KI REPORTS
KIReports.org
CLINICAL RESEARCH

Machine Learning to Identify Dialysis Patients at High Death Risk

Oguz Akbilgic[1,2], Yoshitsugu Obi[3], Praveen K. Potukuchi[4], Ibrahim Karabayir[1,5], Danh V. Nguyen[6], Melissa Soohoo[3], Elani Streja[3], Miklos Z. Molnar[4,7,8,9], Connie M. Rhee[3], Kamyar Kalantar-Zadeh[3] and Csaba P. Kovesdy[4]

[1]Center for Biomedical Informatics, Department of Pediatrics, University of Tennessee Health Science Center, Memphis, Tennessee, USA; [2]Department of Health Informatics and Data Science, Parkinson School of Health Sciences and Public Health, Loyola University Chicago, Maywood, Illinois, USA; [3]Harold Simmons Center for Kidney Disease Research and Epidemiology, Division of Nephrology and Hypertension, University of California Irvine Medical Center, Orange, California, USA; [4]Division of Nephrology, Department of Medicine, University of Tennessee Health Science Center, Memphis, Tennessee, USA; [5]Faculty of Economics and Administrative Sciences, Kirklareli University, Kirklareli, Turkey; [6]Division of General Internal Medicine, University of California Irvine Medical Center, Orange, California, USA; [7]Department of Surgery, Methodist University Hospital Transplant Institute, Memphis, Tennessee, USA; [8]Division of Transplant Surgery, Department of Surgery, University of Tennessee Health Science Center, Memphis, Tennessee, USA; and [9]Department of Transplantation and Surgery, Semmelweis University, Budapest, Hungary

Introduction: Given the high mortality rate within the first year of dialysis initiation, an accurate estimation of postdialysis mortality could help patients and clinicians in decision making about initiation of dialysis. We aimed to use machine learning (ML) by incorporating complex information from electronic health records to predict patients at risk for postdialysis short-term mortality.

Methods: This study was carried out on a contemporary cohort of 27,615 US veterans with incident end-stage renal disease (ESRD). We implemented a random forest method on 49 variables obtained before dialysis transition to predict outcomes of 30-, 90-, 180-, and 365-day all-cause mortality after dialysis initiation.

Results: The mean (±SD) age of our cohort was 68.7 ± 11.2 years, 98.1% of patients were men, 29.4% were African American, and 71.4% were diabetic. The final random forest model provided C-statistics (95% confidence intervals) of 0.7185 (0.6994–0.7377), 0.7446 (0.7346–0.7546), 0.7504 (0.7425–0.7583), and 0.7488 (0.7421–0.7554) for predicting risk of death within the 4 different time windows. The models showed good internal validity and replicated well in patients with various demographic and clinical characteristics and provided similar or better performance compared with other ML algorithms. Results may not be generalizable to non-veterans. Use of predictors available in electronic medical records has limited the assessment of number of predictors.

Conclusion: We implemented and ML-based method to accurately predict short-term postdialysis mortality in patients with incident ESRD. Our models could aid patients and clinicians in better decision making about the best course of action in patients approaching ESRD.

Kidney Int Rep (2019) **4,** 1219–1229; https://doi.org/10.1016/j.ekir.2019.06.009
KEYWORDS: chronic kidney disease; dialysis; end-stage renal disease; mortality; random forest

Patients with ESRD on dialysis represent a growing group characterized by extremely high morbidity and mortality.[1] Mortality is especially high in the first few months and up to a year after dialysis transition.[1,2] Patients transitioning to dialysis are often acutely ill, and suffer from major comorbid conditions that portend a poor short-term survival,[3] yet dialysis is frequently initiated by health care providers without considering short-term outcomes in their discussions with patients and relatives.[4] Better knowledge about chances to survive in the immediate aftermath of dialysis transition would allow patients to weigh their desire to extend their life to a certain duration versus their quality of life during this time, and more informed patients may elect to pursue other modalities such as conservative or palliative care as opposed to aggressive and invasive interventions.

Several studies have attempted to create prediction models that would allow the quantification of postdialysis mortality to aid informed decision making at

Correspondence: Csaba P. Kovesdy, Nephrology Section, Memphis VA Medical Center, 1030 Jefferson Avenue, Memphis, Tennessee 38104, USA. E-mail: ckovesdy@uthsc.edu
Received 13 February 2019; revised 30 April 2019; accepted 10 June 2019; published online 22 June 2019

the time of dialysis initiation.[5–14] Most of these studies used data from patients who have already started dialysis,[5–13] and hence their utility in aiding pre-transition decisions about starting or deferring dialysis is questionable. In addition, many of these studies had other significant limitations, such as limited sample size,[5,8,14] the use of restricted populations (e.g., elderly[6,7,11,14]), or a lack of information about data with potentially important predictive potential, such as race/ethnicity or various laboratory parameters.[5–10,12–14]

We recently developed a prediction model from a large contemporary cohort of US veterans with incident ESRD, using detailed patient information form the pre-ESRD period, and showing improved diagnostic performance in predicting short-term posttransition mortality.[15] In addition to improvement in population size, composition, and available data, methodological advances in prediction modeling also allow for improvement in prediction parameters. Our aim was to improve on the predictive characteristics of our previous model by implementing an ML algorithm using a random forest method that allows the development of a generalizable model and is also transparent and provides information about the importance of individual variables included in the model.

MATERIALS AND METHODS

Study Population and Patient Characteristics

We used data from a historic cohort of 85,505 US veterans with incident ESRD (Transition of Care in Chronic Kidney Disease, TC-CKD) who transitioned to dialysis treatment between October 1, 2007, and March 31, 2014. The inclusion and exclusion criteria and the data sources used for generating the analytical data set for this study were previously described,[15] and the cohort selection is shown in Supplementary Figure S1. Briefly, we excluded patients with missing information on race/ethnicity, on estimated glomerular filtration rate (eGFR) in the 12 months before ESRD transition, on the cause of their ESRD, and on the diagnostic code for renal disease. For each outcome, we further excluded 35, 64, 108, and 317 patients due to missing 30-, 90-, 180-, and 365-day complete follow-up, respectively. A total of 49 patient characteristics (including demographics, comorbidities, vital signs, vital status, and laboratory characteristics) were identified from a combination of US Renal Data System, Centers for Medicare and Medicaid Services databases and Department of Veterans Affairs administrative databases, as previously described.[16,17] All laboratory parameters (including eGFR) and vital signs (systolic and diastolic blood pressure and body mass index) were obtained before dialysis transition, and the value recorded closest to transition was used for analyses. eGFR was estimated using the Chronic Kidney Diseases Epidemiology Collaboration formula.[18] We further excluded 8263 patients (~23%) with missing information on systolic and diastolic blood pressure and blood sugar. The final analytical cohort consisted of 27,615 patients with complete data for the 49 baseline characteristics. We did not implement any transformation or normalization for continuous predictor variables. In a sensitivity analysis, we imputed missing laboratory data and vital signs by using their mean values. The outcome of interest was all-cause mortality occurring within 30, 90, 180, and 365 days of transitioning to dialysis. Deaths were identified from Department of Veterans Affairs Vital Status Files, which have 98% sensitivity and specificity when compared with the National Death Index.[19]

Predictive Modeling

In traditional Cox proportional hazard regression, the outcome variable is the time to an event. In our approach, we implemented a classification task by treating the outcome as a binary variable representing the occurrence of the event, death in our case, within a certain prediction window. We constructed 4 different prediction models for mortality (corresponding to deaths within 30, 90, 180, and 365 days of dialysis transition) from survival data up to 14 months following transition to dialysis. We implemented ML with a random forest method as our main predictive model to identify patients with ESRD at high risk of death within the first year of dialysis transition. Random forest is an iterative ML method that ensembles multiple simple decision trees to provide the final outcome. At each iteration, a random subset of the predictors is selected, and a decision tree is built using selected predictors. By repeating this multiple times, different simple decision trees are obtained and embedded into the random forest framework.[20] In our models, predictors of mortality consisted of demographics, laboratory results, and comorbid conditions. We executed random forest models each inducing 500 decision trees for each outcome of interest by using 49 risk factors as predictors. We report predictive discrimination as C-statistics. We compared the results from our random forest method with logistic regression and to other commonly used ML algorithms, such as artificial neural networks (ANN), support vector machines, and k-nearest neighborhood, by treating our outcome as a binary variable, and Cox proportional hazard regression and random forest survival models by using an outcome variable representing time from initiation of dialysis to death. As a type of ML, ANN is a broad category of methods that can be used for classification, clustering, and prediction. In this study, we implemented a supervised feed-forward multilayer network model. In a

feed-forward ANN model, the information (input or predictors) is exposed to nonlinear transformation through hidden layers and then connected to the outcome (dependent variable) layer. Support vector machines are supervised ML models that can be used for binary classification by finding the optimal hyperplane separating data coming from 2 distinct categories. Finally, k-nearest neighborhoods are supervised, nonparametric, and instance-based algorithms that memorize input-output associations from training and then estimate the best outcome for a given new input based on its similarity to input data within the training set.

For model building and validation, we split our data into training and testing based on geographical splitting.[21] This layer of cross-validation is to address possible geography-based overfitting. The number of subjects within zip codes starting with 0 to 9 were 1794, 2711, 3374, 2853, 5400, 281, 829, 1909, 3871, 1840, and 2753, respectively. We selected cases with zip codes starting with 0 to 8 for the training data and we built a 10-fold cross-validated (internal validation) model. This layer of cross-validation was to address overfitting due to random splitting of the data. We then implemented the cross-validated model on the test set (external validation) that consisted of cases from zip codes starting with 9. We repeated this process 10 times by changing the test data from zip codes 9 to 0 for all 4 prediction windows. In a sensitivity analysis, we also split the data randomly in a ratio of 80/20, corresponding to cohorts of 22,092 and 5523 patients for training and testing, respectively, and repeated this process 100 times. Once the generalizability of the random forest–based predictive models was verified via both internal and external validation, we built final random forest models using the entire data set with 10-fold cross-validation. We examined the C-statistics from the final models both overall and in relevant subgroups of patients to determine consistency of the results. To assess how well the predicted probabilities reflect the actual survival rate, we split all predicted probability values from final models into 5 clusters based on increasing intervals of predicted risk: 0 to $<$5th, 5th to $<$35th, 35th to $<$65th, 65th to $<$95th, and 95th to $\leq$100th percentiles. We annotated subjects within these 5 categories as "Low Risk," "Mid-Low Risk," "Mid Risk," "Mid-High Risk," and "High Risk," respectively. For each risk group and for each prediction window, we calculated the survival rate and compared it graphically with the observed survival rate. We also used independent sample t and Mann-Whitney U tests to compare predicted risk of death between groups of patients who died and those who survived.

Using the final models, we implemented a variable importance analysis to identify which specific predictors are the main predictors of mortality (Figure 1). The variable importance analysis was based on the mean decrease in accuracy by exclusion of a specific variable. To do so, each variable was dropped from the final model and the performance of the new model was compared with the original model. Predictors causing a large decrease in accuracy when dropped are considered more important than the variables causing less decrease in accuracy when dropped from the model. We normalized the variable importance to (0–100) intervals, where 0 represents the minimum importance (lowest decrease in the accuracy when excluded) and 100 represents the maximum importance (highest decrease in the accuracy when excluded).

To obtain simpler predictive models, we carried out another simulation study based on the variable importance ranking for 30-day mortality variables. In this simulation study, we built 49 different random forest models with 10-fold cross-validation by introducing new risk factors into the model step-by-step based on their importance. First, we built a model using the variable found to be the most important predictor. Second, we used the first and second most important variable, third, the first 3 most important variables, and so on. Finally, the last model was built using all predictors, and the 10-fold cross-validated area under the curve (AUC) corresponding to each of the 49 models is shown graphically in Figure 2, to indicate the point where the introduction of additional variables does not result in substantial further increase in AUC.

To make our approach more applicable in clinical practice, we proceeded to building predictive models using a smaller number of variables without loss of predictive performance. To do that, we ranked the risk factors based on their importance and built new models by introducing new risk factors step-by-step into the model. We built separate 10-fold cross-validated models for each prediction window (i) by using the risk factors that provided at least 0.001 increment in AUC when they were introduced into the model, (ii) by retaining only the top 15 most important predictor variables (since the increment in AUC stabilized for each prediction window by addition of the 15th most important predictor), and (iii) by applying a combination of these 2 approaches. Because the accuracies of the models were comparable for these 3 approaches (data not shown), and because the same top 15 variables ended up being used for all prediction windows, we applied the approach using the 15 top predictor variables and built final compact models using the entire data set. The R-code of our model is available on request for interested researchers.

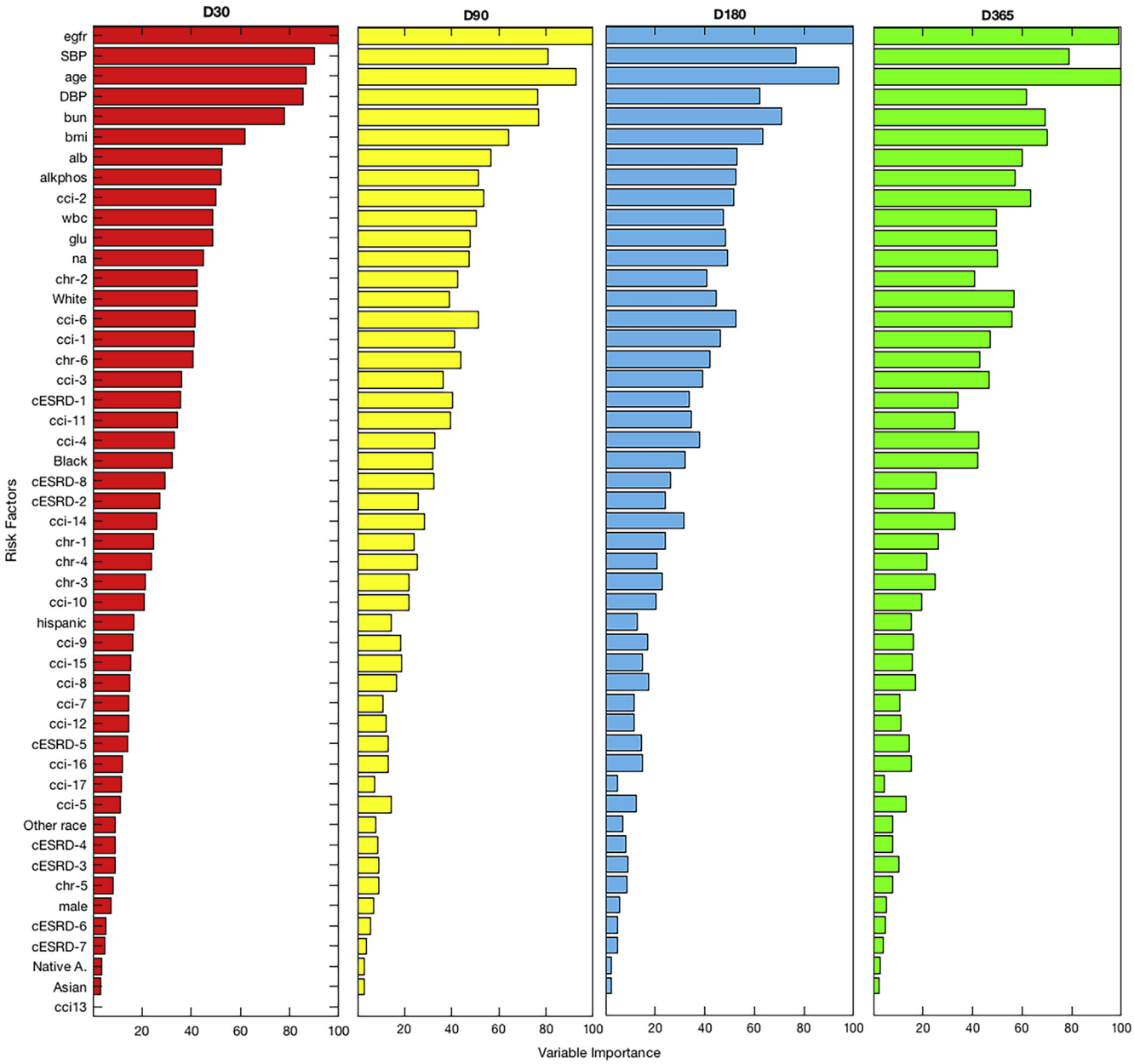

Figure 1. Variable importance analysis. Results indicate the decrease in accuracy of the final model on exclusion of each specific variable, quantified on a relative scale of 0 to 100, where 0 represents the minimum importance (lowest decrease in the accuracy when excluded) and 100 represents the maximum importance (highest decrease in the accuracy when excluded). alb, serum albumin; alkphos, serum alkaline phosphatase; bmi, body mass index; bun, blood urea nitrogen; cci-1, myocardial infarction; cci-2, congestive heart failure; cci-3, peripheral vascular disease; cci-4, cerebrovascular disease; cci-5, dementia; cci-6, chronic pulmonary disease; cci-7, connective tissue disease/rheumatic disease; cci-8, peptic ulcer disease; cci-9, mild liver disease; cci-10, diabetes without complications; cci-11, diabetes with complications; cci-12, paraplegia/hemiplegia; cci-13, renal disease; cci-14, nonmetastatic cancer; cci-15, moderate or severe liver disease; cci-16, metastatic carcinoma; cci-17, HIV/AIDS; cESRD-1, diabetes; cESRD-2, hypertension/large-vessel disease; cESRD-3, primary glomerulonephritis; cESRD-4, interstitial nephritis/pyelonephritis; cESRD-5, neoplasm/tumors; cESRD-6, cystic/hereditary/congenital diseases; cESRD-7, secondary glomerulonephritis; cESRD-8, miscellaneous conditions; chr-1, anemia; chr-2, atrial fibrillation; chr-3, depression; chr-4, hyperlipidemia; chr-5, hypertension; chr-6, ischemic heart disease; DBP, diastolic blood pressure; D30, 30-day mortality; D90, 90-day mortality; D180, 180-day mortality; D365, 365-day mortality; glu, blood glucose; na, serum sodium; Native A, Native American; SBP, systolic blood pressure; wbc, white blood cell count.

The study was approved by the institutional review boards of the Memphis and Long Beach Veterans Affairs Medical Centers, with waiver of written informed consent. All analyses were performed on R version 3.4.1 using "randomForest,"[22] "class,"[23] "e1071,"[24] "neuralnet,"[25] and "ggrandomForestSRC"[26] for various ML algorithms, and figures were generated using MATLAB (MathWorks, Natick,

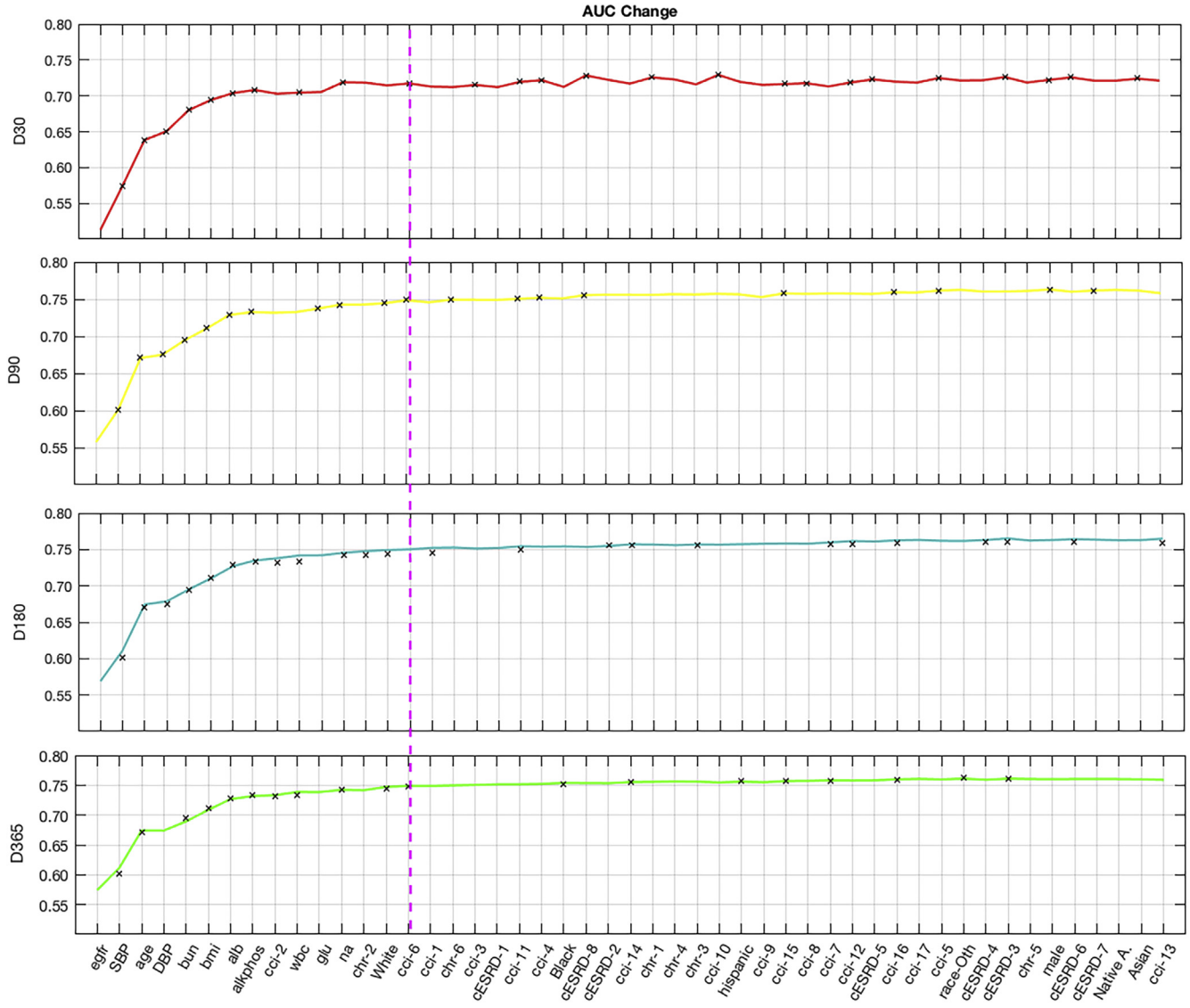

Figure 2. Incremental change in 10-fold cross-validated area under the curve (AUC) on entering individual variables one by one into the predictive model. Variables were entered in the order established by the variable importance analysis for 30-day mortality (Figure 1) as long as they increased the AUC by at least 0.001. alb, serum albumin; alkphos, serum alkaline phosphatase; bmi, body mass index; bun, blood urea nitrogen; cci-1, myocardial infarction; cci-2, congestive heart failure; cci-3, peripheral vascular disease; cci-4, cerebrovascular disease; cci-5, dementia; cci-6, chronic pulmonary disease; cci-7, connective tissue disease/rheumatic disease; cci-8, peptic ulcer disease; cci-9, mild liver disease; cci-10, diabetes without complications; cci-11, diabetes with complications; cci-12, paraplegia/hemiplegia; cci-13, renal disease; cci-14, nonmetastatic cancer; cci-15, moderate or severe liver disease; cci-16, metastatic carcinoma; cci-17, HIV/AIDS; cESRD-1, diabetes; cESRD-2, hypertension/large-vessel disease; cESRD-3, primary glomerulonephritis; cESRD-4, interstitial nephritis/pyelonephritis; cESRD-5, neoplasm/tumors; cESRD-6, cystic/hereditary/congenital diseases; cESRD-7, secondary glomerulonephritis; cESRD-8, miscellaneous conditions; chr-1, anemia; chr-2, atrial fibrillation; chr-3, depression; chr-4, hyperlipidemia; chr-5, hypertension; chr-6, ischemic heart disease; DBP, diastolic blood pressure; D30, 30-day mortality; D90, 90-day mortality; D180, 180-day mortality; D365, 365-day mortality; glu, blood glucose; na, serum sodium; Native A, Native American; SBP, systolic blood pressure; wbc, white blood cell count.

MA) version 2018a and IBM SPSS Statistics version 25 (Armonk, NY).

RESULTS

Table 1 shows baseline characteristics of the 27,615 patients included in the final cohort. Mean (±SD) age was 68.7 ± 11.2 years, 98.1% of patients were men, 29.4% were African American, and 71.4% were diabetic.

Table 2 summarizes the average C-statistics with 95% confidence intervals (CIs) of the random forest model using all 49 baseline characteristics as predictors in the training (internal validation: these are 10-fold cross-validation results over training data) and testing sets (external validation) defined based on the zip codes of the individual patients' home residence address. Results were similar with sensitivity analysis, whereas the training and testing sets were based on random

Table 1. Baseline characteristics ($N = 27,615$)

Age, yr	68.7 ± 11.2
Male sex	27,101 (98.1)
Race	
White	18,162 (65.8)
Black	8127 (29.4)
Native American	182 (0.7)
Asian	150 (0.5)
Other	994 (3.6)
Hispanic ethnicity	2119 (7.7)
Cause of ESRD	
Diabetes	13,414 (48.6)
Hypertension/large-vessel disease	8113 (29.4)
Primary glomerulonephritis	1396 (5.1)
Interstitial nephritis/pyelonephritis	713 (2.6)
Neoplasm/tumors	689 (2.5)
Cystic/hereditary/congenital diseases	454 (1.7)
Secondary glomerulonephritis	243 (0.9)
Miscellaneous conditions	2591 (9.4)
Comorbidities	
Myocardial infarction	6884 (24.8)
Congestive heart failure	14,895 (53.9)
Peripheral vascular disease	10,179 (36.9)
Cerebrovascular disease	8259 (29.9)
Dementia	609 (2.2)
Chronic pulmonary disease	11,143 (40.4)
Connective tissue disease/rheumatic disease	1008 (3.7)
Peptic ulcer disease	1747 (6.3)
Mild liver disease	2514 (9.1)
Diabetes without complications	4982 (18.0)
Diabetes with complications	14,754 (53.4)
Paraplegia/hemiplegia	844 (3.1)
Renal disease	27,615 (100)
Nonmetastatic cancer	5635 (20.4)
Moderate or severe liver disease	616 (2.2)
Metastatic carcinoma	610 (2.2)
HIV/AIDS	315 (1.1)
Anemia	20,272 (73.4)
Atrial fibrillation	4196 (15.2)
Depression	7210 (26.1)
Hyperlipidemia	22,432 (81.2)
Hypertension	27,105 (98.2)
Ischemic heart disease	15,780 (57.1)
Body mass index (kg/m^2)	29.9 ± 6.7
Vital signs	
Systolic blood pressure (mm Hg)	140.3 ± 23.2
Diastolic blood pressure (mm Hg)	72.5 ± 13.7
Laboratory characteristics	
Last eGFR (ml/min per 1.73 m^2)	11.7 (8.3–17.3)
White blood cells (×10^3/μl)	7.8 ± 3.5
Serum sodium (mEq/l)	138.9 ± 3.8
Serum albumin (g/dl)	3.4 ± 0.7
Serum urea nitrogen (mg/dl)	66.9 ± 29.3
Serum alkaline phosphatase (IU/l)	96.3 ± 71.8
Glucose (mg/dl)	129.6 ± 59.0

eGFR, estimated glomerular filtration rate; ESRD, end-stage renal disease.
Values are provided as means ± SD, median (25th–75th percentile), and number (%).

splits of 80%/20% (Supplementary Table S1). All the C-statistics were similar in the training and testing sets, indicating that the random forest models were generalizable within the Veteran Affairs cohort for all 4 outcomes of interest. The final model for each outcome using the entire US veteran dataset showed C-statistics (95% CI) of 0.718 (0.708–0.718), 0.758 (0.749–0.766), 0.762 (0.754–0.770), and 0.760 (0.751–0.769) for 30-day, 90-day, 180-day, and 365-day mortality, respectively.

Figure 1 shows results from the variable importance analysis using the final cross-validated models for the 4 outcomes. eGFR, systolic blood pressure, and age were consistently the 3 most important predictors of outcome for each prediction window. Other variables that were important for prediction in the different models were diastolic blood pressure, blood urea nitrogen, body mass index, serum alkaline phosphatase, congestive heart failure, and chronic pulmonary disease. Figure 2 shows the contribution of the individual patient characteristics to the 10-fold cross-validated AUC of the predictive model when entered one by one into the model. The AUC showed no substantial increase after the addition of the 15th most important predictor. The C-statistic and 95% CI for the simplified model that included the top 15 predictive variables were 0.719 (0.699–0.738), 0.745 (0.735–0.755), 0.750 (0.743–0.758), and 0.749 (0.742–0.755) for the 30-day, 90-day, 180-day, and 365-day mortality outcomes, respectively, with Figure 3 showing the corresponding receiver operating characteristic curves. Table 3 shows the C-statistics and 95% CIs of the final compact predictive model in select subgroups of patients, indicating consistent results in all subgroups, except for a better predictive performance in patients whose pretransition eGFR was <15 ml/min per 1.73 m^2 compared with those with eGFR ≥ 15 ml/min per 1.73 m^2. Supplementary Figures S2 and S3 and Supplementary Table S2 show the comparison of predicted and observed all-cause mortality, indicating consistent results when categorizing patients by quartiles of predicted mortality rate from low risk to high risk (Supplementary Figure S2) and when comparing predicted mortality rates between patients who survived and those who died (Supplementary Figure S3 and Supplementary Table S2).

For comparison, we implemented ANN, support vector machines, k-nearest neighborhood, and logistic regression methods to predict risk for death within 1 year of initiation of dialysis by implementing the same geographic location–based cross-validation protocol. The multilayer ANN architecture with a single hidden layer including 100 neurons and using an activation function of sigmoid provided C-statistics (95% CI) of 0.606 (0.561–0.645). A support vector machines model provided C-statistics (95% CI) of 0.719 (0.708–0.730). When we implemented a k-nearest neighborhood algorithm with $k = 10$, it provided C-statistics (95% CI) of 0.654 (0.642–0.666),

Table 2. Predictive model performance in cross-validated models split by geographic location

	C-statistics with 95% confidence interval			
	30-day mortality	90-day mortality	180-day mortality	365-day mortality
Training	0.7179 (0.7144–0.7214)	0.7594 (0.7578–0.7611)	0.7628 (0.7606–0.7649)	0.7596 (0.7582–0.7610)
Test	0.7361 (0.7149–0.7573)	0.7637 (0.7535–0.7739)	0.7604 (0.747458–0.7751)	0.7574 (0.7463–0.7686)

The initial training set was defined based on individual patient residence zip codes starting with 0 to 8 and the test set was defined based on zip codes starting with 9. We repeated this process 10 times by changing the test data zip codes from 9 to 0 for all for prediction windows. The values presented are the average C-statistics from all models.

while logistic regression providing C-statistics (95% CI) of 0.682 (0.649–0.714). Our random forest survival analysis using an outcome variable representing the time to death from initiation of dialysis including 200 individual trees provided C-statistics (95% CI) of 0.688 (0.656–0.710). We also ran Cox regression and obtained C-statistics (95% CI) of 0.714 (0.706–0.722). Supplementary Figure S4 shows comparisons of all models considered.

The results were similar when using a patient cohort of 35,878 patients with imputed values for missing laboratory data and vital signs (C-statistic and 95% CI for 1-year mortality: 0.749 [0.744–0.755]).

DISCUSSION

We developed a new risk score to predict postdialysis all-cause mortality in patients with incident ESRD

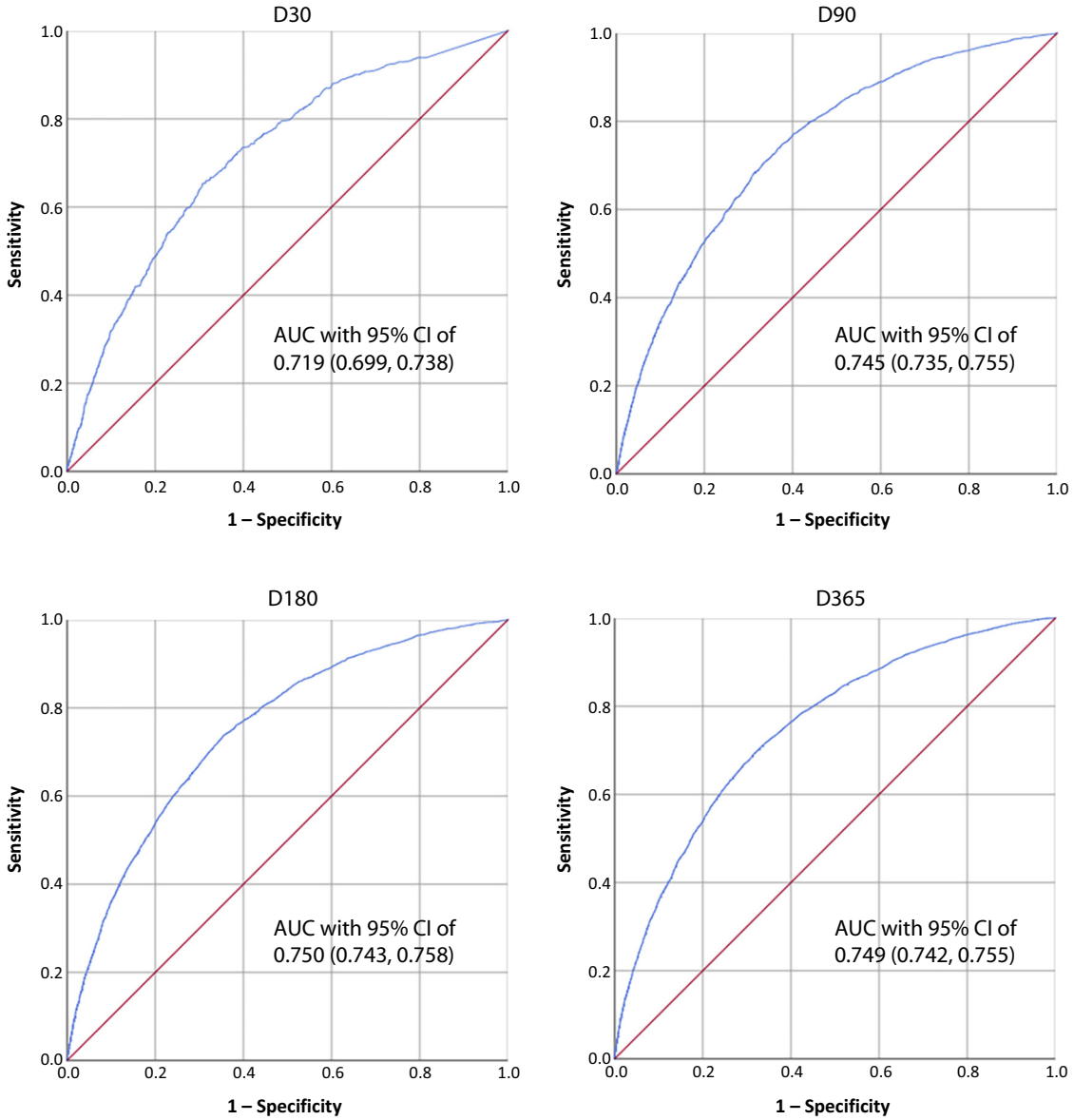

Figure 3. Receiver operator curves of the final compact predictive model based on the top 15 predictors of 30-day, 90-day, 180-day, and 365-day mortality. AUC, area under the curve; CI, confidence interval; D30, 30-day mortality; D90, 90-day mortality; D180, 180-day mortality; D365, 365-day mortality.

Table 3. C-statistics and 95% confidence intervals in select subgroups of patients

	D30		D90		D180		D365	
	n	AUC	*n*	AUC	*n*	AUC	*n*	AUC
Overall	27,580	0.719 (0.699, 0.738)	27,551	0.745 (0.735, 0.755)	27,507	0.750 (0.743, 0.758)	27,298	0.749 (0.742, 0.755)
Age								
<65	11,306	0.696 (0.648, 0.743)	11,298	0.727 (0.704, 0.750)	11,291	0.731 (0.714, 0.748)	11,227	0.730 (0.717, 0.743)
≥65	16,274	0.689 (0.667, 0.711)	16,253	0.710 (0.698, 0.722)	16,216	0.717 (0.708, 0.727)	16,071	0.717 (0.709, 0.725)
Race								
Black	8,119	0.716 (0.666, 0.766)	8108	0.768 (0.746, 0.791)	8098	0.771 (0.754, 0.789)	8051	0.756 (0.742, 0.771)
White	18,136	0.698 (0.677, 0.720)	18,119	0.721 (0.709, 0.733)	18,087	0.726 (0.717, 0.736)	17,932	0.725 (0.717, 0.733)
Sex								
Female	514	0.701 (0.550, 0.852)	513	0.701 (0.614, 0.787)	513	0.752 (0.682, 0.822)	505	0.753 (0.693, 0.812)
Male	27,066	0.719 (0.699, 0.738)	27,038	0.745 (0.735, 0.755)	26,994	0.750 (0.742, 0.758)	26,793	0.748 (0.741, 0.755)
Type of dialysis								
Hemodialysis	25,976	0.713 (0.694, 0.733)	25,949	0.740 (0.730, 0.750)	25,908	0.745 (0.737, 0.754)	25,709	0.744 (0.738, 0.751)
Peritoneal dialysis	1363	0.790 (0.654, 0.925)	1361	0.777 (0.706, 0.848)	1358	0.794 (0.742, 0.845)	1352	0.767 (0.727, 0.808)
Cause of ESRD								
Diabetes								
Yes	13,399	0.700 (0.667, 0.734)	13,388	0.730 (0.713, 0.747)	13,375	0.735 (0.722, 0.748)	13,285	0.738 (0.728, 0.748)
No	14,181	0.719 (0.695, 0.742)	14,163	0.742 (0.730, 0.755)	14,132	0.751 (0.741, 0.761)	14,013	0.750 (0.741, 0.758)
Hypertension								
Yes	8099	0.707 (0.676, 0.738)	8089	0.742 (0.725, 0.759)	8069	0.752 (0.738, 0.766)	7992	0.748 (0.737, 0.760)
No	19,481	0.721 (0.697–0.745)	19,462	0.743 (0.731–0.756)	190,438	0.748 (0.738–0.758)	19,306	0.747 (0.739, 0.756)
Comorbidities								
Congestive heart failure								
Yes	14,876	0.683 (0.660–0.706)	14,862	0.716 (0.703–0.729)	14,840	0.725 (0.715–0.735)	14,713	0.725 (0.717, 0.734)
No	12,704	0.729 (0.690–0.767)	12,689	0.746 (0.727–0.764)	12,667	0.746 (0.731–0.760)	12,585	0.734 (0.722, 0.745)
Chronic pulmonary disease								
Yes	11,133	0.678 (0.651–0.705)	11,119	0.700 (0.686–0.715)	11,102	0.710 (0.698–0.722)	11,000	0.716 (0.706, 0.726)
No	16,447	0.731 (0.702–0.760)	16,432	0.754 (0.739–0.769)	16,405	0.753 (0.741–0.765)	16,298	0.745 (0.735, 0.754)
eGFR								
<15	18,528	0.720 (0.690, 0.751)	18,512	0.733 (0.717–0.749)	18,489	0.737 (0.725–0.748)	18,386	0.728 (0.719–0.738)
≥15	9052	0.643 (0.615, 0.671)	9039	0.675 (0.660–0.690)	9018	0.686 (0.674–0.699)	8912	0.699 (0.688–0.710)

AUC, area under the curve; D, day; ESRD, end-stage renal disease.
Results represent C-statistics and 95% confidence intervals based on the final compact predictive model including the 15 most important variables.

using exclusively information available before dialysis transition in a large national cohort of US veterans. Our random forest–based predictive model showed good validity when tested in cohorts divided geographically and discriminated well in various subgroups. Our model also provided similar or better performance in terms of C-statistic when compared with methods such as ANN, logistic regression, support vector machines, k-nearest neighborhood, Cox regression, and random forest survival. However, we have not compared our results with more novel survival approaches such as Deep Survival Analysis[27,28] because of computational limitations within our systems.

Interestingly, our 1-year mortality prediction was better than the 30-days mortality prediction. This may be because all-cause mortality may be a result of different causes of death at different time points, and the covariates included in our models may be better at predicting the types of deaths occurring at 1 year.

The predictive performance of our model (C-statistics with 95% CI of 0.76; 0.75–0.77) were superior to the performance of more traditionally used Cox regression models (C-statistics with 95% CI of 0.714; 0.706–0.722). This may be due to the distribution- and assumption-free nature of the random forest model allowing simultaneous use of predictors with high correlation. We also note that the performance of the current Cox regression model was slightly better than the performance of the Cox regression model (0.71; 0.70–0.72 for eGFR <15 and 0.66; 0.65–0.67 for eGFR ≥15) previously built on the same cohort.[15] This may be because of our inclusion of additional predictors (blood pressure and blood glucose) in our current model. Two other major studies examined 6-month postdialysis all-cause mortality, both implementing a logistic regression model.[7,11] In terms of AUC comparison for 6-month postdialysis mortality prediction, the C-statistic of 0.76 we obtained was higher compared with the C-statistic of 0.69 in the study by Thamer *et al.*[11] and 0.70 in the study by Couchoud *et al.*[7] Besides differences in the applied prediction methods, differences from these studies also could be explained by differences in the studied patient

population and differences in the available predictor variables.

ESRD is a complex disease state, with complex interactions among demographic, comorbid, and laboratory characteristics. As opposed to traditional methods, the ML-based random forest method used by us to develop our prediction model is ideally suited to incorporate complex interactions between numerous outcome predictors, and hence could result in superior predictive performance. Two main critiques of ML-based predictive models in clinical decision making are that they are not intuitive (they are often referred to as "black box" methods), and that they may result in overfitting causing limited generalizability. In our study, we used the random forest method as the main predictive model algorithm because (i) it is suitable for variable importance analysis providing information on which specific risk factors contribute more to the outcome,[20,29,30] and (ii) it usually provides more generalizable models compared with other ML algorithms.[31–33]

Our models are based on predialysis patient characteristics, and hence they may aid decision making before the initiation of dialytic therapies by informing patients, caregivers, and health care providers about the likely short-term risk of mortality if a patient were to choose such a path. The knowledge of a numeric risk could make it easier for patients to decide between a dialytic modality and more conservative interventions, such as medications-only, or palliative therapy, and could result in better patient-centered outcomes such as quality of life for individuals with high short-term mortality risk, and also in health care savings.[4,7] The application by clinicians of our model or other similar models to predict future mortality and to advise patients and families about outcomes will need to be done with proper recognition of the remaining limitations of such models. Notwithstanding the moderate accuracy of currently available prediction models, the ability to present a numerical risk of death still represents an advance that could aid in the discussion between caregivers and patients and family members.

We used data that are available in electronic medical records to develop our predictive models, which makes system-wide scaling and automation of our method possible in health care systems using computerized patient records. Although the use of data from electronic medical records is an advantage from the standpoint of scalability and generalizability across systems, it also represents a limitation due to the lack of patient characteristics that are not available in electronic medical records, such as various socioeconomic and clinical characteristics known to have an important role in ESRD (e.g., severity of various disease states, nutritional status). The incorporation of such additional characteristics in future studies promises to further improve our ability to predict mortality in this population.

Our study has limitations that need to be recognized when interpreting its findings. We used a nationwide US veteran cohort to develop our predictive models, which consisted of mostly male patients with incident ESRD who all transitioned to dialysis. Therefore, our model may not be applicable to female patients or to patients at earlier stages of chronic kidney disease, although previously a different predictive model using the same veteran population validated well in an independent cohort with equal representation of women.[15] We used predictors available in electronic medical records, which limited the assessment of predictors; for example, the use of International Classification of Diseases codes for the assessment of comorbid conditions can have low sensitivity and does not allow for the determination of stages of severity for most of them, and certain conditions cannot be characterized at all. Furthermore, we excluded a large proportion of the source cohort due to missing data, which could limit our ability to extend conclusions to the entire US veteran population. Another limitation is exclusion of several variables that have been shown to predict outcomes in dialysis patients, such as vascular access at the time of dialysis transition,[34] the use of certain medications[35] or the length of care received in the predialysis period,[36] proteinuria,[37] serum phosphorus,[38] or hepatitis C seropositivity.[39] Our exclusion of these variables was motivated by a desire to use solely patient characteristics assessed before reaching a decision about renal replacement therapy, and those that are not primarily representative of health care system or practitioner behavior. Other variables such as laboratory tests with known predictive ability (e.g., proteinuria, serum phosphorus, or hepatitis C seropositivity) were excluded due to high degrees of combined missingness in our cohort. Incorporation of this and other clinical information has the potential to further improve predictive performance, which can be tested in future studies. Finally, although our results were internally valid, we did not have access to an independent data set to perform external validation of our results. We will therefore provide our R-code to researchers who intend to test our approach on their own data.

To conclude, we used ML methods to predict short-term all-cause mortality in incident dialysis patients based on predialysis information alone. Our models could help informing patients and caregivers about short-term postdialysis mortality and may aid decision

making in patients who are faced with ESRD. Our online risk calculator will also allow the immediate application of our risk scores in clinical practice.

DISCLOSURE

All the authors declared no competing interests.

ACKNOWLEDGMENTS

CK and KKZ are employees of the US Department of Veterans Affairs. The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as official policy or interpretation of the US Department of Veterans Affairs or the US government.

This study is supported by grant U01-DK102163 from the National Institutes of Health to KKZ and CK and by resources from the US Department of Veterans Affairs (VA). The data reported here have been supplied in part by the US Renal Data System (USRDS). Support for VA/Centers for Medicare and Medicaid Services (CMS) data is provided by the Veterans Health Administration, Office of Research and Development, Health Services Research and Development, and VA Information Resource Center (project numbers SDR 02–237 and 98–004).

STATEMENT OF ETHICS

This manuscript reports results based on a secondary analysis of retrospective data, which does not require collection of written informed consent. The study protocol has been approved by the research institute's committee on human research. There was no animal experiment carried out.

AUTHOR CONTRIBUTIONS

Study concept and design: all authors. Acquisition of data: YO, ES, MS, PKP, and CPK.

Analysis and interpretation of data: OA, IK, and CPK. Drafting of the manuscript: OA and CPK. Critical revision of the manuscript for important intellectual content and approval of the final version: all authors. Statistical analysis: OA, IK, and CPK. Obtained funding: KKZ and CPK. Administrative, technical, or material support: YO, PKP, and CPK. Study supervision: KKZ and CPK.

SUPPLEMENTARY MATERIAL

Supplementary File (Word)

Table S1. Predictive model performance in cross-validated models split randomly.

Table S2. Comparison of predicted mortality rates between patients who survived and those who died.

Figure S1. Cohort selection study flowchart of the Veterans Affairs cohort.

Figure S2. Comparison of predicted and observed 1-year survival rates. Patients are categorized according to categories of predicted survival rates based on the compact prediction model containing 15 predictors.

Figure S3. Predicted and survival rates between patients who survived and those who died during 30-day, 90-day, 180-day, and 365-day observation windows.

Figure S4. Area under the curve with 95% confidence intervals for models in comparison predicting 1-year postdialysis mortality. ANN, artificial neural networks; Cox, Cox regression; Cox-eGFR<15, our team's previously published model on the same cohort not including vital signs for patients with eGFR <15; Cox-eGFR>15, our team's previously published model on the same cohort not including vital signs for patients with eGFR >15; KNN, K-nearest neighborhood; LR, Logistic regression; RF-Surv, Random forest survival; RF-49, The random forest model built on all 49 predictors; RF-15, The random forest model built on top 15 predictors; SVM, support vector machines.

REFERENCES

1. Saran R, Robinson B, Abbott K, et al. US Renal Data System 2016 Annual Data Report: Epidemiology of Kidney Disease in the United States. *Am J Kidney Dis*. 2017;69:A7–A8.
2. Arif FM, Sumida K, Molnar MZ, et al. Early mortality associated with inpatient versus outpatient hemodialysis initiation in a large cohort of US Veterans with incident end-stage renal disease. *Nephron*. 2017;137:15–22.
3. Kovesdy CP, Naseer A, Sumida K, et al. Abrupt decline in kidney function precipitating initiation of chronic renal replacement therapy. *Kidney Int Rep*. 2018;3:602–609.
4. Wong SPY, McFarland LV, Liu CF, et al. Care practices for patients with advanced kidney disease who forgo maintenance dialysis. *JAMA Intern Med*. 2019;179:305–313.
5. Barrett BJ, Parfrey PS, Morgan J, et al. Prediction of early death in end-stage renal disease patients starting dialysis. *Am J Kidney Dis*. 1997;29:214–222.
6. Couchoud C, Hemmelgarn B, Kotanko P, et al. Supportive care: time to change our prognostic tools and their use in CKD. *Clin J Am Soc Nephrol*. 2016;11:1892–1901.
7. Couchoud C, Labeeuw M, Moranne O, Allot V. A clinical score to predict 6-month prognosis in elderly patients starting dialysis for end-stage renal disease. *Nephrol Dial Transplant*. 2009;24:1553–1561.
8. Foley RN, Parfrey PS, Hefferton D, et al. Advance prediction of early death in patients starting maintenance dialysis. *Am J Kidney Dis*. 1994;23:836–845.
9. Ivory SE, Polkinghorne KR, Khandakar Y, et al. Predicting 6-month mortality risk of patients commencing dialysis treatment for end-stage kidney disease. *Nephrol Dial Transplant*. 2017;32:1558–1565.
10. Liu J, Huang Z, Gilbertson DT, et al. An improved comorbidity index for outcome analyses among dialysis patients. *Kidney Int*. 2010;77:141–151.
11. Thamer M, Kaufman JS, Zhang Y, et al. Predicting early death among elderly dialysis patients: development and validation of a risk score to assist shared decision making

for dialysis initiation. *Am J Kidney Dis*. 2015;66:1024–1032.

12. Thijssen S, Usvyat L, Kotanko P. Prediction of mortality in the first two years of hemodialysis: results from a validation study. *Blood Purif*. 2012;33:165–170.
13. Wagner M, Ansell D, Kent DM, et al. Predicting mortality in incident dialysis patients: an analysis of the United Kingdom renal registry. *Am J Kidney Dis*. 2011;57:894–902.
14. Wick JP, Turin TC, Faris PD, et al. A clinical risk prediction tool for 6-month mortality after dialysis initiation among older adults. *Am J Kidney Dis*. 2017;69:568–575.
15. Obi Y, Nguyen DVD, Zhou H, et al. Development and validation of prediction scores for early mortality at transition to dialysis. *Mayo Clin Proc*. 2018;93:1224–1235.
16. Lu JL, Molnar MZ, Sumida K, et al. Association of the frequency of pre-end-stage renal disease medical care with post-end-stage renal disease mortality and hospitalization. *Nephrol Dial Transplant*. 2018;33:789–795.
17. Sumida K, Molnar MZ, Potukuchi PK, et al. Association of slopes of estimated glomerular filtration rate with post–end-stage renal disease mortality in patients with advanced chronic kidney disease transitioning to dialysis. *Mayo Clin Proc*. 2016;91:196–207.
18. Levey AS, Stevens L. A new equation to estimate glomerular filtration rate. *Ann Intern Med*. 2009;150:604–612.
19. Sohn MW, Arnold N, Maynard C, Hynes DM. Accuracy and completeness of mortality data in the Department of Veterans Affairs. *Popul Health Metr*. 2006;4:2.
20. Breiman L. Random forest. *Machine Learning*. 2001;45:5–32.
21. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *Eur Urol*. 2015;67:1142–1151.
22. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2:18–22.
23. Venables WN, Ripley BD. *Modern Applied Statistics With S, Fourth Edition*. New York, NY: Springer-Verlag; 2002.
24. Dimitriadou E, Hornik K, Leisch MF, et al., R Package 'E1071' [computer program]. Version 1.5-25. Published March 5, 2011. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.191.317&rep=rep1&type=pdf. Accessed July 11, 2019.
25. Günther F, Fritsch S. neuralnet: Training of Neural Networks. *R J*. 2019;2:30.
26. Ehrlinger J, Blackstone EH. ggRandomForests: Survival with Random Forests. Published 2015. Available at: https://www.semanticscholar.org/paper/ggRandomForests-%3A-Survival-with-Random-Forests-Ehrlinger-Blackstone/386fd1e44fae06e0ab73e8cc24f7b75fd6935ac3. Accessed July 11, 2019.
27. Katzman J, Shaham U, Cloninger A, et al. Deep Survival: A Deep Cox Proportional Hazards Network. *arXiv*. 2016;1606.00931v1 [stat.ML]. Accessed June 1, 2018.
28. Yousefi S, Amrollahi F, Amgad M, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep*. 2017;7:11707.
29. Grömping U. Variable importance assessment in regression: linear regression versus random forest. *Am Stat*. 2009;63:308–319.
30. Janitza S, Tutz G, Boulesteix AL. Random forest for ordinal responses: prediction and variable selection. *Comput Stat Data Anal*. 2016;96:57–73.
31. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920–1930.
32. Kullarni VY, Sinha PK. Random Forest classifier: a survey and future research directions. *Int J Adv Comput*. 2013;36:1144–1156.
33. Mahajan R, Kamaleswaran R, Howe A, Akbilgic O. Cardiac rhythm classification from a short single lead ECG recording via Random Forests. *Computing in Cardiology*. 2017;44:1–4.
34. Saleh T, Sumida K, Molnar MZ, et al. Effect of age on the association of vascular access type with mortality in a cohort of incident end-stage renal disease patients. *Nephron*. 2017;137:57–63.
35. Streja E, Gosmanova EO, Molnar MZ, et al. Association of continuation of statin therapy initiated before transition to chronic dialysis therapy with mortality after dialysis initiation. *JAMA Netw Open*. 2018;1:e182311.
36. Liu P, Quinn RR, Oliver MJ, et al. Association between duration of predialysis care and mortality after dialysis start. *Clin J Am Soc Nephrol*. 2018;13:893–899.
37. Kovesdy CP, Lott EH, Lu JL, et al. Outcomes associated with microalbuminuria: effect modification by chronic kidney disease. *J Am Coll Cardiol*. 2013;61:1626–1633.
38. Kovesdy CP, Anderson JE, Kalantar-Zadeh K. Outcomes associated with serum phosphorus level in males with non-dialysis dependent chronic kidney disease. *Clin Nephrol*. 2010;73:268–275.
39. Molnar MZ, Alhourani HM, Wall BM, et al. Association of hepatitis C viral infection with incidence and progression of chronic kidney disease in a large cohort of US veterans. *Hepatology*. 2015;62:1495–1502.