Pena Michelle (Orcid ID: 0000-0003-3340-2893)
Heerspink Hiddo (Orcid ID: 0000-0002-3126-3730)
Belur Nagaraj Sunil (Orcid ID: 0000-0002-6409-4101)

Machine Learning based Early Prediction of End-stage Renal Disease in Patients with Diabetic Kidney Disease using Clinical Trials Data

Sunil Belur Nagaraj, PhD[1*], Michelle J Pena, PhD[1], Wenjun Ju, PhD[2], Hiddo L Heerspink, PhD[1,3] on behalf of the BEAt-DKD consortium

[1] Department of Clinical Pharmacy & Pharmacology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[2] University of Michigan, Ann Arbor, MI USA

[3] The George Institute for Global Health, Sydney, Australia

*Corresponding author

Sunil Belur Nagaraj

Department of Clinical Pharmacy and Pharmacology

University of Groningen,

University Medical Center Groningen,

De Brug 1D – 1- 019

9700AD, Groningen, The Netherlands.

s.belur.nagaraj@umcg.nl

Number of words in the abstract: 246

Number of words in the main text: 2650

Number of figures: 4

Number of tables: 1

Number of supplementary figures: 1

Number of supplementary tables: 4

ABSTRACT

Aims: Predicting long-term renal risk in patients with type 2 diabetes is of importance in clinical practice and clinical trials. We hypothesize that by using multiple baseline demographic and clinical characteristics, machine learning models can accurately predict end-stage renal disease (ESRD).

Materials and methods: In total 11789 patients from three clinical trials: RENAAL (N = 1513), IDNT (N = 1715), and ALTITUDE (N = 8561) with type 2 diabetes and nephropathy were used in this study. Eighteen baseline demographic and clinical characteristics were used as predictors to train machine learning models to predict ESRD (doubling of serum creatinine and/or end-stage renal disease). We used the area under the receiver operator curve (AUC) to assess the prediction performance of models and compared against traditional Cox proportional hazard regression and kidney failure risk equation models.

Results: The feed forward neural network model predicted ESRD with an AUC of 0.82 (0.76-0.87), 0.81 (0.75-0.86), and 0.84 (0.79 – 0.90) in RENAAL, IDNT and ALTITUDE, respectively. The feed forward neural network model selected UACR, serum albumin, uric acid and serum creatinine as important predictors and obtained the state-of-the-art performance to predict the long-term ESRD.

Conclusions: Despite large inter-patient variability, nonlinear machine learning models can be used to predict long term ESRD in patients with type 2 diabetes and nephropathy using baseline demographic and clinical characteristics. The proposed method offers a potential to create accurate and multiple outcome prediction automated models to identify high-risk patients who could benefit from therapy in clinical practice.

INTRODUCTION

Diabetic kidney disease (DKD) is the leading cause of end-stage renal disease (ESRD) [1] . Blood pressure lowering with angiotensin-converting enzyme inhibitors (ACEi) and angiotensin receptor blockers (ARB) are guideline recommended treatment to slow down the progression of diabetic kidney disease [2–4]. However, individual patients show a large variation in disease progression likely due to the complex heterogenous nature of the disease. There is a need for a robust and efficient tool to identify patients at highest risk for developing ESRD and who require stringent monitoring and treatment intensification.

In current practice, albuminuria [5] and eGFR[6] are the main predictors of progression of diabetic kidney disease. However, a recent study suggests that the margin of error for all eGFR formulae is high, thus making it a less reliable tool to access overall renal function [7]. The primary reason is that the coefficients used in current eGFR formulae are population-based and are less efficient at an individual level. Various renal risk scores have been developed using traditional epidemiological tools (Cox regression or logistic regression) for predicting ESRD [8,9]. The past decade has seen a boom in computational processes for predictive analytics using machine learning techniques. Unlike traditional statistical approaches where pre-selected clinical characteristics are used in prediction, machine learning techniques can automatically identify important characteristics to predict ESRD. Several methods have already been developed to predict ESRD from electronic health records using machine learning techniques [10–15]. However, these methods use observational data and lack external validation: models were trained and validated within the same dataset and are not likely to generalize well due to patient heterogenity and demographic differences [16].

In this work, we developed and validated a machine learning framework to predict long term ESRD in patients with type 2 diabetes and nephropathy using baseline clinical characteristics of 11,789 patients participating in past clinical trials. We hypothesized that including several baseline clinical characteristics in a machine learning model can accurately identify patients at high risk to develop ESRD. We specifically used clinical trial data to train and validate our models as it benefits from (i) rigorous data and endpoint collection through independent adjudication committees using rigorous definitions and procedures; (ii) central laboratory measurements minimizing inter-laboratory assay variability; and (iii) international reach which increases generalizability to various populations. We externally validated the performance of the machine learning models to address the problem of inter-patient variability.

MATERIALS AND METHODS

Study population

For the present study, we used data from three clinical trials: RENAAL (*N*=1513), IDNT (*N*=1715), and ALTITUDE (*N*=8561). The detailed design, rationale, and study outcomes for these trials have been previously published [2,3,17]. In RENAAL and IDNT, the effect of angiotensin receptor blockers; losartan and irbesartan, respectively, on renal outcomes were investigated. Inclusion criteria in RENAAL and IDNT were similar with only minor differences. Patients with type 2 diabetes, hypertension, and nephropathy aged 30-70 years were eligible for both trials. Serum creatinine levels ranged between 1.0 mg/dL and 3.0 mg/dL. All patients had proteinuria, defined as a urinary albumin to creatinine ratio (UACR) of >300 mg/g based on single first morning void or a 24-hour urinary protein excretion of >500mg/day in the RENAAL trial and >900 mg/day in the IDNT trial. In both trials the glomerular filtration rate was estimated (eGFR) using the Modification of Diet in Renal Disease (MDRD) Study formula [18]. Exclusion criteria for both trials were type 1 diabetes or non-diabetic renal disease.

Patients in the RENAAL trial were randomly allocated to treatment with losartan 100 mg/day or matched placebo. Patients in the IDNT trial were randomly allocated to treatment with irbesartan 300 mg/day or matched placebo. The IDNT trial additionally included a calcium channel blocker treatment arm (amlodipine 10 mg/day). The trials were designed to keep the dose of the ARB stable during follow-up. Additional antihypertensive agents (other than ACEi or ARB in RENAAL, and ACEi, ARB, or calcium channel blockers in IDNT) were allowed during the trial to achieve the target level of 135/85 mm Hg or less for RENAAL or 140/90 mm Hg or less for IDNT.

In the ALTITUDE trial, 8561 type 2 diabetes patients with a high risk of renal and cardiovascular events from 854 centers in 36 countries were included. Patients were randomly allocated to treatment with aliskiren 300 mg/day or matched placebo. The median follow-up duration was 32.9 months. Patients with UACR ≥200 mg/g, eGFR ≥30 and ≤ 60 ml/min/1.73 m^2, or a history of cardiovascular disease were included in the trial.

All trials were approved by local medical ethics committees and conducted according to the guidelines of the declaration of Helsinki.

Clinical variables

Eighteen baseline clinical variables were used as predictors to train the models: age, sex, body mass index (BMI), smoking status, diastolic blood pressure (DBP), systolic blood pressure (SBP), serum creatinine, serum potassium, haemoglobin, glycated hemoglobin (HbA1c), serum albumin, serum calcium, phosphorous, serum uric acid, high-density lipoprotein (HDL), low-density lipoprotein (LDL), UACR , and history of carviovascular diseases. Each trial measured all serum and urine samples in a central laboratory. It should be noted that though we did not use eGFR directly as an input variable to the machine learning model, we used all variables that is used for eGFR calculation: serum creatinine, age and sex in the machine learning model. By this way, the ML model identifies non-linear relationship between these variables with other variables for ESRD prediction instead of linear relationship used in traditional eGFR calculations.

Clinical Outcomes

For all trials, the primary renal endpoint was a composite of ESRD, defined as chronic dialysis or renal transplantation, or a confirmed doubling of serum creatinine from baseline. All renal endpoints were adjudicated by a blinded independent endpoint committee using rigorous guidelines and definitions.

Performance evaluation metric

We used the area under the receiver operator characteristic curve (AUC) as the metric to evaluate the performance of the model. AUC = 1 indicates that the model can accurately distinguish between high and low risk patients; AUC = 0.5 indicates that the model's performance is equivalent to random chance performance. In addition, we also estimated the following performance measures for all models:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}},\ Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}},\ \text{and}\ F - score = 2 \times \frac{\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision+recall}},$$

where *TP* (true positive) = number of correctly classified patients with ESRD, *FP* (false positive) = number of incorrectly classified patients with ESRD, and *FN* (false negative) = number of incorrectly classified patients without ESRD. Similar to the AUC, precision, recall and F-score values of 1.0 indicates accurate classification. In addition, we also obtained calibration plots of best performing models to assess the relationship between predicted probabilities and the observed ESRD outcomes [19].

Statistical significance was obtained using a paired t-test on the probability output of the prediction models. A *p*-value < 0.05 was considered as significant.

Statistical Analysis

The architecture of the proposed machine learning based ESRD prediction system is shown in figure 1. First, we used the *k*-nearest neighbor algorithm [20] to impute missing variables in both training and testing sets. The percentage number of variables imputed using this technique is summarized in supplementary table 1. Since the training dataset consisted of unequal number of patients from two groups (with and without an event), a class imbalance problem is created which could severely bias the performance of the system. Due to this, we created a balanced training set by using Synthetic Minority Over-sampling Technique (SMOTE) algorithm [21]. Variables in the training set were standardized by subtracting the mean and dividing by the standard deviation to have unit mean and standard deviation. Testing set variables were standardized with respect to the mean and standard deviation of the training set. We then performed 5-fold cross-validation within the training set (80% subset for training the model and the remaining unseen 20% subset for validation) to identify the optimal combination of variables (feature selection) using elastic-net regularization algorithm and tune hyperparameters of the machine learning models (see supplementary table 2).

Since five different classification models were obtained due to 5-fold cross-validation, we repeated this process 1000 times to obtain 5000 models (1000 iterations of 5-fold cross-validation). Since different classification models are obtained for every hyperparameter combination and during every training fold, the model that provided the highest AUC on the validation set was used as the final model and was trained on all of the training data. The final trained optimal model was then used to estimate the probability of ESRD for each patient in the testing set. Through this process, we obtained nearly an unbiased estimate of the classification model since only training data was used for optimizing classifier models which are completely independent of the testing set.

We compared the performance of four classical machine learning algorithms: logistic regression, support vector machine with Gaussian kernel, random forest, and feed forward neural networks to predict ESRD. We performed following experiments to evaluate the performance of our models: train on RENAAL + IDNT, test on ALTITUDE; train on RENAAL + ALTITUDE, test on IDNT; and train on IDNT + ALTITUDE, test on RENAAL. In all experiments, we combined data from two clinical trials and tested on the third clinical trial to include a large number of patients with ESRD for training the model. We also compared the performance of machine learning models with the traditional Cox proportional hazards regression and Kidney Failure Risk Equation models (KFRE) [9]. In the KFRE model, we use age, sex, UACR, eGFR, bicarbonate, phosphorus, albumin and calcium variables to estimate the ESRD probability. Since bicarbonate was not present in the ALTITUDE data, we did not estimate KFRE ESRD probability in these data.

All of the coding and analysis were performed using the MATLAB 2018a scripting language (Natick, MA, USA). All results are reported as mean (95% confidence interval (CI)) unless stated otherwise. We used bootstrapping with 1000 samplings to estimate the 95% confidence interval. Paired t-test was used to estimate the statistical significance.

RESULTS

In total, there were 489, 283, and 508 patients with ESRD in RENAAL (median follow-up of 3.7 years), IDNT (median follow-up of 2.6 years), and ALTITUDE (median follow-up of 2.7 years) trials, respectively. Figure 2 illustrates the performance of individual clinical variables for ESRD prediction. UACR had the highest prediction performance in RENAAL:

AUC = 0.72 (0.69 – 0.74) and IDNT: 0.65 (0.63 – 0.67), respectively. In ALTITUDE, UACR (AUC = 0.77 [0.74 – 0.79]) and hemoglobin (AUC = 0.77 [0.72 – 0.80]) provided highest prediction performance when compared to other variables.

Table 1 summarizes the prediction performance of the proposed approach using machine learning models for all training-testing combinations. The performance of the feed forward neural networks (FNN) model (single layer, 50 neurons, activation function = sigmoid, loss function = binary cross entropy, regularization parameter = 0.0001, solver = adam, learning rate = 0.01) outperformed other machine learning models and achieved the highest AUC of 0.82 (0.76 – 0.87), 0.81 (0.75 – 0.86), and 0.84 (0.79 – 0.90) for predicting ESRD in RENAAL, IDNT, and ALTITUDE respectively. The performance of the FNN model was significantly better (*p-value* < 0.05) than the traditional Cox regression and KFRE models in all three datasets. Additional performance metrics are provided in supplementary table 3. The distribution of ESRD probability in individuals with and without an ESRD event predicted by the FNN and Cox models is shown in figure 3. We set a probability threshold of 0.5 for equal weightage for the two groups and estimated the mean Euclidean distance [22] between the probability scores < 0.5 (without ESRD) and probability scores ≥ 0.5 (with ESRD). The separation of predicted probabilities between two groups using FNN (Euclidean distance: RENAAL = 0.66, IDNT = 0.68) was higher when compared to the KFRE (Euclidean distance: RENAAL = 0.49, IDNT = 0.52). Figure 4 compares the calibration plots of FNN and KFRE. The calibration plot of FNN more closely follows the diagonal line when compared to the KFRE in both RENAAL and IDNT. However, there was no significant difference between the calibration plots of FNN and KFRE (*p-value* = 0.1 and 0.2 for RENAAL and IDNT, respectively).

Supplementary figure 1 shows the heatmap of variables selected by the elastic-net regularization algorithm. Different number of variables were selected by the algorithm for different training and validation steps, and in total 7 (age, UACR, serum albumin, serum uric acid, haemoglobin, SBP and serum creatinine), 8 (age, UACR, serum albumin, phosphorous, serum uric acid, haemoglobin, SBP and serum creatinine), and 5 (UACR, serum albumin, phosphorous, haemoglobin and serum creatinine) variables were selected when the algorithm was trained on RENAAL + IDNT, RENAAL + ALTITUDE, and IDNT + ALTITUDE, respectively. UACR, serum albumin, serum uric acid and serum creatinine were selected as important predictive variables (normalized weight > 0.3) in all three training combinations (the normalized weight > 0.3 was used a convention the importance interpretation).

To evaluate the impact of treatment assignment to placebo or active intervention, we tested the performance of the FNN model separately on placebo and treatment arms. Supplementary table 4 summarizes the prediction performance. There was no significant difference (*p-value* > 0.05) in the final prediction performance of the FNN model irrespective of treatment assignment.

To evaluate how much internal cross-validation biases the performance of the machine learning models when compared to external validation, we pooled RENAAL, IDNT and ALTITUDE trial data and performed 10-fold cross-validation using the pooled dataset. The FNN model resulted in an overall AUC of 0.90 (0.85 – 0.93), much better than the AUC obtained during external validation. This increase in the prediction performance was due to the random inclusion of few patients from the testing set during model training process which can severely bias the prediction performance.

DISCUSSION

We present a framework to assess and compare the performance of various machine learning techniques to predict long term ESRD risk using baseline information. The FNN based ESRD prediction model demonstrated good prediction ability (AUC's > 0.8 in three clinical trials) and outperformed other machine learning and traditional risk prediction models which were validated in the same dataset. Accordingly, the FNN model accurately identified high-risk patients who could benefit from therapy using baseline clinical information. The consistent performance of the FNN model in three clinical trials suggests that the proposed framework avoids model overfitting and is likely to generalize well on the new dataset. Such a model can also be used as a early prediction tool to identify patients who could benefit from intensified therapy in clinical practice.

Findings of this study have four important implications. First, we demonstrate that individual clinical variables are not sufficient to accurately predict long-term ESRD outcome. Second, machine learning techniques incorporating multiple clinical variables can predict ESRD much better than the existing traditional logistic or Cox regression methods, or better than the KFRE renal risk score. Third, UACR, serum albumin, serum uric acid and serum creatinine were selected by the elastic net regularization technique in all three clinical trials making them important biomarkers to predict ESRD. Fourth, machine learning algorithms are not sensitive whether the patient was treated with placebo or ARBs, suggesting that the developed algorithm can be used to predict ESRD for any individual regardless of the RAAS intervention background medication.

The machine learning framework developed in this study has several advantages. First, it uses a data-driven approach to identify multiple (and novel) risk markers associated with ESRD instead of the traditional hypothesis driven approach. Second, it can be used as a personalized ESRD monitoring tool where the machine learning model is repeatedly retrained with the new clinical assessments at different time points, thus calibrating it for the underlying patient. Third, the framework can also be used as a screening tool for patient inclusion/exclusion in clinical trials. Enriching trials with patients with a high probability of developing long-term ESRD can reduce sample size requirements and lead to shorter more efficient clincial trials.

Though several machine learning-based methods have already been developed to predict individuals' renal diseases with CKD [10–14], a fair comparison is difficult due to (i) variability within datasets, (ii) methodological differences to develop prediction models, and (iii) external validation. Differences in datasets can be due to heterogeneity of disease severity and drug response either from observational studies or clinical trials. Methodological differences can rise due to improper tuning of machine learning hyperparameters which can severely bias the prediction performance. Hyperparameter tuning is essential for robust and stable performance of the machine learning model and we achieved this by performing exhaustive grid search over a broad range of hyperparameters using only training data which resulted in a consistent performance (AUC > 0.8) when validated in all three clinical trials. Our results also confirm the importance of external validation of the prediction model when compared to cross-validation within the same dataset which can result in optimistic performance. This kind of external validation is important to evaluate the robustness and generalizability of the model when used for prediction on a new dataset. We recommend using internal cross-validation for model development and external validation for evaluating the stability of the prediction performance of the model.

Despite obtaining good ESRD prediction using machine learning algorithms, there are several limitations to our study. Firstly, a sample size of 11,789 patients may not be sufficient to capture large heterogenity of disease severity seen in patients. Secondly, we used data from clinical trials which is both a strength and a limitation of our study. A strength due to minimal variability in the clinical measurements; random assignment of patients to the treatment; timely assessment of end points; inclusion of patients from multiple countries and centers capturing demographic heterogeneity. A limitation since the developed machine learning model does not take into account the variability in medication adherence which are commonly seen in observational data. Thirdly, the machine learning model did not achieve perfect prediction performance (AUC = 1.0). We hypothesize that further improvements can be obtained by including (i) additional molecular and cellular biomarkers, and (ii) increasing the overall sample size for training the FNN model. Fourthly, these data are analyzed only in a clinical trial setting. Validating the algorithm's in a real world setting should be addressed in the future in order to determine it's true generalizabilty to a non-clinical trial, type 2 diabetes general population.

To conclude, we evaluated the performance of several machine learning algorithms using baseline demographic and clinical variables for predicting the ESRD in individual patients with type 2 diabetes and nephropathy. The performance of the FNN model was superior when compared to other machine learning models. Findings of this study pave the way to develop accurate and stable next-generation machine learning based ESRD predicition systems for clinical practice to identify high-risk patients who could benefit from therapy.

CONFLICTOF INTERESTS

Dr. L Heerspink reports grants and other from Abbvie, grants and other from Astra Zeneca, grants and other from Boehringer Ingelheim, other from Dimerix, other from Merck, other from MundiPharma, other from Mitsubishi Tanabe, other from Retrophin, other from Chinook, grants and other from Janssen, outside the submitted work. Other authors have no conflicts of interest to declare.

FUNDING

This study was funded from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 115974 (BEAt-DKD).

REFERENCES

1. Ghaderian SB, Hayati F, Shayanpour S, Beladi Mousavi SS. Diabetes and end-stage renal disease; a review article on new concepts. *J Ren Inj Prev*. 2015;4(2):28-33. doi:10.12861/jrip.2015.07

2. Brenner BM, Cooper ME, de Zeeuw D, et al. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *N Engl J Med*. 2001;345(12):861-869. doi:10.1056/NEJMoa011161

3. Lewis EJ, Hunsicker LG, Clarke WR, et al. Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *N*

Engl J Med. 2001;345(12):851-860. doi:10.1056/NEJMoa011303

4. Patel A, ADVANCE Collaborative Group, MacMahon S, et al. Effects of a fixed combination of perindopril and indapamide on macrovascular and microvascular outcomes in patients with type 2 diabetes mellitus (the ADVANCE trial): a randomised controlled trial. *Lancet Lond Engl*. 2007;370(9590):829-840. doi:10.1016/S0140-6736(07)61303-8

5. Heerspink HJL, Greene T, Tighiouart H, et al. Change in albuminuria as a surrogate endpoint for progression of kidney disease: a meta-analysis of treatment effects in randomised clinical trials. *Lancet Diabetes Endocrinol*. 2019;7(2):128-139. doi:10.1016/S2213-8587(18)30314-0

6. Greene T, Ying J, Vonesh EF, et al. Performance of GFR Slope as a Surrogate End Point for Kidney Disease Progression in Clinical Trials: A Statistical Simulation. *J Am Soc Nephrol JASN*. 2019;30(9):1756-1769. doi:10.1681/ASN.2019010009

7. Porrini E, Ruggenenti P, Luis-Lima S, et al. Estimated GFR: time for a critical appraisal. *Nat Rev Nephrol*. 2019;15(3):177-190. doi:10.1038/s41581-018-0080-9

8. Lin C-C, Li C-I, Liu C-S, et al. Development and validation of a risk prediction model for end-stage renal disease in patients with type 2 diabetes. *Sci Rep*. 2017;7(1):10177. doi:10.1038/s41598-017-09243-9

9. Tangri N, Stevens LA, Griffith J, et al. A predictive model for progression of chronic kidney disease to kidney failure. *JAMA*. 2011;305(15):1553-1559. doi:10.1001/jama.2011.451

10. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, et al. Application of Machine Learning Techniques to High-Dimensional Clinical Data to Forecast Postoperative Complications. *PloS One*. 2016;11(5):e0155705. doi:10.1371/journal.pone.0155705

11. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc JAMIA*. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030

12. Kate RJ, Perez RM, Mazumdar D, Pasupathy KS, Nilakantan V. Prediction and detection models for acute kidney injury in hospitalized older adults. *BMC Med Inform Decis Mak*. 2016;16:39. doi:10.1186/s12911-016-0277-4

13. Nadkarni GN, Fleming F, McCullough JR, et al. Prediction of rapid kidney function decline using machine learning combining blood biomarkers and electronic health record data. *bioRxiv*. Published online 2019:587774.

14. Ravizza S, Huschto T, Adamov A, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat Med*. 2019;25(1):57-59.

doi:10.1038/s41591-018-0239-8

15. Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol*. 2018;12(2):295-302. doi:10.1177/1932296817706375

16. Di Tanna GL, Wirtz H, Burrows KL, Globe G. Evaluating risk prediction models for adults with heart failure: A systematic literature review. *PloS One*. 2020;15(1):e0224135. doi:10.1371/journal.pone.0224135

17. Parving H-H, Brenner BM, McMurray JJV, et al. Cardiorenal end points in a trial of aliskiren for type 2 diabetes. *N Engl J Med*. 2012;367(23):2204-2213. doi:10.1056/NEJMoa1208799

18. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann Intern Med*. 1999;130(6):461-470.

19. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning*. ; 2005:625–632.

20. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak*. 2016;16(Suppl 3). doi:10.1186/s12911-016-0318-z

21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357.

22. Hu L-Y, Huang M-W, Ke S-W, Tsai C-F. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*. 2016;5(1):1304. doi:10.1186/s40064-016-2941-7

FIGURE LEGENDS

Figure 1: Architecture of the proposed ESRD prediction system. Rigorous cross-validation was performed to identify optimal model to predict renal risk in the testing set. Abbreviations: *k*-NN = *k* nearest neighbor; SMOTE = synthetic minority oversampling technique; CV = cross-validation; ESRD = end-stage renal disease.

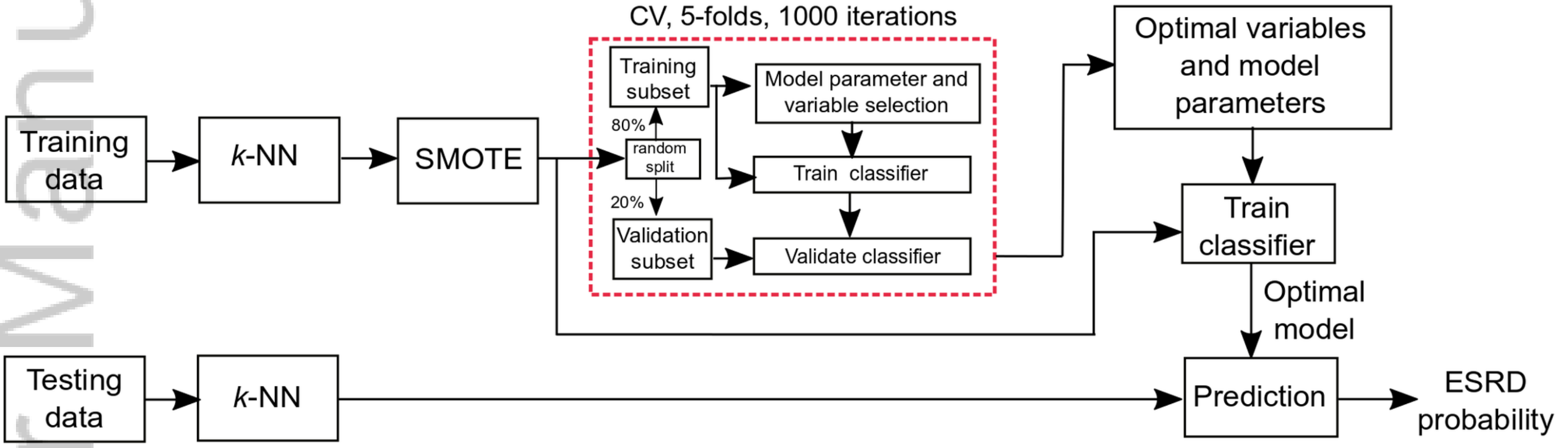
Figure 2: The distribution of AUC (mean (95% CI)) to predict ESRD using individual variables in all three clinical trials. Solid vertical black line corresponds to the mean AUC and rectangular box represents the standard deviation. Abbreviations: AUC = area under the receiver operator characteristic curve; BMI = body mass index; ACR = urine albumin-creatinine ratio; Albumin = serum albumin; Phos = phosphorous;SP = serum potassium; UA = serum uric acid; Scr = serum creatinine; DBP = diastolic blood pressure; SBP = systolic blood pressure; Hb = hemoglobin; Smoking = current/past smoker;CVD = history of cardiovascular diseases.

Figure 3: Plot showing the distribution of the predicted ESRD risk probability in patients with and without ESRD events for all three clinical trials. Jittering was performed for the ESRD event for better visualization. The best performing machine learning model (FNN) is compared with the best performing traditional KFRE model. To quantify the separation between two clusters, we estimated the mean Euclidean distance between the probability scores < 0.5 (without ESRD) and probability scores ≥ 0.5 (with ESRD). The mean Euclidean distance for FNN and KFRE models were 0.66 and 0.5, respectively. Abbreviations: ESRD = end-stage renal disease; KFRE = Kidney Failure Risk Equation; FNN = Feed-forward neural network.

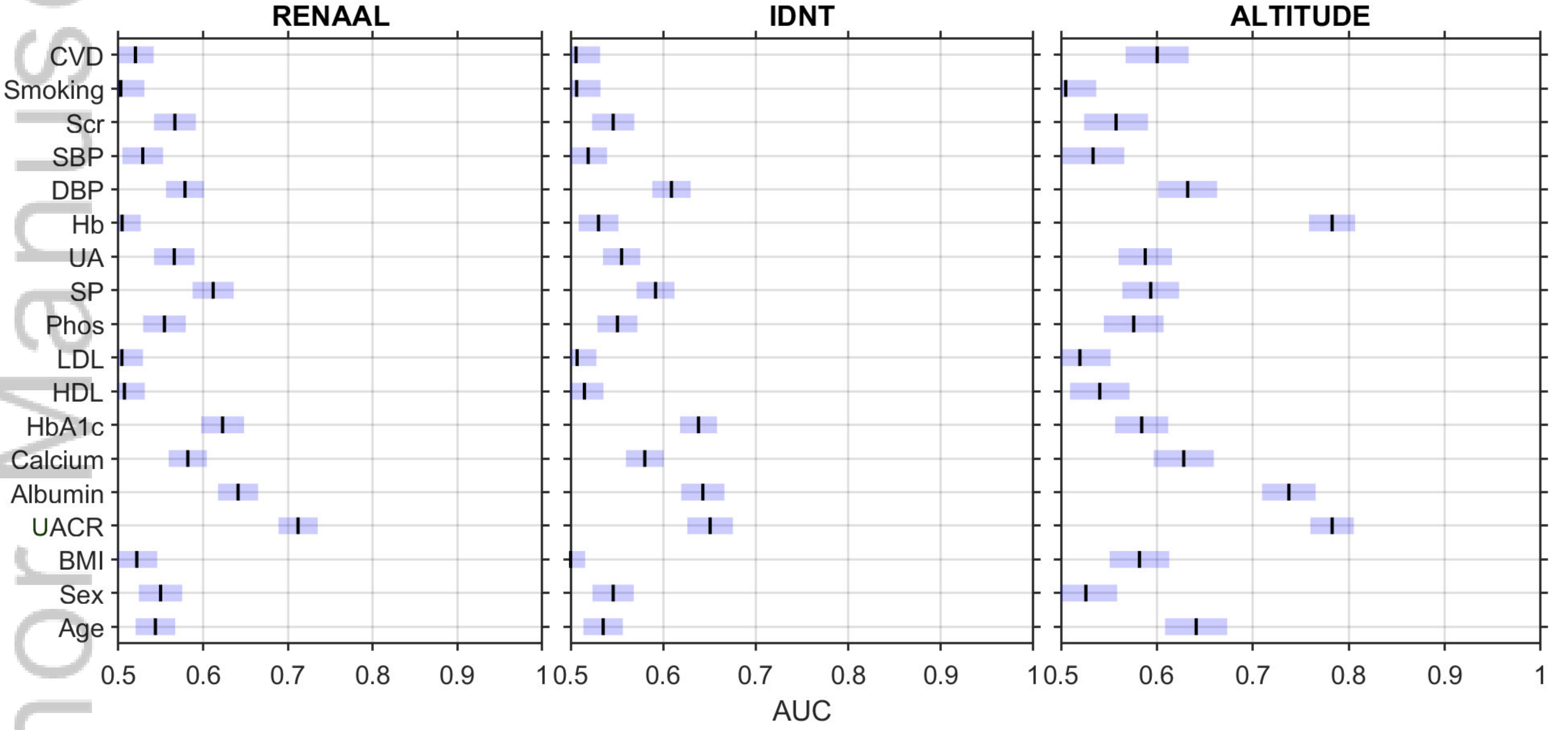
Figure 4: Risk calibration plots for FNN and KFRE models to predict ESRD events in RENAAL and IDNT trials. The calibration plot of FNN model is closer to the identity (or

diagonal) when compared to the KFRE model. Abbreviations: ESRD = end-stage renal disease; KFRE = Kidney Failure Risk Equation; FNN = Feed-forward neural network.

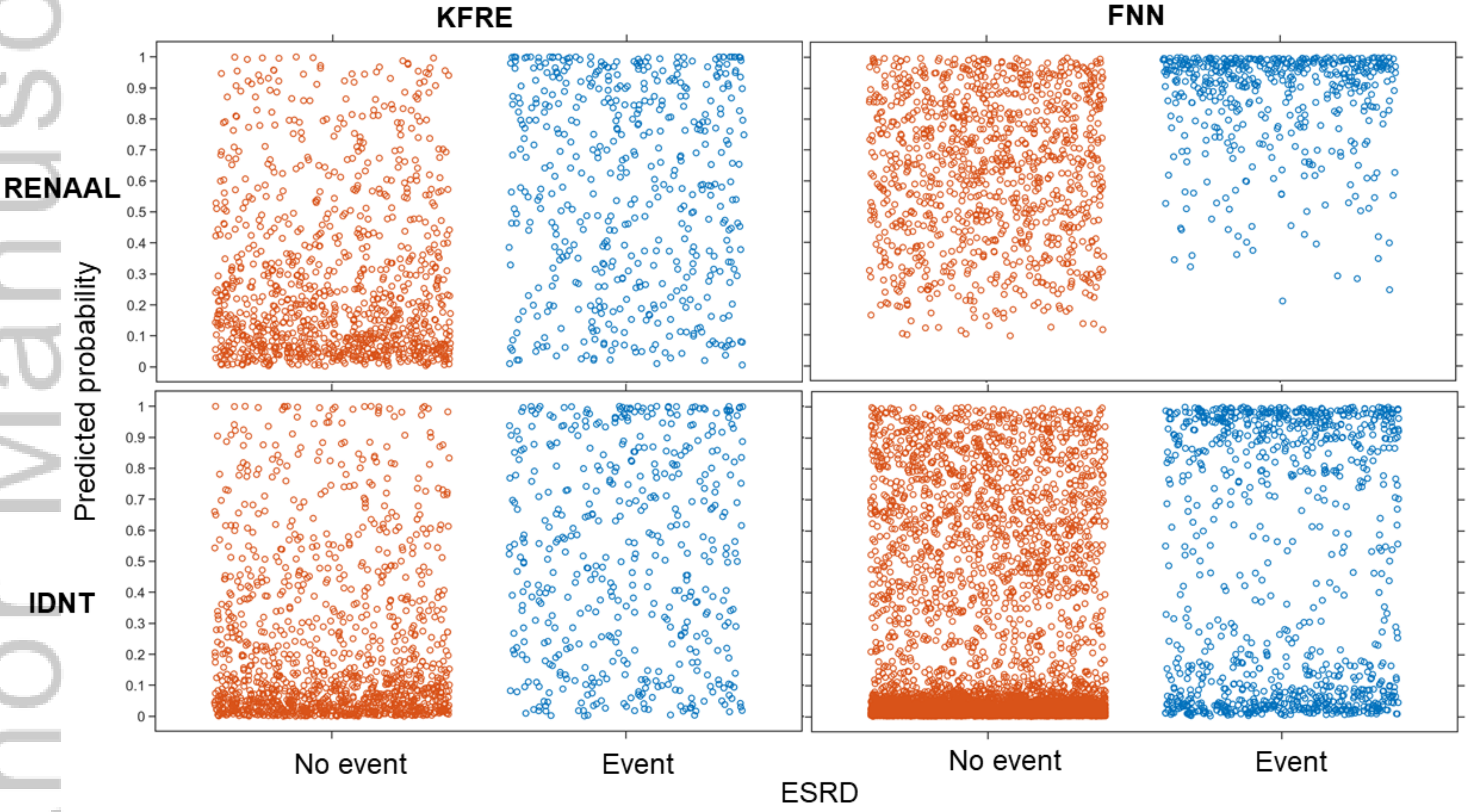
Supplementary figure 1: Heatmap illustrating the weights assigned to individual variables by the elastic-net regularization algorithm. The color bar indicates weights (normalized to 1 for the purpose of illustration) assigned by elastic-net algorithm: higher the intensity more predictive is the variable. Variables selected by the EN algorithm are represented by vertical bars in blue color. Unselected variables are shown as white bars. Abbreviations: IA = model trained on IDNT +ALTITUDE; RA = model trained on RENAAL +ALTITUDE; RI = model trained on RENAAL+ IDNT.



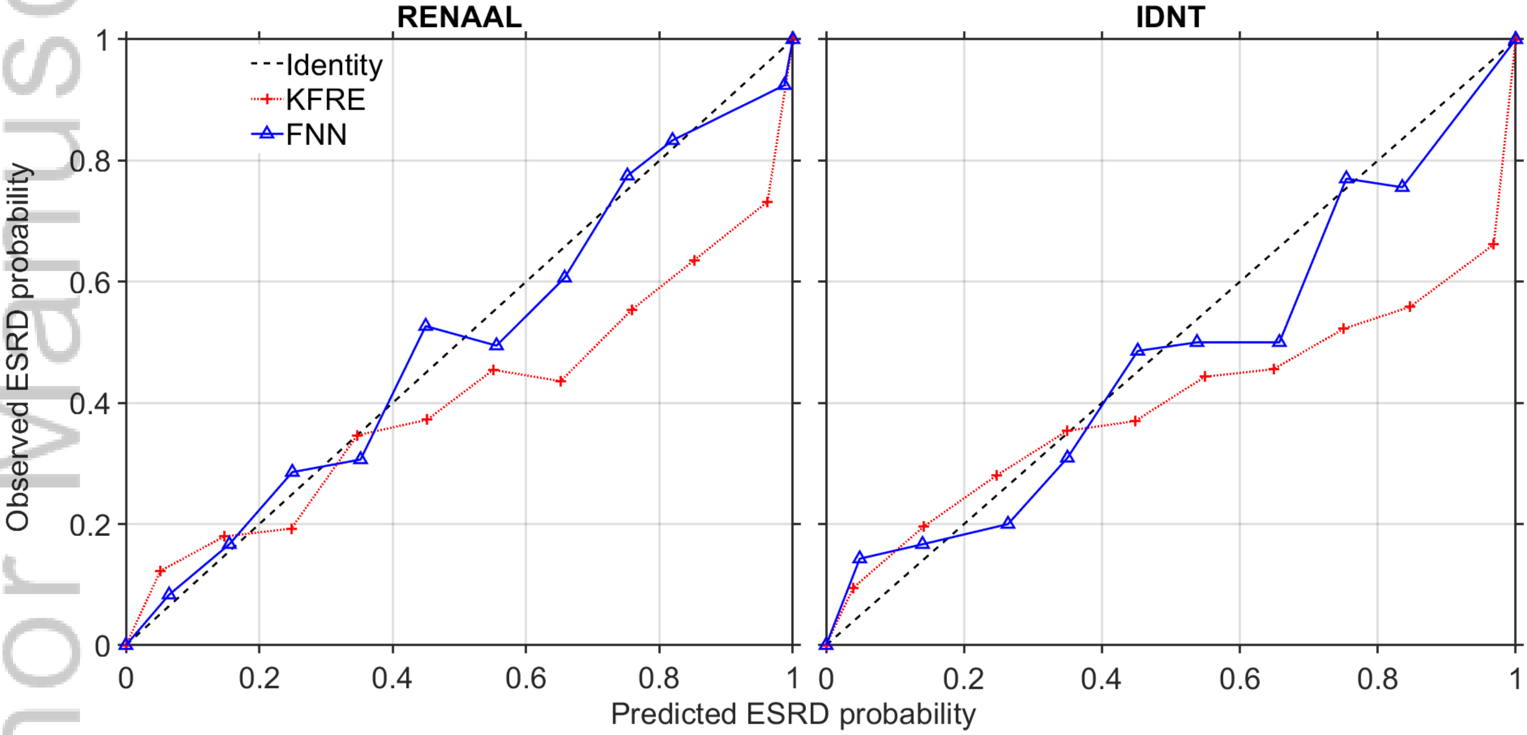
DOM_14178_dom-20-0672-op-File002.tif



DOM_14178_dom-20-0672-op-File003.tif



DOM_14178_dom-20-0672-op-File004.tif



DOM_14178_dom-20-0672-op-File005.tif

Table 1: Comparison of renal risk prediction performance (mean AUC (95% CI)) using classical machine learning algorithms for different datasets. The feed-forward neural network model significantly outperformed other machine learning and traditional techniques using baseline clinical variables. Due to unavailability of serum bicarbonate, we could not predict renal risk using KFRE model in the ALTITUDE trial. The performance of the feed forward neural network model was significantly better than the Cox proportional hazard regression (p-value=0.007, 0.006 and 0.01) and KFRE (p-value=0.001, 0.003, and NA) models for RENAAL, IDNT and ALTITUDE, respectively.

Testing data **Classifier**	**RENAAL**	**IDNT**	**ALTITUDE**
Logistic regression	0.77 (0.72-0.82)	0.76 (0.68-0.81)	0.78 (0.74 – 0.85)
Support vector machine	0.78 (0.71-0.85)	0.78 (0.70-0.83)	0.81 (0.71 – 0.85)
Random forest	0.80 (0.72-0.86)	0.79 (0.71-0.83)	0.82 (0.71 – 0.89)
Feed-forward neural network	**0.82** **(0.76-0.87)**	**0.81** **(0.75-0.86)**	**0.84** **(0.79 – 0.90)**
Cox proportional hazard regression	0.74 (0.73-0.75)	0.74 (0.73-0.75)	0.78 (0.77-0.79)
KFRE model	0.77 (0.74-0.79)	0.76 (0.73-0.79)	NA