ORIGINAL ARTICLE

WILEY

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] (ORCID) | Michelle J. Pena PhD[1] (ORCID) | Wenjun Ju PhD[2] |
Hiddo L. Heerspink PhD[1,3] (ORCID) | 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, Michigan

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

Correspondence
Sunil Belur Nagaraj, PhD, Department of Clinical Pharmacy and Pharmacology, University of Groningen, University Medical Center Groningen, De Brug 1D – 1-019 9700AD, Groningen, the Netherlands.
Email: sbn1984@gmail.com

Abstract

Aim: To predict end-stage renal disease (ESRD) in patients with type 2 diabetes by using machine-learning models with multiple baseline demographic and clinical characteristics.

Materials and methods: In total, 11 789 patients with type 2 diabetes and nephropathy from three clinical trials, RENAAL (n = 1513), IDNT (n = 1715) and ALTITUDE (n = 8561), 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 ESRD). We used the area under the receiver operator curve (AUC) to assess the prediction performance of models and compared this with 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 the RENAAL, IDNT and ALTITUDE trials, respectively. The feed forward neural network model selected urinary albumin to creatinine ratio, serum albumin, uric acid and serum creatinine as important predictors and obtained a state-of-the-art performance for predicting long-term ESRD.

Conclusions: Despite large inter-patient variability, non-linear 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 has the potential to create accurate and multiple outcome prediction automated models to identify high-risk patients who could benefit from therapy in clinical practice.

KEYWORDS
clinical trial, cohort study, diabetes complications, diabetic nephropathy, type 2 diabetes

1 | INTRODUCTION

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

In current practice, albuminuria[5] and estimated glomerular filtration rate (eGFR)[6] are the main predictors of progression of DKD. However, a recent study suggests that the margin of error for all eGFR formulae is high, thus making it a less reliable tool with which to assess 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 last decade has seen a major rise in computational processes for predictive analytics using machine-learning techniques. Unlike traditional statistical approaches where preselected 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 trained and validated within the same dataset are unlikely to generalize well because of patient heterogenity and demographic differences.[16]

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

2 | MATERIALS AND METHODS

2.1 | Study population

For the present study, we used data from three clinical trials, namely, RENAAL (n = 1513), IDNT (n = 1715) and ALTITUDE (n = 8561).

The detailed design, rationale and study outcomes for these trials have been published.[2,3,17] In RENAAL and IDNT, the effect of two ARBs, losartan and irbesartan, upon renal outcomes was 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 and 3.0 mg/dL. All patients had proteinuria, defined as a urinary albumin to creatinine ratio (UACR) of more than 300 mg/g based on single first morning void or a 24-hour urinary protein excretion of more than 500 mg/day in the RENAAL trial and more than 900 mg/day in the IDNT trial. In both trials eGFR was calculated using the Modification of Diet in Renal Disease 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 ACEis or ARBs in RENAAL, or ACEis, ARBs or calcium channel blockers in IDNT) were allowed during the trial to achieve the target level of 135/85 mmHg or less for RENAAL or 140/90 mmHg or less for IDNT.

In the ALTITUDE trial, 8561 type 2 diabetes patients with a high risk of renal and cardiovascular events from 854 centres 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/1.73m^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.

2.2 | 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, HbA1c, serum albumin, serum calcium, phosphorous, serum uric acid, high-density lipoprotein (HDL), low-density lipoprotein (LDL), UACR and a history of cardiovascular diseases. Each trial measured all serum and urine samples in a central laboratory. It should be noted that although we did not use eGFR directly as an input variable to the machine-learning model, we did use all the variables which are used for eGFR calculations, that is, serum creatinine, age and sex in the machine-learning model. In this way, the machine-learning model identifies a non-linear relationship between these variables and other variables for predicting ESRD, instead of a linear relationship as used in traditional eGFR calculations.

2.3 | 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.

2.4 | 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:

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

where true positive (TP) is the number of correctly classified patients with ESRD, false positive (FP) is the number of incorrectly classified patients with ESRD and false negative (FN) is the number of incorrectly classified patients without ESRD. Similar to the AUC, precision, recall and F-score values of 1.0 indicate accurate classification. In addition, we also obtained calibration points of the 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 of less than .05 was considered significant.

2.5 | Statistical analysis

The architecture of the proposed machine-learning–based ESRD prediction system is shown in Figure 1. First, we used the k-nearest neighbour algorithm[20] to impute missing variables in both the training and testing sets. The percentage number of variables imputed using this technique is summarized in Table S1. Because the training dataset consisted of an 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. Because of this, we created a balanced training set by using the Synthetic Minority Oversampling Technique (SMOTE) algorithm.[21] Variables in the training set were standardized by subtracting the mean and dividing by the standard deviation to calculate the unit mean and standard deviation. Testing set variables were standardized with respect to the mean and the 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 an elastic-net regularization algorithm to tune the hyperparameters of the machine-learning models (Table S2).

Because five different classification models were obtained as a result of 5-fold cross-validation, we repeated this process 1000 times to obtain 5000 models (1000 iterations of 5-fold cross-validation). Because different classification models are obtained for every hyperparameter combination and during every training fold, the model which 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 an almost unbiased estimate of the classification model as only training data were 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, a support vector machine with Gaussian kernel, and random forest and feed forward neural networks (FNN) to predict ESRD. We performed the 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 so as to include a large number of patients with ESRD for training
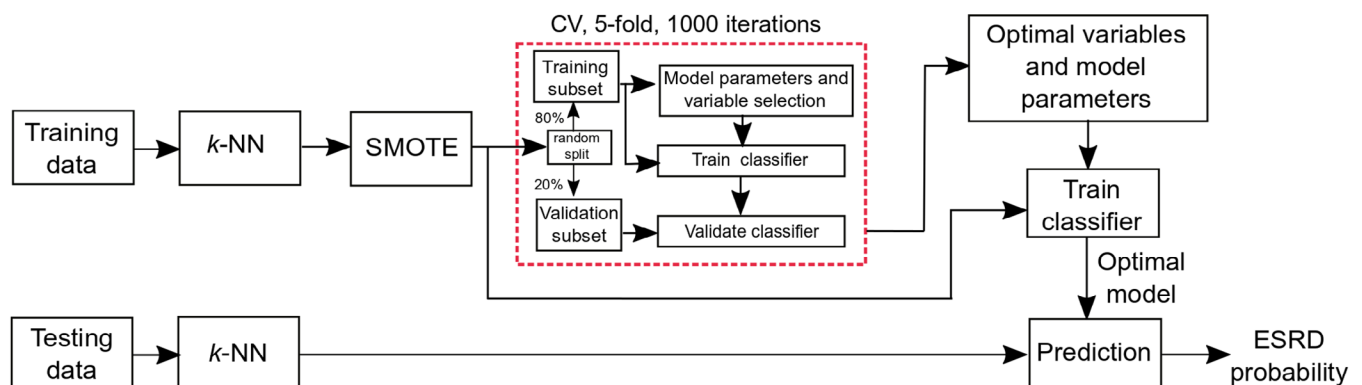

FIGURE 1 Architecture of the proposed ESRD prediction system. Rigorous cross-validation was performed to identify optimal model to predict renal risk in the testing set. CV, cross-validation; ESRD, end-stage renal disease; k-NN, k nearest neighbour; SMOTE, synthetic minority oversampling technique

the model. We also compared the performance of machine-learning models with traditional Cox proportional hazards regression and kidney failure risk equation (KFRE) models.[9] In the KFRE model, we used age, sex, UACR, eGFR, bicarbonate, phosphorus, albumin and calcium variables to estimate the ESRD probability. Because bicarbonate was not present in the ALTITUDE data, we did not estimate KFRE ESRD probability in those data.

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

3 | RESULTS

In total, there were 489, 283 and 508 patients with ESRD in the 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 (AUC = 0.65 [0.63-0.67]). In ALTITUDE, UACR (AUC = 0.77 [0.74-0.79]) and haemoglobin (AUC = 0.77 [0.72-0.80]) provided the best prediction performance compared with 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 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 the 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 <.05) than the traditional Cox regression and KFRE models in all three datasets. Additional performance metrics are provided in Table S3. 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 .5 for equal weightage for the two groups and estimated the mean Euclidean distance[22] between the probability scores of less than .5 (without ESRD) and probability scores of .5 or higher (with ESRD). The separation of predicted probabilities between two groups using FNN (Euclidean distance: RENAAL = 0.66, IDNT = 0.68) was higher compared with that of 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 compared with the KFRE in both RENAAL and IDNT. However, there was no significant difference between the calibration plots of FNN and KFRE (P-value = .1 and .2 for RENAAL and IDNT, respectively).

Figure S1 shows the heatmap of variables selected by the elastic-net regularization algorithm. Different numbers of variables were selected by the algorithm for different training and validation steps, and in total seven (age, UACR, serum albumin, serum uric acid, haemoglobin, SBP and serum creatinine), eight (age, UACR, serum albumin,
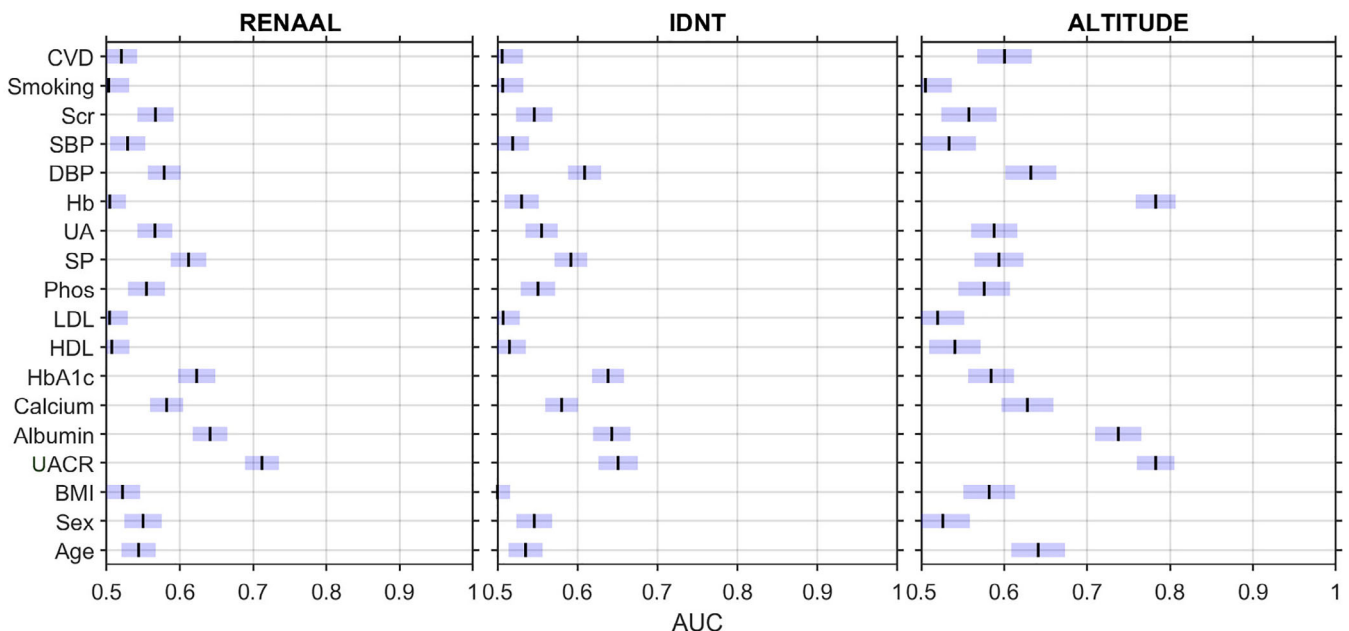

FIGURE 2 The distribution of AUC (mean [95% CI]) to predict ESRD using individual variables in all three clinical trials. Solid vertical black line corresponds to the mean AUC and rectangular box represents the standard deviation. Albumin, serum albumin; ACR, urine albumin-creatinine ratio; AUC, area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

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. Because of the 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 = .007, .006 and .01) and KFRE (P-value = .001, .003 and NA) models for RENAAL, IDNT and ALTITUDE, respectively

| Classifier	Testing data		
	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
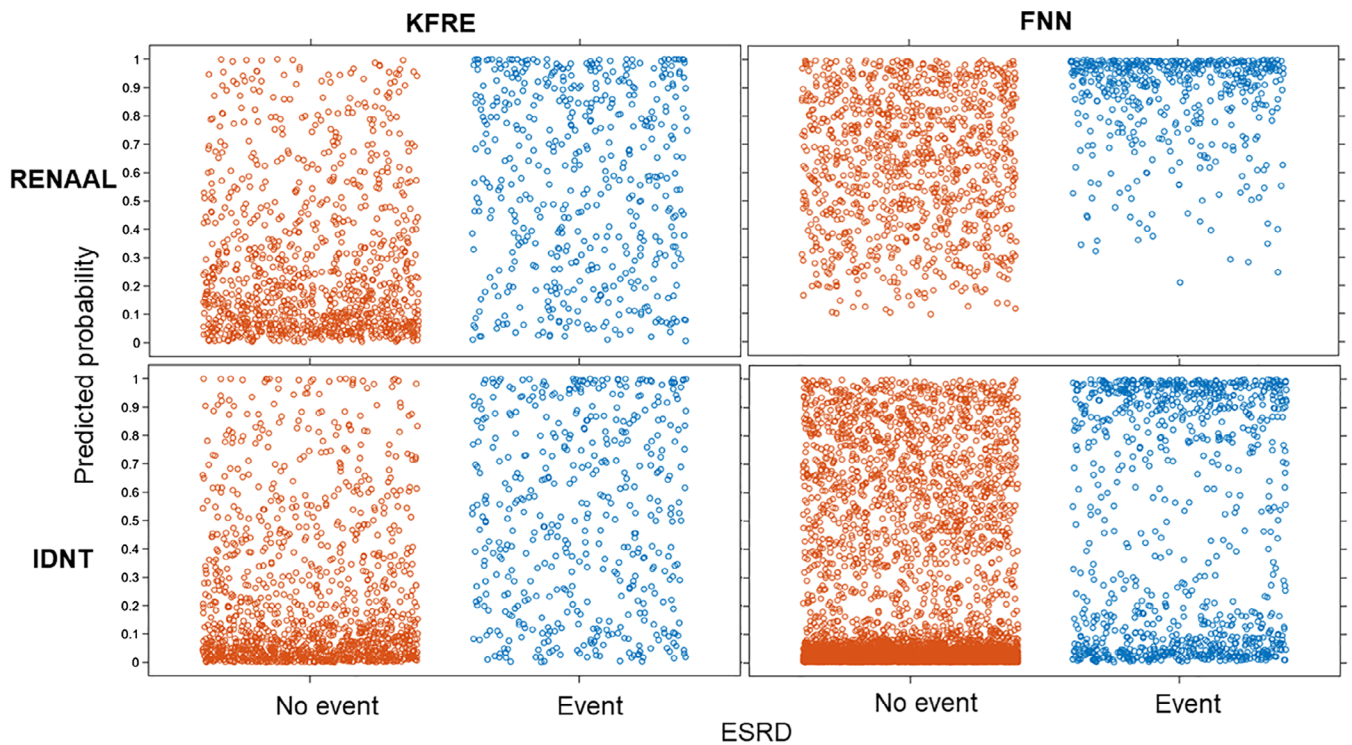
FIGURE 3 Plot showing the distribution of the predicted ESRD risk probability in patients with and without ESRD events for all three clinical trials. Jittering was performed for the ESRD event for better visualization. The best performing machine-learning model (FNN) is compared with the best performing 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. ESRD, end-stage renal disease; FNN, feed-forward neural network; KFRE, kidney failure risk equation

phosphorous, serum uric acid, haemoglobin, SBP and serum creatinine) and five (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 of >0.3 was used as per the convention of important 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. Table S4 summarizes the prediction performance. There was no significant difference (P-value >.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 with external
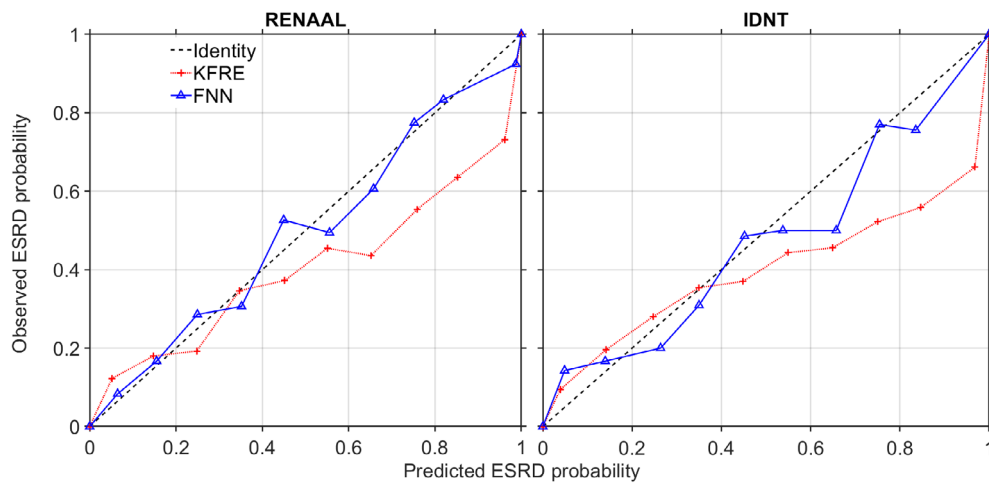
FIGURE 4 Risk calibration plots for FNN and KFRE models to predict ESRD events in RENAAL and IDNT trials. The calibration plot of FNN model is closer to the identity (or diagonal) when compared with the KFRE model. ESRD, end-stage renal disease; FNN, feed-forward neural network; KFRE, kidney failure risk equation

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), which was much better than the AUC obtained during external validation. This increase in the prediction performance was caused by the random inclusion of few patients from the testing set during the model training process, which can severely bias the prediction performance.

4 | 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 showed good prediction ability (AUCs greater than 0.8 in three clinical trials) and outperformed other machine-learning and traditional risk prediction models that 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 will probably generalize well on the new dataset. Such a model can also be used as an early prediction tool to identify patients who could benefit from intensified therapy in clinical practice.

The findings of this study have four important implications. First, we show that individual clinical variables are not sufficient to accurately predict long-term ESRD outcomes. 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 were not sensitive whether the patient was treated with placebo or ARBs, suggesting that the developed algorithm can be used

for predicting ESRD for any individual regardless of the renin-angiotensin-aldosterone system 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, clinical trials.

Although several machine-learning–based methods have already been developed to predict renal diseases in individuals with CKD,[10–14] a fair comparison is difficult because of (a) variability within datasets, (b) methodological differences to develop prediction models and (c) external validation. Differences in datasets can be attributed to the heterogeneity of disease severity and drug response, either from observational studies or clinical trials. Methodological differences can arise because of 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 an exhaustive grid search over a wide 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 compared with 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. First, a sample

size of 11 789 patients may not be sufficient to capture the large heterogeneity of disease severity seen in patients. Second, we used data from clinical trials, which is both a strength of and a limitation to our study. It represented a strength because of minimal variability in the clinical measurements, random assignment of patients to the treatment, timely assessment of endpoints, and inclusion of patients from multiple countries and centres capturing demographic heterogeneity. However, this was also a limitation because the developed machine-learning model does not take into account the variability in medication adherence which is commonly seen in observational data. Third, the machine-learning model did not achieve perfect prediction performance (i.e. AUC = 1.0). We hypothesize that further improvements can be obtained by including (a) additional molecular and cellular biomarkers and (b) increasing the overall sample size for training the FNN model. Fourth, these data are only analysed in a clinical trial setting. Validating the algorithms in a real-world setting should be addressed in future to determine the true generalizabilty to a non-clinical trial, type 2 diabetes general population.

In conclusion, 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 compared with other machine-learning models. The findings of this study pave the way to develop accurate and stable next-generation machine-learning–based ESRD prediction systems for clinical practice to identify high-risk patients who could benefit from therapy.

CONFLICT OF INTEREST

HLH 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. The other authors have no conflicts of interest to declare.

AUTHOR CONTRIBUTIONS

SBN designed and developed the machine learning analysis and algorithm development. SBN and MJP wrote the first draft of the manuscript. SBN, MJP, WJ and HLH participated in the interpretation of the data and revised the manuscript critically. SBN performed statistical analysis. All authors gave final approval to submit the article for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/dom.14178.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study. The data that support the findings of this study are available upon reasonable request from the steering committee of RENAAL, IDNT, and ALTITUDE clinical trial studies.

ORCID

Sunil Belur Nagaraj https://orcid.org/0000-0002-6409-4101
Michelle J. Pena https://orcid.org/0000-0003-3340-2893
Hiddo L. Heerspink https://orcid.org/0000-0002-3126-3730

REFERENCES

1. Ghaderian SB, Hayati F, Shayanpour S, Beladi Mousavi SS. Diabetes and end-stage renal disease; a review article on new concepts. *J Renal Inj Prev.* 2015;4(2):28-33.
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.
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.
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.* 2007;370(9590):829-840.
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.
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.* 2019;30(9):1756-1769.
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.
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.
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.
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.
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.* 2017;24(6):1052-1061.
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.
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.* 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.
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.
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.
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.
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, Association for Computing Machinery, New York, NY.

20. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak.* 2016;16(Suppl 3):197-208.

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.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Belur Nagaraj S, Pena MJ, Ju W, Heerspink HL, The BEAt-DKD Consortium. Machine-learning–based early prediction of end-stage renal disease in patients with diabetic kidney disease using clinical trials data. *Diabetes Obes Metab.* 2020;22:2479–2486. https://doi.org/10.1111/dom.14178