validation with 10 repeats. Model tuning was performed through an automatic grid search, testing 10 different values for each algorithm parameter.[23] For each model, we used the test datasets to evaluate performance. The evaluation metrics included the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, accuracy, precision, and F1 score.

Feature selection and final model determination

We combined model-driven, data-driven, and clinical-driven methods to identify appropriate features that ensure both robust model performance and clinical relevance: (1) Model-driven methods: Scaled variable importance (VI) scores, ranging from 0 to 100, were calculated for each model using the "caret" package in R.[33] Higher scores indicated greater contributions of the variables. (2) Data-driven methods: Spearman correlation was used to examine the relationships between pairs of variables, with a particular focus on reducing multicollinearity. When two features were highly correlated, the one less associated with the outcome was removed from the dataset.[18] (3) Clinical-driven methods: Clinically meaningful variables related to AKI or CKD were identified based on prior studies and expert input from physicians. An example of how we utilised these three methods to identify feature selection for eGFR and CrCL is provided in Supplementary Note 4.

Using the feature selection strategy above, we reduced the prediction models for AKI in 1 month, CKD in 1 month, AKI in 1 year, and CKD in 1 year from 69 to 3 features, based on VI rankings and clinical relevance for each outcome. The final models with the best predictive performance during the feature reduction process were selected separately for further analysis of the four outcomes. AUROCs were used to assess predictive performance, with the optimal cutoff value determined by maximizing the Youden index (sensitivity + specificity − 1).[34] Features of the selected machine learning model were gradually reduced

until a significant decrease in AUROC was observed. The final reduced models, which maintained relatively good performance and included a reasonable number of features, were chosen to enhance practical applicability for physicians.

Webpage deployment tool building

With the four final models selected for predicting the incidence risk of AKI/CKD in 1 month/year, we developed a web application to enhance the utility in clinical settings. The application was built using the R Shiny Web Application Framework[35] and integrated all four predictive models. When the values for each required feature are correctly entered, the application simultaneously provides predicted low/high risk assessments for AKI in 1 month, CKD in 1 month, AKI in 1 year, and CKD in 1 year.

Statistical and coding statement

Demographics, lab test results, vital signs, comorbidities, medications, COVID-19-related variables, and hospitalization histories were summarised across the non-case group, AKI in 1 month, AKI in 1 year, CKD in 1 month, and CKD in 1 year groups. Continuous variables were presented as mean (SD) values, and categorical variables were summarised as number and percentages. The DeLong test was employed to compare differences between the two AUROCs when necessary.[36] Data manipulation was conducted with SAS software version 9.4 (SAS Institute, Cary, North Carolina), and machine learning was performed with R software version 3.6.2 (R Foundation for Statistical Computing, Vienna, Austria). A two-tailed P value < 0.05 was considered statistically significant.

Role of funders

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

A total of 104,565 patients were included after applying the exclusion criteria. Among them, 101,870 (age (mean, [SD]), 52.6 [17.8]; female (%), 57.4%) patients had no incidence of kidney outcomes during follow-up, 366 (age, 64.1 [15.2]; female, 43.7%) patients developed AKI within 1 month, 332 (age, 69.6 [12.7]; female, 46.7%) patients developed CKD within 1 month, 1475 (age, 64.6 [15.3]; female, 46.2%) patients developed AKI within 1 year, and 1642 (age, 69.1 [13.0]; female, 53.8%) patients developed CKD within 1 year. Patients with CKD had the lowest eGFR in both the 1-month prediction window (mL/min/1.73 m^2, mean [SD], 51.8 [20.1]) and the 1-year prediction window (58.6 [20.3]), compared to those with AKI in the 1-month prediction window (65.0 [25.5]), the 1-year prediction window (72.2 [23.8]), and the non-cases group (91.9 [20.9]). 69 variables were extracted from the EHR, and details of all variable lists and characteristics are provided in Supplementary Table S1.

Eight machine learning algorithms were applied to the training datasets using all 69 variables for full model training, separately for AKI and CKD within both the 1-month and 1-year prediction windows, resulting in a total of 32 models. The performance metrics for these models are summarised in Supplementary Table S2. Assessed with scaled variable importance (VI) scores, the most important variable for predicting AKI in both the 1-month and 1-year prediction windows was the number of inpatient visits in the previous year (VI for 1-month window, 100; VI for 1-year window, 100), followed by eGFR (29.0; 52.8) and BUN (15.6; 33.8). For CKD in both the 1-month and 1-year prediction windows, the top three variables were eGFR (100; 100), creatinine clearance (CrCL) (59.8; 35.4), and number of inpatient visits (45.0; 40.4). In the evaluated subgroups categorised by variables, besides lab test results and hospitalization-related variables, the following factors also demonstrated importance within their respective categories: BMI, systolic blood pressure, diastolic blood pressure, hypertension, age, sex, and the number of COVID-19 infections in the previous year (Fig. 3).

Based on the model-driven, data-driven, and clinical-driven methods, we reduced the number of variables (features) from 69 to 3 (with 69, 31, 15, 12, 9, and 3) and evaluated the ML performance separately for each set. The list of features and performance for the top 4 ML models are shown in Fig. 4. We first compared each model with AUROC and found that models using 9 features maintained or even improved AUROC compared to those with more than 9 features. Specifically, at 9 features, the XGBoost models showed the best performance for predicting AKI in 1 year, AKI in 1 month, and CKD in 1 year. The Random Forest model was the best for predicting CKD in 1 month with 9 features as well. The models with the best AUROC at 9 features were then selected to assess accuracy, F1 score, sensitivity, and specificity. These models exhibited promising performance, showing no significant decline and even outperforming some models with more than 9 features across all evaluated metrics. Therefore, we selected 9 features for the final models, including 5 patient demographic information (age, sex, BMI, diastolic blood pressure, and systolic blood pressure), 2 lab test results (eGFR (mL/min/1.73 m^2), and BUN (mg/dL)), and 2 previous 1-year history (number of COVID-19 infections, and number of inpatient visits).

Overall, the four machine learning models—AdaBoost, XGBoost, Neural Network, and Random Forest—demonstrated strong performance using the 9 selected features to predict the four outcomes. Model selection was primarily guided by the AUROC. Additionally, we sought a balanced performance across other criteria, including accuracy, F1 score, sensitivity and specificity,

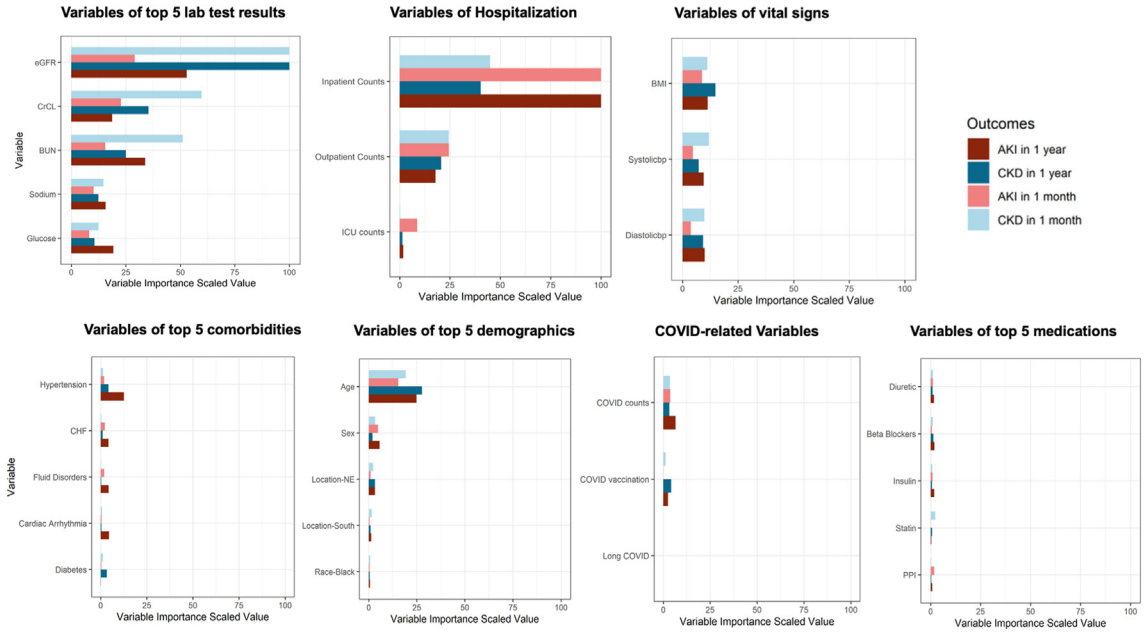

Fig. 3: Scaled variable importance plots for selected variables, with four predicted outcomes: Incidence of AKI in 1 year, Incidence of CKD in 1 year, Incidence of AKI in 1 month, and Incidence of CKD in 1 month.

to determine the final models. XGBoost was selected for predicting (1) the incidence of AKI in 1-month window, achieving an AUROC of 0.803 (95% CI: 0.770–0.835), accuracy of 0.749 (0.728–0.771), F1 score of 0.848 (0.829–0.867), sensitivity of 0.751 (0.732–0.771), and specificity of 0.775 (0.756–0.793); (2) the incidence of AKI in 1-year window, with an AUROC of 0.799 (0.777–0.822), accuracy of 0.740 (0.727–0.760), F1 score of 0.816 (0.798–0.834), sensitivity of 0.746 (0.723–0.769), and specificity of 0.747 (0.724–0.771); (3) the incidence of CKD in 1-year window, with an AUROC of 0.894 (0.879–0.910), accuracy of 0.832 (0.815–0.848), F1 score of 0.883 (0.870–0.899), sensitivity of 0.841 (0.821–0.855), and specificity of 0.803 (0.789–0.818). Random Forest was selected for the incidence of CKD in 1-month window, with an AUROC of 0.896 (0.864–0.928), accuracy of 0.841 (0.822–0.858), F1 score of 0.909 (0.881–0.924), sensitivity of 0.842 (0.821–0.864), and specificity of 0.828 (0.803–0.823) (Fig. 5).

A webpage application, named AIBI APP, was developed incorporating the final models for predicting AKI and CKD in 1 month or 1 year, with the aim of facilitating its utility in clinical settings (Supplementary Figure S2). Users can obtain prediction results (high/low risk) by inputting the corresponding values for the 9 selected variables. A reminder note is included: "This prediction is generated by a machine learning model to assist your clinical decision-making. It should not replace your professional judgment in evaluating the patient." This note aims to remind healthcare providers that the application offers predictive assistance but cannot replace clinical diagnosis. Access to the application is available at the following link: https://zackzhang1993.shinyapps.io/aibi_app/.

Discussion

In this study, we utilised national EHR datasets in the US with a large sample size, emulating a prospective cohort design with rigorous cohort definition and variable identification. We applied multiple ML algorithms to predict the incidence of AKI and CKD in both the short term (1 month) and long term (1 year) during the post-COVID-19 pandemic. Our final models demonstrated strong predictive performance, with an AUROC exceeding 0.89 for CKD prediction and approximately 0.80 for AKI prediction. Additionally, we developed a user-friendly web-based application incorporating the final selected ML models, which include input variables of age, sex, BMI, diastolic blood pressure, systolic blood pressure, eGFR, BUN, number of inpatient admissions, and number of COVID-19 infections, to facilitate real-world clinical application.

In our study, after evaluating eight different machine learning models, we selected XGBoost to predict the risk of AKI in 1 month and 1 year, as well as CKD in 1 year. Additionally, we selected Random Forest to predict CKD in 1 month as the final model. Recently, other studies have also utilised XGBoost,[19,37] and Random Forest[18,38] to predict CKD and AKI. AdaBoost[15,39] and Neural Network[13,16] have also been reported to perform well in

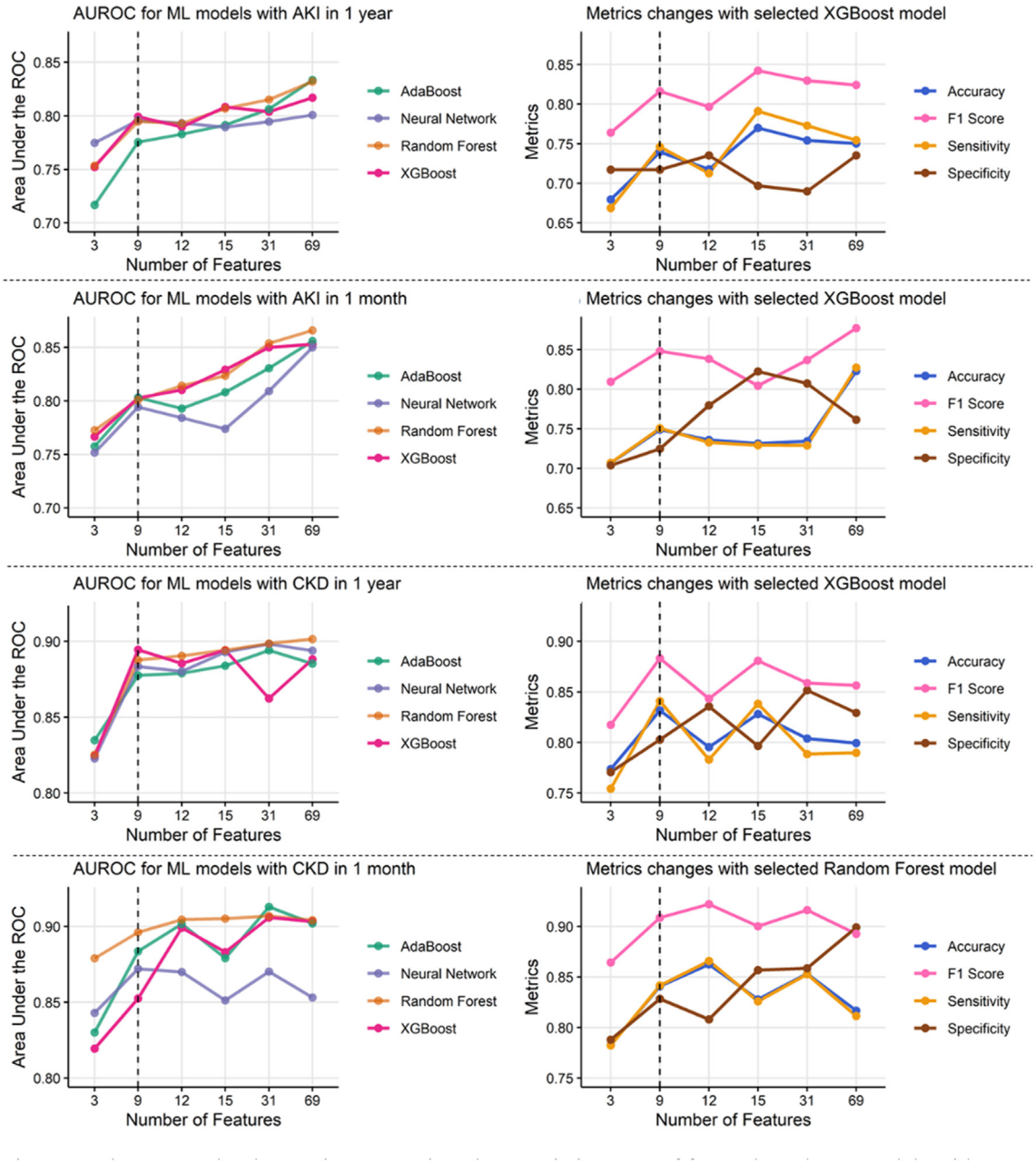

Fig. 4: The area under the receiver operating characteristic curve of four selected ML models with varying number of features, and performance of the selected model with varying number of features: Incidence of AKI in 1 year, Incidence of AKI in 1 month, Incidence of CKD in 1 year, and Incidence of CKD in 1 month. List of features: (1) 31: Age, beta-blocker use, Blood Urea Nitrogen (BUN), body mass index (BMI), Cardiac Arrhythmia, Creatinine Clearance (CrCL), Congestive Heart Failure, diastolic blood pressure, diabetes, diuretic use, estimated Glomerular Filtration Rate (eGFR), glucose, hypertension, ICU count, inpatient count, insulin use, Long COVID, Location in South, Location in Northeast, Location in West, outpatient count, proton pump inhibitor (PPI), Fluid and Electrolyte Disorders, Race in Black or African American, Race in Asian, Race in others, sodium, statin use, systolic blood pressure, COVID-19 counts, and sex. (2) 15: Age, BUN, BMI, Cardiac Arrhythmia, CrCL, Congestive Heart Failure, diastolic blood pressure, diabetes, eGFR, ICU count, inpatient count, outpatient count, systolic blood pressure COVID-19 counts, and sex. (3) 12: Age, BUN, BMI, CrCL, diastolic blood pressure, eGFR, ICU count, inpatient count, outpatient count, systolic blood pressure, COVID-19 counts, and sex. (4) 9: Age, BUN, BMI, diastolic blood pressure, eGFR, inpatient count, systolic blood pressure, COVID-19 counts, and sex. (5) 3: Age, eGFR, inpatient count.

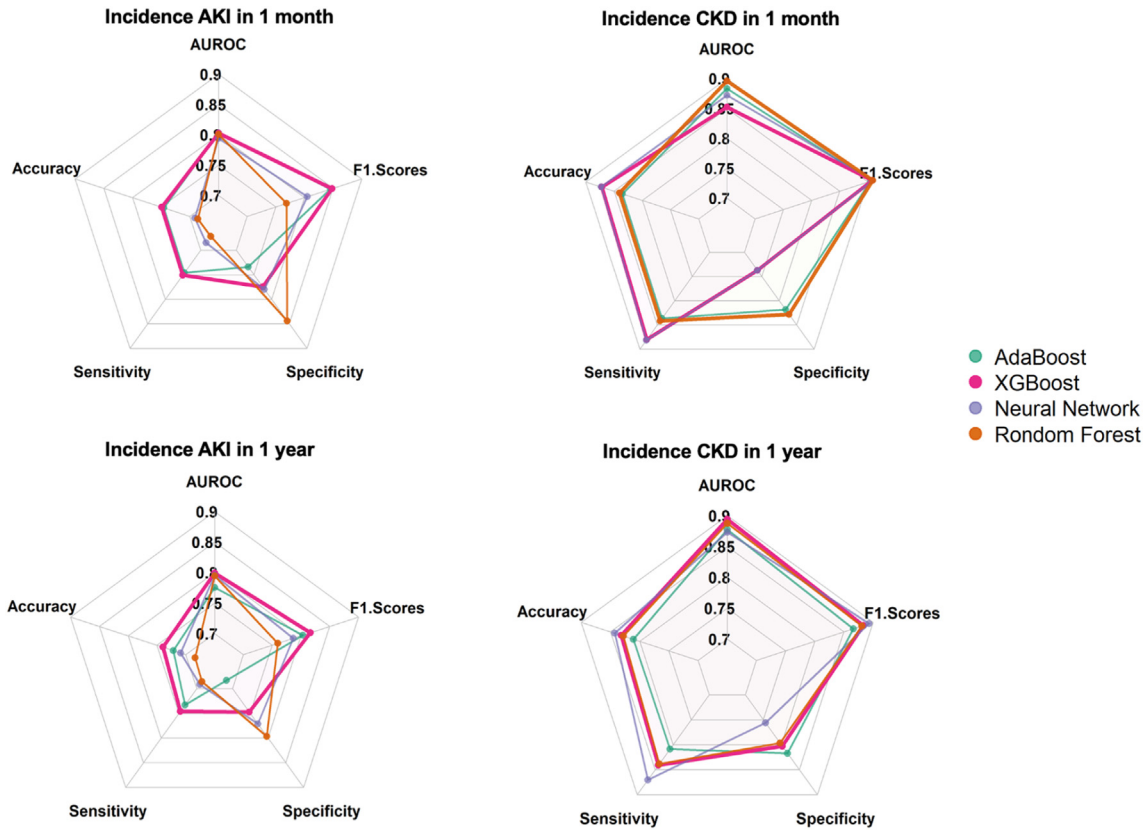

Fig. 5: Performance of machine learning models with selected 9 covariates in predicting outcomes: Incidence of AKI in 1 year, Incidence of CKD in 1 year, Incidence of AKI in 1 month, and Incidence of CKD in 1 month.

predicting kidney disease outcomes. Therefore, in this study, when selecting final models, we specifically compared XGBoost, Random Forest, AdaBoost, and Neural Network together, as all have previously demonstrated effectiveness as ML models for this purpose. In our results, XGBoost and Random Forest demonstrated superior performance for different outcomes, which may be largely attributed to differences in the datasets used, compared to previous studies. We utilised national EHR data comprising 104,565 patients selected from a pool of 1,312,610, whereas prior studies primarily relied on cohort studies with typically only a few hundred patients. Our study confirmed the feasibility of using large-scale EHR data and advanced ML models to effectively predict the risk of kidney diseases.

This study uses the same dataset to predict the incidence of both AKI and CKD in the short term (1 month) and long term (1 year). Regarding the performance of the ML models, we observed that: the prediction of CKD (AUROC in 1 month: 0.896; in 1 year: 0.894) outperformed the prediction of AKI (AUROC in 1 month: 0.803; in 1 year: 0.799). This difference may be attributed to unmeasured or complex risk factors, as well as the multifactorial nature of AKI. The risk factors for AKI include environmental variables, the process of care, acute exposures, and patient-specific factors.[4] In our study, the prediction models primarily utilised patient-specific data, such as demographic information and healthcare visit histories. However, certain environmental variables, such as inadequate hydration, air pollution, or exposure to wastewater, and acute exposures, such as nephrotoxic drugs or unexpected incidents leading to ICU admission, are difficult to capture with the existing datasets, which likely impacted the prediction performance for AKI.

This study aims to predict the risk of kidney outcomes during the post-COVID-19 pandemic, with the number of COVID-19 infections in the previous year included in the final models. Multiple published papers have demonstrated the significant impact of COVID-19 infections on AKI and CKD.[8,11,12,40,41] Moreover, some recent studies also highlight the impact of COVID-19 reinfection on multiple symptoms, including kidney diseases.[42,43] Therefore, we incorporated the number of COVID-19 infections as a variable in our ML models.

We assessed the necessity of including COVID-19 in the final model by evaluating the model's performance (AUROC) after excluding COVID-19 as a variable.

Additionally, we tested the final model with the inclusion of Long COVID to determine whether incorporating Long COVID would improve model performance (Supplementary Figure S3). We observed a significant decrease in AUROC when COVID-19 was excluded from the ML models for predicting AKI in 1 year, CKD in 1 month, and CKD in 1 year. Although a decrease in AUROC was also noted for AKI in 1 month, this reduction was not statistically significant. A probable explanation is that the short-term effects of COVID-19 on AKI may be accounted for by inpatient hospitalization.[44,45] Moreover, when Long COVID was added to the final models, the AUROC significantly decreased for all prediction outcomes. This decline can be attributed to collinearity between COVID-19 infections and Long COVID (Pearson correlation coefficient = 0.56, $p < 0.001$). Another reason for not including Long COVID in our model is the potential misclassification bias associated with its diagnosis, as it often presents with vague symptoms that overlap with other illnesses.[46,47]

We also evaluated the performance of our selected ML models using testing datasets that included only patients without a COVID-19 diagnosis in the year prior to the index date. Compared to the testing datasets of all eligible patients, the model performance with testing datasets of non-COVID-19 patients showed a slight reduction in AUROC, accuracy, sensitivity, specificity, and F1 scores across all four predicted outcomes (Supplementary Table S3). However, our models still demonstrated a high non-misclassification rate (accuracy) in each model: 73.7% for AKI in 1 month, 72.8% for AKI in 1 year, 83.0% for CKD in 1 month, and 82.2% for CKD in 1 year. Our results indicated that our models also perform well for individuals without a COVID-19 diagnosis, which is crucial for assessment in the post-COVID-19 era and provides insights for future research. At present, with access to data across COVID-19 periods, we believe it is essential to consider COVID-19 as a predictor in AKI/CKD models. This aligns with current research, as COVID-19 has been identified as a significant risk factor for kidney disease.[8–11] However, as COVID-19 cases continue to decline,[48] its relevance in future prediction models needs to be re-evaluated. Based on our findings, we suggest that the decision to include COVID-19 as a predictor should depend on (1) whether it remains a significant independent risk factor in future cohort studies, (2) the extent to which its inclusion enhances model performance, and (3) the availability of reliable COVID-19-related data.

We compared our final selected model to LR models, and our final models demonstrated significantly better performance based on AUROC, as assessed by the DeLong test: AKI in 1 month (AUROC: selected model = 0.803, LR = 0.771, $p = 0.025$); AKI in 1 year (AUROC: selected model = 0.799, LR = 0.762, $p = 0.012$); CKD in 1 month (AUROC: selected model = 0.896, LR = 0.872, $p = 0.040$); and CKD in 1 year (AUROC: selected model = 0.894, LR = 0.841, $p < 0.001$). Our findings align with previous systematic reviews comparing ML models to logistic regression for AKI prediction. While traditional ML models often yield similar performance to LR, gradient boosting models have been shown to outperform LR in terms of AUROC and other metrics.[49]

To enhance the utility and applicability of the webpage application, we balanced maintaining strong performance while limiting the number of required variables to a reasonable level. This approach makes the application more accessible and useable in generalised clinical settings. For our final nine selected variables, five demographic variables (age, sex, BMI, diastolic blood pressure, and systolic blood pressure) can be easily obtained during routine nurse pre-checks. eGFR and BUN are common blood tests typically included in standard health check-ups. The remaining two variables—number of COVID-19 infection and number of inpatient visits in the previous 1 year—can be extracted from patient histories or recalled directly by patients. Given this streamlined approach, we believe our AIBI app is highly user-friendly and practical for clinical settings.

Our study has some limitations. First, we acknowledge the potential limitations of the TriNetX dataset in terms of generalizability and data availability. Our sample consists of patients who had access to the healthcare system and were drawn from approximately 120 healthcare organizations across the United States. Therefore, while our dataset is derived from a national source, it may not be fully representative of all patients in the U.S. Additionally, unlike traditional cohort studies that conduct regular monitoring of kidney function, our study extracted kidney function data from laboratory test results recorded in EHR, which may introduce biases inherent to EHR data. Second, after data manipulation, the final cohort of eligible patients for analysis may not fully represent the original TriNetX dataset. Compared to the overall TriNetX data, these patients include a higher proportion of females, White patients, and those from the southern United States (Supplementary Table S4). Third, comparing to clinical trial or prospective cohorts, we acknowledge that the use of EHR data may introduce inherent misclassification bias related to the identification of comorbidities, COVID-19 infections (especially mild cases), and the measurement of CKD and AKI as our prediction outcomes. Fourth, although ICD-10-CM codes allow for the identification of CKD severity (e.g., N18.1 for stage one CKD, N18.2 for stage two CKD), in our dataset, most CKD cases were coded simply as N18. As a result, it is unfortunate that we were unable to predict different stages of kidney disease. Lastly, we used the most recent data to train and internally test our models to focus on the post-COVID-19 pandemic period. However, finding external datasets from this timeframe has been

challenging. Our next steps will involve training and validating our application with datasets beyond TriNetX.

Conclusion

In this study, we utilised large national EHR data and advanced ML models to predict the risks of AKI and CKD in both the short term and long term during the post-pandemic period. Our final prediction models demonstrated strong predictive performance, with eGFR, number of inpatients, number of COVID-19 infections, and six other variables. Incorporating the number of COVID-19 infections in the past year showed improved prediction performance and should be considered in future models for kidney disease prediction. We developed a user-friendly and practical application for clinical settings to assist clinicians in predicting the risk of kidney disease in individual patients and providing timely surveillance.

Contributors

Designed research (project conception, development of overall research plan): YZ, DMB, NG, and VMC. Data extraction and study oversight: YZ, and DMB. Accessed and verified the underlying data: YZ, and DMB. Analysed data: YZ, and DMB. Performed statistical analysis: YZ. Data Visualization: YZ, and RL. Webpage Application Building: YZ, and RL. Wrote the first draft of the manuscript: YZ. Review and editing: YZ, NG, RL, VMC, and DMB. All authors have read and approved the final manuscript. This study is part of the YZ's doctoral dissertation research project with the Penn State College of Medicine, United States of America.

Data sharing statement

The data for this study were obtained from a third-party partner, TriNetX, and cannot be publicly shared. However, readers with access to TriNetX can replicate the patient selection process by following the methodology outlined in our study.

Declaration of interests

The authors have no conflicts of interest to disclose.

Acknowledgements

This study was supported by Artificial Intelligence and Biomedical Informatics Pilot Funding, Penn State College of Medicine.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2025.105726.

References

1 Kidney disease statistics for the United States | NIDDK. National Institute of diabetes and digestive and kidney diseases. https://www.niddk.nih.gov/health-information/health-statistics/kidney-disease. Accessed November 15, 2022.
2 Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. *Kidney Int Suppl*. 2022;12(1):7–11. https://doi.org/10.1016/j.kisu.2021.11.003.
3 Goyal A, Daneshpajouhnejad P, Hashmi MF, Bashir K. Acute kidney injury. In: *StatPearls [internet]*. StatPearls Publishing; 2023. https://www.ncbi.nlm.nih.gov/books/NBK441896/. Accessed August 30, 2024.
4 Kellum JA, Romagnani P, Ashuntantang G, Ronco C, Zarbock A, Anders HJ. Acute kidney injury. *Nat Rev Dis Primer*. 2021;7(1):1–17. https://doi.org/10.1038/s41572-021-00284-z.
5 Rifkin DE, Coca SG, Kalantar-Zadeh K. Does AKI truly lead to CKD? *J Am Soc Nephrol JASN*. 2012;23(6):979–984. https://doi.org/10.1681/ASN.2011121185.
6 Chen TK, Knicely DH, Grams ME. Chronic kidney disease diagnosis and management. *JAMA*. 2019;322(13):1294–1304. https://doi.org/10.1001/jama.2019.14745.
7 Sanmarchi F, Fanconi C, Golinelli D, Gori D, Hernandez-Boussard T, Capodici A. Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. *J Nephrol*. 2023;36(4):1101–1117. https://doi.org/10.1007/s40620-023-01573-4.
8 Bowe B, Xie Y, Xu E, Al-Aly Z. Kidney outcomes in long COVID. *J Am Soc Nephrol*. 2021;32(11):2851–2862. https://doi.org/10.1681/ASN.2021060734.
9 Groff D, Sun A, Ssentongo AE, et al. Short-term and long-term rates of postacute sequelae of SARS-CoV-2 infection: a systematic review. *JAMA Netw Open*. 2021;4(10):e2128568. https://doi.org/10.1001/jamanetworkopen.2021.28568.
10 Chung EYM, Palmer SC, Natale P, et al. Incidence and outcomes of COVID-19 in people with CKD: a systematic review and meta-analysis. *Am J Kidney Dis*. 2021;78(6):804–815. https://doi.org/10.1053/j.ajkd.2021.07.003.
11 Bowe B, Xie Y, Al-Aly Z. Postacute sequelae of COVID-19 at 2 years. *Nat Med*. 2023;29(9):2347–2357. https://doi.org/10.1038/s41591-023-02521-2.
12 Zhang Y, Ba DM, Risher K, et al. Effects of ACE inhibitor/ARB therapy and long COVID on kidney disease: a retrospective cohort study using real-world data. *Clin Kidney J*. 2024;17(7):sfae164. https://doi.org/10.1093/ckj/sfae164.
13 Zhu Y, Bi D, Saunders M, Ji Y. Prediction of chronic kidney disease progression using recurrent neural network and electronic health records. *Sci Rep*. 2023;13(1):22091. https://doi.org/10.1038/s41598-023-49271-2.
14 Takkavatakarn K, Oh W, Cheng E, Nadkarni GN, Chan L. Machine learning models to predict end-stage kidney disease in chronic kidney disease stage 4. *BMC Nephrol*. 2023;24(1):376. https://doi.org/10.1186/s12882-023-03424-7.
15 Suganthi NM,VMJ, Rama P, Chandralekha E. Chronic kidney disease detection using AdaBoosting ensemble method and K-fold cross validation. In: *2022 international conference on automation, computing and renewable systems (ICACRS)*. 2022:979–983. https://doi.org/10.1109/ICACRS55517.2022.10029047.
16 Singh V, Asari VK, Rajasekaran R. A deep neural network for early detection and prediction of chronic kidney disease. *Diagnostics*. 2022;12(1):116. https://doi.org/10.3390/diagnostics12010116.
17 Bai Q, Su C, Tang W, Li Y. Machine learning to predict end stage kidney disease in chronic kidney disease. *Sci Rep*. 2022;12(1):8377. https://doi.org/10.1038/s41598-022-12316-z.
18 Hu J, Xu J, Li M, et al. Identification and validation of an explainable prediction model of acute kidney injury with prognostic implications in critically ill children: a prospective multicenter cohort study. *eClinicalMedicine*. 2024;68:102409. https://doi.org/10.1016/j.eclinm.2023.102409.
19 Ghosh SK, Khandoker AH. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci Rep*. 2024;14(1):3687. https://doi.org/10.1038/s41598-024-54375-4.
20 Collins GS, Moons KGM, Dhiman P, et al. *TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods*. 2024. https://doi.org/10.1136/bmj-2023-078378.
21 Publication Guidelines. TriNetX. https://trinetx.com/real-world-resources/case-studies-publications/trinetx-publication-guidelines/. Accessed August 13, 2024.
22 RxNorm overview. https://www.nlm.nih.gov/research/umls/rxnorm/overview.html. Accessed August 15, 2024.
23 Abedi V, Avula V, Chaudhary D, et al. Prediction of long-term stroke recurrence using machine learning models. *J Clin Med*. 2021;10(6):1286. https://doi.org/10.3390/jcm10061286.
24 Inker LA, Eneanya ND, Coresh J, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. *N Engl J Med*. 2021;385(19):1737–1749. https://doi.org/10.1056/NEJMoa2102953.
25 Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. *Nephron*. 1976;16(1):31–41. https://doi.org/10.1159/000180580.
26 Li J, Guo S, Ma R, et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Med Res Methodol*. 2024;24(1):41. https://doi.org/10.1186/s12874-024-02173-x.
27 Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: a simulation

study. *Stat Methods Med Res*. 2023;32(8):1461–1477. https://doi.org/10.1177/09622802231165001.
28 Buuren S van, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45:1–67. https://doi.org/10.18637/jss.v045.i03.
29 White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377–399. https://doi.org/10.1002/sim.4067.
30 Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. *Am J Epidemiol*. 2010;172(9):1070–1076. https://doi.org/10.1093/aje/kwq260.
31 Deng Y, Lumley T. Multiple imputation through XGBoost. *J Comput Graph Stat*. 2024;33(2):352–363. https://doi.org/10.1080/10618600.2023.2252501.
32 Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357. https://doi.org/10.1613/jair.953.
33 Kuhn M. 15 variable importance | the caret package. https://topepo.github.io/caret/variable-importance.html. Accessed August 16, 2024.
34 Unal I. Defining an optimal cut-point value in ROC analysis: an alternative approach. *Comput Math Methods Med*. 2017;2017:3762651. https://doi.org/10.1155/2017/3762651.
35 Chang W, Cheng J, Allaire JJ, et al. Shiny: web application framework for R. https://cran.r-project.org/web/packages/shiny/index.html; 2024. Accessed August 16, 2024.
36 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845. https://doi.org/10.2307/2531595.
37 Raihan MJ, Khan MAM, Kee SH, Nahid AA. Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Sci Rep*. 2023;13(1):6263. https://doi.org/10.1038/s41598-023-33525-0.
38 Subasi A, Alickovic E, Kevric J. Diagnosis of chronic kidney disease by using random forest. In: Badnjevic A, ed. *CMBEBIH 2017*. Springer; 2017:589–594. https://doi.org/10.1007/978-981-10-4166-2_89.
39 Asra T, Setiadi A, Safudin M, Lestari EW, Hardi N, Alamsyah DP. Implementation of AdaBoost algorithm in prediction of chronic kidney disease. In: *2021 7th International Conference on engineering, applied sciences and technology (ICEAST)*. 2021:264–268. https://doi.org/10.1109/ICEAST52143.2021.9426291.
40 Bowe B, Xie Y, Al-Aly Z. Acute and postacute sequelae associated with SARS-CoV-2 reinfection. *Nat Med*. 2022;28(11):2398–2405. https://doi.org/10.1038/s41591-022-02051-3.
41 Cai M, Xie Y, Topol EJ, Al-Aly Z. Three-year outcomes of post-acute sequelae of COVID-19. *Nat Med*. 2024;30(6):1564–1573. https://doi.org/10.1038/s41591-024-02987-8.
42 Government of Canada SC. Experiences of Canadians with long-term symptoms following COVID-19. https://www150.statcan.gc.ca/n1/pub/75-006-x/2023001/article/00015-eng.htm; 2023. Accessed August 15, 2024.
43 Wei J, Stoesser N, Matthews PC, et al. Risk of SARS-CoV-2 reinfection during multiple Omicron variant waves in the UK general population. *Nat Commun*. 2024;15(1):1008. https://doi.org/10.1038/s41467-024-44973-1.
44 Hirsch JS, Ng JH, Ross DW, et al. Acute kidney injury in patients hospitalized with COVID-19. *Kidney Int*. 2020;98(1):209–218. https://doi.org/10.1016/j.kint.2020.05.006.
45 Chan L, Chaudhary K, Saha A, et al. AKI in hospitalized patients with COVID-19. *J Am Soc Nephrol*. 2021;32(1):151–160. https://doi.org/10.1681/ASN.2020050615.
46 Archambault PM, Rosychuk RJ, Audet M, et al. Post-COVID-19 condition symptoms among emergency department patients tested for SARS-CoV-2 infection. *Nat Commun*. 2024;15(1):8449. https://doi.org/10.1038/s41467-024-52404-4.
47 Høeg TB, Ladhani S, Prasad V. How methodological pitfalls have created widespread misunderstanding about long COVID. *BMJ Evid Based Med*. 2024;29(3):142–146. https://doi.org/10.1136/bmjebm-2023-112338.
48 Preliminary estimates of COVID-19 burden for 2024-2025 | COVID-19 | CDC. https://www.cdc.gov/covid/php/surveillance/burden-estimates.html. Accessed March 14, 2025.
49 Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: a systematic review and meta-analysis. *Int J Med Inf*. 2021;151:104484. https://doi.org/10.1016/j.ijmedinf.2021.104484.