Clinical Kidney Journal, 2023, vol. 16, no. 3, 549–559

https://doi.org/10.1093/ckj/sfac252
Advance Access Publication Date: 7 December 2022
Original Article

ORIGINAL ARTICLE

Comparison of a chronic kidney disease predictive model for type 2 diabetes mellitus in Malaysia using Cox regression versus machine learning approach

Ruth Sim[1], Chun Wie Chong[1], Navin Kumar Loganadan[2], Noor Lita Adam[3], Zanariah Hussein[4] and Shaun Wen Huey Lee[1,5,6,7]

[1]School of Pharmacy, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, Subang Jaya, Selangor, Malaysia, [2]Department of Pharmacy, Putrajaya Hospital, Ministry of Health Malaysia, Jalan P9, Presint 7, Putrajaya, Wilayah Persekutuan Putrajaya, Malaysia, [3]Department of Medicine, Hospital Tuanku Jaafar, Ministry of Health Malaysia, Jalan Rasah, Bukit Rasah, Seremban, Negeri Sembilan, Malaysia, [4]Department of Medicine, Putrajaya Hospital, Ministry of Health Malaysia, Jalan P9, Presint 7, Putrajaya, Wilayah Persekutuan Putrajaya, Malaysia, [5]School of Pharmacy, Faculty of Health and Medical Sciences, Taylor's University, 1, Jln Taylors, Subang Jaya, Selangor, Selangor, Malaysia, [6]Asian Centre for Evidence Synthesis in Population, Implementation and Clinical Outcomes (PICO), Health and Well-being Cluster, Global Asia in the 21st Century (GA21) Platform, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, Subang Jaya, Selangor, Malaysia and [7]Center for Global Health, University of Pennsylvania, Philadelphia, PA, USA

Correspondence to: Shaun Wen Huey Lee; E-mail: shaun.lee@monash.edu

ABSTRACT

Background. Diabetes is one of the leading causes of chronic kidney disease (CKD) and end-stage renal disease. This study aims to develop and validate different risk predictive models for incident CKD and CKD progression in people with type 2 diabetes (T2D).

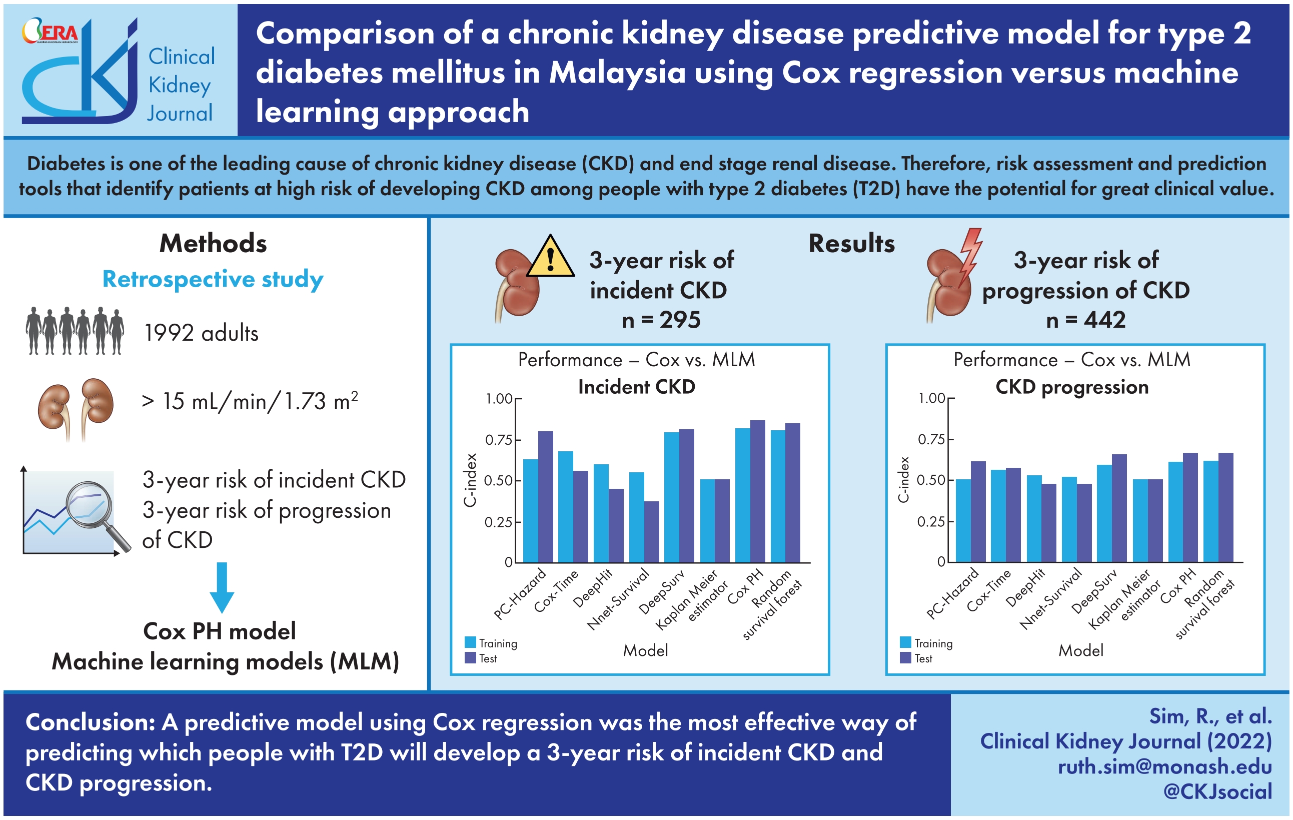
Methods. We reviewed a cohort of people with T2D seeking care from two tertiary hospitals in the metropolitan cities of the state of Selangor and Negeri Sembilan from January 2012 to May 2021. To identify the 3-year predictor of developing CKD (primary outcome) and CKD progression (secondary outcome), the dataset was randomly split into a training and test set. A Cox proportional hazards (CoxPH) model was developed to identify predictors of developing CKD. The resultant CoxPH model was compared with other machine learning models on their performance using C-statistic.

Results. The cohorts included 1992 participants, of which 295 had developed CKD and 442 reported worsening of kidney function. Equation for the 3-year risk of developing CKD included gender, haemoglobin A1c, triglyceride and serum creatinine levels, estimated glomerular filtration rate, history of cardiovascular disease and diabetes duration. For risk of CKD progression, the model included systolic blood pressure, retinopathy and proteinuria. The CoxPH model was better at prediction compared with other machine learning models examined for incident CKD (C-statistic: training 0.826; test 0.874) and CKD progression (C-statistic: training 0.611; test 0.655). The risk calculator can be found at https://rs59.shinyapps.io/071221/.

Received: 6.7.2022; Editorial decision: 7.11.2022

Conclusions. The Cox regression model was the best performing model to predict people with T2D who will develop a 3-year risk of incident CKD and CKD progression in a Malaysian cohort.

GRAPHICAL ABSTRACT



Clinical Kidney Journal

Comparison of a chronic kidney disease predictive model for type 2 diabetes mellitus in Malaysia using Cox regression versus machine learning approach

Diabetes is one of the leading cause of chronic kidney disease (CKD) and end stage renal disease. Therefore, risk assessment and prediction tools that identify patients at high risk of developing CKD among people with type 2 diabetes (T2D) have the potential for great clinical value.

Methods

Retrospective study

1992 adults

> 15 mL/min/1.73 m^2

3-year risk of incident CKD
3-year risk of progression of CKD

Cox PH model
Machine learning models (MLM)

Results

3-year risk of incident CKD
n = 295

3-year risk of progression of CKD
n = 442

Conclusion: A predictive model using Cox regression was the most effective way of predicting which people with T2D will develop a 3-year risk of incident CKD and CKD progression.

Sim, R., et al.
Clinical Kidney Journal (2022)
ruth.sim@monash.edu
@CKJsocial

Keywords: chronic kidney disease, diabetes, machine learning, prediction

INTRODUCTION

Diabetes is one of the most common metabolic disorders worldwide, affecting an estimated 537 million people in 2021 [1, 2]. In Malaysia, diabetes currently affects one in every six individuals or 3.9 million adults aged 18 and above, with type 2 diabetes (T2D) accounting for >90% of the people affected [3]. In people with T2D, there is a potential risk of developing long-term microvascular and macrovascular complications [3, 4]. Importantly, diabetes is the leading cause of chronic kidney disease (CKD) and end-stage renal disease in Malaysia, resulting in a significant economic burden to the healthcare system [5–7]. Compared with the general population, people with T2D are at a higher risk of developing nephropathy. As such, the ability to identify people with diabetes at risk of developing CKD may prevent adverse health outcomes associated with CKD. The development of risk prediction tools and screening of CKD for T2D patients allows concerted effort and prioritization of healthcare services to this highest risk group. The risk equation tool can be used to help physicians identify people at risk of reduced kidney function, enabling early and appropriate patient care through better targeted strategies.

Several prediction models to identify individuals at risk of developing CKD have been developed [8–10]. These studies reported that individuals who were older in age, of Black, American Indian, Hispanic or Asian race, who had high albuminuria levels, high haemoglobin A1c (HbA1c), and comorbidities including hypertension, dyslipidaemia, obesity and smoking history were at higher risk of developing CKD [11, 12]. However, very few studies have developed a prediction model for Asians with T2D [13, 14].

Recently, the advancement in machine learning methods, especially deep learning techniques, has provided researchers with alternative methods to examine the complex relationships between variables and hazard prediction, in addition to conventional survival models [15]. However, the performance of these models compared with traditional regression methods has not been explored, especially for an Asian population. In this study, we aim to develop a risk prediction model for incident CKD and CKD progression, and compare the resultant prediction model against those using machine learning for people with T2D.

MATERIALS AND METHODS

Study population

The demographic information of people who sought care between January 2012 and 10 May 2021 from two tertiary hospitals was extracted from medical records. These two tertiary hospitals were located in metropolitan cities of the state of Selangor and Negeri Sembilan, which provide diabetes clinics with endocrinologists and diabetes pharmacists. To be eligible for inclusion, they had to be diagnosed with T2D; aged 18 years and above; on oral glucose-lowering therapies such as sulfonylureas, dipeptidyl peptidase-4 inhibitors or sodium-glucose cotransporter 2 inhibitors; with an estimated glomerular filtration rate (eGFR) >15 mL/min/1.73 m^2; and meet the research quality standard with linked laboratory data. Diagnosis of T2D was based on Malaysia Clinical Practice Guidelines for Management of T2D [16]. People with type 1 diabetes, those with a history of gestational diabetes, or a history of pregnancy within the last 12 months were excluded. Patients were followed until the outcomes of interest were developed, or if they were lost to follow-up prior to 10 May 2021, and the rest were censored on 10 May 2021.

Exposure variables

Potential candidate variables for predicting risk of incident CKD or CKD progression among people with T2D were identified from literature [12–14, 17, 18]. Variables identified include: age, family history of diabetes, hypertension, dyslipidaemia, smoking status, retinopathy, neuropathy, years since diagnosis of T2D, proteinuria (positive urine dipstick of ≥1+), race, weight, body mass index (BMI), glycated haemoglobin (HbA1c), high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglycerides (TG), total cholesterol (TC), urea, serum creatinine, eGFR, systolic blood pressure (SBP), diastolic blood pressure (DBP), sex, history of established cardiovascular disease (CVD) [ischaemic heart disease (IHD), myocardial infarction (MI), heart failure (HF), stroke (CVA)], renin-angiotensin-aldosterone system inhibitors (RAASi) and nonsteroidal anti-inflammatory drug (NSAIDS). Baseline characteristics were recorded within 365 days from the first eGFR measurement.

Missing data

To account for missing data, we used multiple imputation with chained equations under the assumption of data being missing at random (MAR). We generated five imputed datasets using classification and regression tree 'cart'. For missingness in continuous variables, years since diagnosis of T2D, weight, BMI, HbA1c, HDL, LDL, TG, TC, urea, serum creatinine, eGFR, SBP and DBP, we imputed the variables on the continuous scale and then converted the variables to categorical variables after imputation (missing data and data after multiple imputation are shown in Supplementary data, Table S4). Several variables including urine albumin–creatinine ratio, urine creatinine and urine albumin were unavailable in >75% of cases as these were often not routinely collected and thus were not considered in the model.

Outcome definition

Incident CKD was defined as the time from index date to an eGFR reading of <60 mL/min/1.73 m^2. The eGFR was calculated using the 2009 CKD Epidemiology Collaboration creatinine equation, as endorsed by Malaysia Clinical Practice Guidelines for Management of T2D [16]. CKD progression was defined as the time from index date to eGFR decline, defined as worsening of eGFR categories (stage 1, ⩾90 mL/min/1.73 m^2; stage 2, 60–89 mL/min/1.73 m^2; stage 3a, 45–59 mL/min/1.73 m^2; stage 3b, 30–44 mL/min/1.73 m^2; stage 4, 15–29 mL/min/1.73 m^2; and stage 5, <15 mL/min/1.73 m^2), coupled with a 25% or more reduction in eGFR from baseline according to the 2012 Kidney Disease: Improving Global Outcomes (KDIGO) Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease [19].

Model development

Data were randomly split into two, where 70% was allocated for developing the prediction model (training set, development set), and 30% for the evaluation of model performance (test set, validation set). For each set, summary statistics for patient demographics were presented with as means with standard deviations (SD) or median with interquartile range (IQRs) for continuous data and total numbers and percentages for categorical data. Continuous variables were compared using the Student's *t*-test, while categorical variables were compared using the χ^2 test. The Cox proportional hazard model was used to develop the incident CKD/CKD progression risk prediction models. We used a combination of clinical guidance and forward selection to determine variable selection. For candidate predictors listed above, univariate Cox analysis was applied. Predictors were excluded when *P*-values exceeded 0.2. In the second step, selected variables from the univariate analysis were entered into the multivariate model. Multivariate analysis with backward stepwise elimination was then conducted to develop the subsequent model. The sequential models are presented in Supplementary data, Table S5a and b. Model with the best discrimination and goodness of fit was selected. Hazard ratios (HRs) were calculated along with 95% confidence interval (CI). Proportional hazards assumption and linearity assumption were tested through examination of Schoenfeld residuals and Martingale residuals [20, 21]. All statistical analyses were performed using R version 4.1.1., mainly using the packages mice, survival, mlr3proba and survivalmodels [22–24]. Additional details are available in Supplementary data, Methods S1.

Cox prediction model performance

We used a series of methods to evaluate the performance of the models in the development and the validation data sets. Concordance statistics (C-statistics) for survival model developed by Nam were computed as measures of discrimination [25]. We then divided the cohort into quintiles of predicted risk of incident CKD/CKD progression and determined the magnitude of the deviation using the modified Nam and D'Agostino χ^2 statistic (GND test) for model calibration [26]. Calibration plots using training and test data were generated to assess the model's goodness of fit. We then determined the optimal cut-off threshold based upon the sensitivity and specificity, and positive and negative prediction values of the model. We repeated the analysis by conducting the 70/30 train test split based on calendar year. We trained the models on the older patients set and validating on the newer patients set. We also excluded BMI variable from the model for the second sensitivity analysis. For the third sensitivity analysis, we added chronicity criterion to the definition of incident CKD by defining incident CKD as eGFR reading <60 mL/min/1.73 m^2 persisting for >3 months.

Validation of other CKD disease models in the study training and test sets

Validation of CKD disease prediction models by Afghahi *et al*. [27] and Dagliati *et al*. [28] was conducted using our training and test data sets.

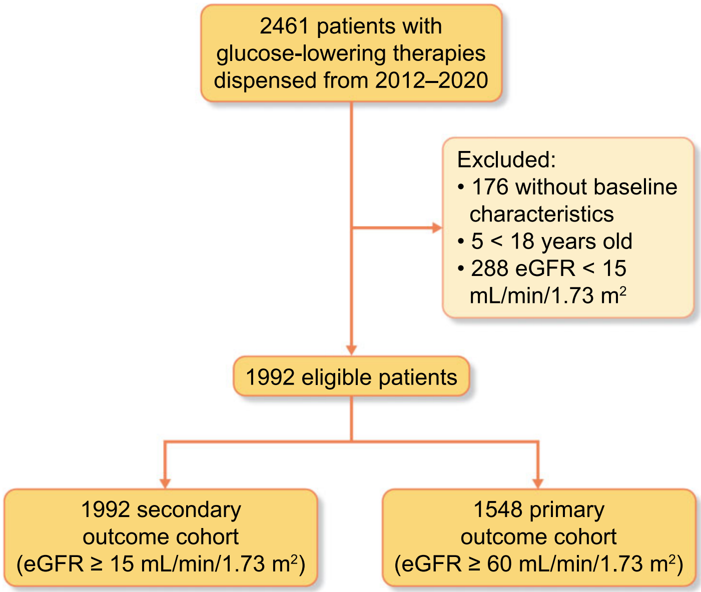

2461 patients with glucose-lowering therapies dispensed from 2012–2020
Excluded:
• 176 without baseline characteristics
• 5 < 18 years old
• 288 eGFR < 15 mL/min/1.73 m²
1992 eligible patients
1992 secondary outcome cohort (eGFR ≥ 15 mL/min/1.73 m²)
1548 primary outcome cohort (eGFR ≥ 60 mL/min/1.73 m²)

Figure 1: Study flowchart and patient selection.

Comparison of Cox model to other machine learning survival models

We then repeated our analyses using the following survival analysis models, including regression analysis (Cox proportional hazards), decision-tree based models (random survival forests), several neural-networks based models (Cox-Time, DeepHit, DeepSurv, Nnet-Survival, PC-Hazard) and the baseline estimator Kaplan–Meier Estimator with the R package mlr3proba and survivalmodels [23, 24]. The final predictors, training and test data included in the following survival models were the same as the final Cox model.

(i) Kaplan–Meier Estimator is a non-parametric model that uses data to estimate survival function without considering covariates in the data and any assumptions of survival curve distribution [29].

(ii) Cox proportional hazards (CoxPH) model is a semi-parametric model used to estimate hazard function considering a set of covariates [30]. The effects of covariates on survival time were evaluated.

(iii) Random forest is a supervised ensemble learning method that aggregates outcome of interest from predictions from many decision trees, each generated from from bootstrap sample of predictors [31]. Random survival forest extends Breiman's random forest by incorporating censoring information into the splitting rules when growing the trees [32].

(iv) DeepSurv is a further adaptation of CoxPH model using deep feed-forward neural network, which is a non-linear extension of CoxPH model that fulfils proportional-hazards assumption [33]. Neural network is a machine learning model whose model structure mimics the structure of animal brain using hidden layers and neurons in those hidden layers.

(v) Nnet-Survival is a fully parametric neural network model used to predict conditional survival probability in discrete time intervals [34].

(vi) Cox-time is a time-dependent neural net CoxPH model that uses neural networks as relative risk function to model relationships between time and covariates [35].

(vii) DeepHit is an advanced non-proportional discrete-time deep neural network that learns the distribution of survival times directly, including survival data with competing risks [36]. The model is able to learn time-varying effects of covariates on survival using the joint probability distribution of possible events at each time point.

(viii) Piecewise Constant Hazard (PC-Hazard) is a model which discretizes the continuous-time hazard function to be piecewise constant and uses neural networks to parameterize the hazard function [37].

Harrells's C-index, Integrated Graf Score, van Houwelingen's alpha calibration and D-calibration were used as performance evaluation metrics to compare these models to the reference Cox proportional hazards model in predicting incident CKD or CKD progression [38–41]. We compared the mean predicted survival curves for different models with the Kaplan–Meier curve using the test set. Additional models description, hyperparameters considered and performance metrics are listed in Supplementary data, Text S1 and Script S1 [38–41]. We added sensitivity analysis by varying the number of epochs while training the neural network models by increasing number of epochs to 100 and 1000.

Ethical and scientific approval

The study was approved by National Medical Research Register (NMRR-20-662-52570) and Monash Human Research Ethics Committee (2020-24900-45575). Ethical approval for this study

Table 1: Characteristics of people in the primary and secondary outcome cohort.

	Primary outcome: incident CKD	Secondary outcome: CKD progression	*P*-value	SMD
n	1548	1992		
Age, years	52.45 (13.20)	54.38 (13.24)	<.001	0.146
Family history of diabetes, *n* (%)	368 (23.8)	411 (20.6)	.028	0.076
Hypertension, *n* (%)	1066 (68.9)	1448 (72.7)	.014	0.084
Dyslipidaemia, *n* (%)	993 (64.1)	1262 (63.4)	.651	0.017
Smoking, *n* (%)	134 (8.7)	154 (7.7)	.349	0.034
Retinopathy, *n* (%)	342 (22.1)	527 (26.5)	.003	0.102
Neuropathy, *n* (%)	377 (24.4)	534 (26.8)	.106	0.056
Diabetes duration, years	10.68 (7.57)	15.06 (83.40)	.054	0.074
Proteinuria, *n* (%)	177 (11.4)	230 (11.5)	.96	0.004
Race, *n* (%)			.978	0.015
Others	56 (3.6)	73 (3.7)		
Indian	320 (20.7)	402 (20.2)		
Chinese	182 (11.8)	241 (12.1)		
Malay	990 (64.0)	1276 (64.1)		
Weight, kg	82.37 (22.06)	81.81 (21.19)	.535	0.026
BMI, kg/m^2	33.75 (8.40)	33.82 (8.55)	.901	0.008
HbA1c, %	8.63 (1.87)	8.62 (1.86)	.864	0.007
HDL, mmol/L	1.23 (0.31)	1.22 (0.30)	.619	0.02
LDL, mmol/L	2.77 (1.04)	2.73 (1.03)	.285	0.044
TG, mmol/L	1.80 (1.26)	1.82 (1.21)	.671	0.017
TC, mmol/L	4.80 (1.22)	4.77 (1.21)	.408	0.032
Urea, mmol/L	4.62 (3.75)	5.59 (5.07)	<.001	0.216
Serum creatinine, μmol/L	72.60 (18.14)	88.97 (43.39)	<.001	0.492
eGFR, mL/min/1.73 m^2	96.90 (27.29)	84.99 (33.19)	<.001	0.392
SBP, mmHg	139.92 (39.85)	140.67 (36.32)	.595	0.02
DBP, mmHg	78.85 (11.28)	78.28 (11.39)	.18	0.05
Male, *n* (%)			.775	0.024
No	800 (51.7)	1011 (50.8)		
Yes	720 (46.5)	940 (47.2)		
NA	28 (1.8)	41 (2.1)		
Established CVD, *n* (%)	383 (24.7)	578 (29.0)	.005	0.097
NSAIDS (%)			.093	0.074
No	1113 (71.9)	1365 (68.5)		
Yes	421 (27.2)	608 (30.5)		
NA	14 (0.9)	19 (1.0)		
RAASi (%)			.894	0.016
No	550 (35.5)	693 (34.8)		
Yes	984 (63.6)	1280 (64.3)		
NA	14 (0.9)	19 (1.0)		

Values are given as mean (SD) or as stated.
SMD, standardized mean difference.

was obtained from the Medical Research and Ethics Committee (MREC), Ministry of Health Malaysia.

RESULTS

Cohort description

A total of 2461 participants with T2D were screened, and after excluding 469 participants who did not fit the inclusion criteria, 1992 participants were included in this study (Fig. 1). Over a median follow-up of 1.7 years, 295 individuals developed CKD while 442 people experienced CKD stage progression (Table 1, Supplementary data, Table S2 and S3). In the incident CKD cohort, 195 of 1253 censored patients (15.6%) were lost to follow-up before the censored date. In the CKD progression prediction, 195 of 1550 censored patients (12.6%) were lost to follow-up before the censored date.

Model prediction performance

Of the 26 variables assessed (Table 2), seven were retained after multivariate analyses with stepwise selection (Table 3). After stepwise selection, a greater risk of incident CKD stages 3–5 was associated with history of cardiovascular disease, gender, HbA1c, TG and serum creatinine level, eGFR and diabetes duration. Similarly, predictors of CKD progression were: retinopathy, proteinuria and established CVD, HbA1c, TG, serum creatinine and SBP level at baseline. The proportional hazards and linearity assumptions were held for all variables (Supplementary data, Fig. S1a, b, and S2a, b).

Model performance

The 3-year risk of primary outcome (incident CKD) and secondary outcome (CKD progression) equations were developed

Table 2: Univariate and multivariate Cox regression for candidate predictor variables.

	Incident CKD				CKD progression			
	Univariate HR (95% CI)	*P*-value	Multivariate HR (95%)	*P*-value	Univariate HR (95%)	*P*-value	Multivariate HR (95% CI)	*P*-value
Age	1.04 (1.03–1.05)	<.001			1.01 (1–1.02)	<.01		
Family history of diabetes	0.458 (0.306–0.687)	<.001	0.71 (0.47–1.08)	.108	0.735 (0.553–0.977)	<.05		
Hypertension	1.47 (1.07–2.03)	<.05			1.43 (1.1–1.87)	<.01	1.25 (0.95–1.64)	.114
Dyslipidaemia	1.05 (0.79–1.4)	.732			0.931 (0.742–1.17)	.539		
Smoking status	0.767 (0.453–1.3)	.323			1.15 (0.784–1.68)	.478		
Retinopathy	1.56 (1.14–2.13)	<.01			1.81 (1.42–2.29)	<.001	1.44 (1.12–1.84)	<.01
Neuropathy	1.4 (1.04–1.9)	<.05			1.25 (0.968–1.61)	.087		
Diabetes duration	1.06 (1.04–1.07)	<.001	1.02 (1–1.04)	<.05	1.02 (1.01–1.04)	<.001		
Proteinuria	1.39 (0.919–2.09)	.119	1.42 (0.94–2.15)	.100	1.75 (1.26–2.43)	<.001	1.64 (1.18–2.29)	<.01
Race	0.947 (0.818–1.09)	.459			0.987 (0.875–1.11)	.835		
Weight	1 (0.993–1.01)	.881			0.994 (0.989–1)	<.05	1 (0.99–1)	.102
BMI	0.995 (0.977–1.01)	.629			0.99 (0.975–1)	.174		
HbA1c	1.05 (0.981–1.13)	.150	1.09 (1.01–1.17)	<.05	1.09 (1.03–1.15)	<.01	1.08 (1.01–1.14)	<.05
HDL	0.715 (0.426–1.2)	.204			1.05 (0.728–1.51)	.794		
LDL	0.977 (0.855–1.12)	.738			1.05 (0.942–1.16)	.404		
TG	1.1 (1.02–1.19)	<.05	1.14 (1.05–1.25)	<.01	1.13 (1.05–1.2)	<.001	1.11 (1.04–1.19)	<.01
TC	1.01 (0.904–1.14)	0.803			1.1 (1–1.2)	<.05		
Urea	1.02 (1–1.03)	<.01			1.01 (1–1.02)	<.01		
Serum creatinine	1.05 (1.04–1.06)	<.001	1.06 (1.04–1.07)	<.001	1.01 (1–1.01)	<.001	1 (1–1.01)	<.001
eGFR	0.936 (0.926–0.945)	<.001	0.98 (0.96–0.99)	<.01	0.991 (0.987–0.995)	<.001		
SBP	1 (0.999–1)	.221			1 (1–1)	<.05	1 (1–1)	<.05
DBP	0.99 (0.977–1)	.137			0.987 (0.977–0.998)	<.05	0.99 (0.98–1)	.082
Male	1.3 (0.991–1.71)	.058	0.3 (0.2–0.47)	<.001	0.918 (0.737–1.14)	.449		
Established CVD	1.71 (1.28–2.29)	<.001	1.41 (1.04–1.91)	<.05	1.6 (1.27–2.02)	<.001	1.37 (1.08–1.74)	<.05
NSAIDS	1.89 (1.43–2.49)	<.001			1.31 (1.04–1.64)	<.05		
RAASi	1.49 (1.1–2.03)	<.05			1.1 (0.867–1.38)	.446		

Table 3: Final multivariate Cox regression model.

	Incident CKD			CKD progression		
	Beta-coefficient	HR (95% CI)	P-value	Beta- coefficient	HR (95% CI)	P-value
HbA1c	0.080 426	1.08 (1.01–1.16)	<.05	0.064592	1.07 (1.01–1.13)	<.05
TG	0.122 041	1.13 (1.04–1.23)	<.01	0.099762	1.1 (1.03–1.19)	<.01
Serum creatinine	0.054342	1.06 (1.04–1.07)	<.001	0.004818	1 (1–1.01)	<.001
Established CVD	0.321619	1.38 (1.02–1.87)	<.05	0.357445	1.43 (1.13–1.82)	<.01
Diabetes duration	0.022046	1.02 (1–1.04)	<.05			
Retinopathy				0.410524	1.51 (1.18–1.93)	<.01
Proteinuria				0.521901	1.69 (1.21–2.35)	<.01
SBP				0.002056	1 (1–1)	.072
eGFR	−0.02404	0.98 (0.96–0.99)	<.001			
Male	−1.19698	0.3 (0.2–0.47)	<.001			

Table 4: Performance of Cox prediction model.

	Incident CKD		CKD progression	
	Training	Test	Training	Test
Discrimination	Concordance = 0.826 (se = 0.015)	Concordance = 0.874 (se = 0.016)	Concordance = 0.611 (se = 0.018)	Concordance = 0.655 (se = 0.029)
Calibration	χ^2 (P) 6.9963487 (0.1360813)	χ^2 (P) 0.5189252 (0.4713005)	χ^2 (P) 7.7028118 (0.5643507)	χ^2 (P) 15.60836405 (0.07552424)
Cut-off value	0.3204644	0.6324474	0.3072828	0.1938746
Sensitivity (%)	0.7368421	0.7558140	0.5125786	0.5806452
Specificity (%)	0.8045714	0.7857143	0.6202414	0.6384778
PPV (%)	0.4722222	0.4452055	0.2837128	0.2962963
NPV (%)	0.9263158	0.9339623	0.8106796	0.8531073

PPV, positive predictive value; NPV, negative predictive value.

based on each final Cox model using coefficients of each predictor (Table 3, Supplementary data, Table S1). The risk calculator can be found at https://rs59.shinyapps.io/071221/. The final model for primary outcome showed good discriminative performance in the training and test data set, with C-statistics of 0.826 [standard error (se) 0.015] and 0.874 (se 0.016), respectively, indicating good performance of the model in distinguishing people with T2D who developed stage 3 CKD from those who did not (Table 4, Fig. 2). However, the model for determining CKD progression had lower discrimination property, with C-statistic between 0.6 and 0.7. The final model for both outcomes showed no evidence of poor calibration ($P > .05$) and no collinearity (Supplementary data, Table S6). Calibration plots show adequate calibration (Supplementary data, Figure S3a–d). Sensitivity analysis showed that HbA1c, serum creatinine and eGFR level, gender, and history of CVD remained significant in the prediction of incident CKD, while HbA1c, TG and serum creatinine level, presence of retinopathy, and established CVD were significant in predicting CKD progression (Supplementary data, Tables S7–S9). Our dataset had comparable performance when validated using similar models from other countries (Supplementary data, Table S10).

Comparison of Cox model to other machine learning models

The Cox regression model achieved the best predictive performance for incident CKD (C-statistic: training 0.822; test 0.870) compared with the other machine learning models evaluated, among all D-calibrated models (Table 5). For CKD progression, the random survival forest gave comparable results to Cox model for C-statistic (C-statistic: training 0.619; test 0.666). Furthermore, Cox model gave the lowest Integrated Graf Score for CKD disease and CKD progression final models (Table 5, Supplementary data, Table S11). The alpha-calibration values of Cox model and random survival forest model were close to one. The mean predicted survival curve for Cox model was close to the Kaplan–Meier curve for both outcomes (Supplementary data, Figure S4a and b). The results remain robust in the sensitivity analysis using different number of epochs (Supplementary data, Table S12).

DISCUSSION

There has been increasing use of risk prediction models in clinical care to improve patient outcomes. Despite these benefits, there has been a lack of easily applicable and validated models for widespread integration. As screening the general population can be laborious and not cost effective, the use of risk prediction models and targeted screening of high-risk patients for CKD among T2D patients allows prioritization of healthcare intervention. This allows for targeted treatment to delay CKD progression among people with T2D that will delay the progression of diabetic nephropathy, which often incurs the highest cost and has greatest impact on patients' quality of life. In this study we developed and validated a set of prediction model for incident CKD and CKD progression for people with T2D in Malaysia. Our study found seven variables routinely collected in clinical care that could be used to predict the risk of incident CKD or CKD progression, suggesting that this can easily be implemented in

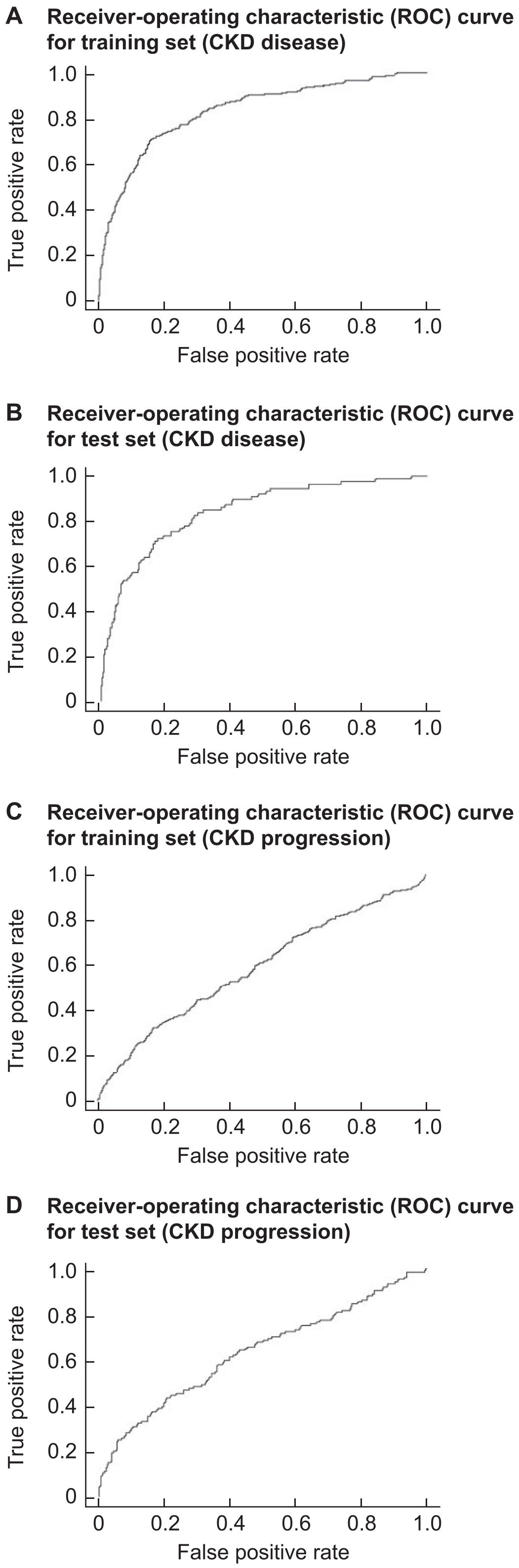


Figure 2: ROC curves for Cox prediction model. (**a**) ROC curve for training set (CKD disease). (**b**) ROC curve for test set (CKD disease). (**c**) ROC curve for training set (CKD progression). (**d**) ROC curve for test set (CKD progression). ROC, receiver-operating characteristic.

clinical practice and be integrated into electronic health records to identify people at increased risk of developing CKD. The results of our study are largely in line with other studies which reported that demographics such as gender as well as laboratory data such as HbA1c, serum creatinine and eGFR could be used in CKD prediction [13, 14, 18, 42]. Our study also identified several other risk factors, such as presence of retinopathy and TG levels, for incident CKD and CKD progression.

CKD prediction studies involving T2D patients in the Asian region had relatively well performance with C-index ranging from 0.80 to 0.89. For example, the study by Low *et al.* [13] in Singapore examined CKD progression prediction among T2D patients using logistic regression while Tuntayothin *et al.* 2021 [14] in Thailand examined CKD disease prediction using Cox model, both of which had a good performance. The performance of our model is comparable to the existing studies for the primary outcome incident CKD model (C-index: 0.61–0.81) and secondary outcome of CKD progression (C-index: 0.73–0.85), albeit slightly lower [10]. One reason for this variation could be differences in the definition of CKD progression.

Prediction models for incident CKD were often developed using the classical Cox model, but several new techniques including neural networks have been used. In our study, we benchmarked our classical Cox model to six additional machine learning models and one Kaplan–Meier model. Interestingly, we noted that the standard Cox regression showed the best C-statistic performance, despite various literature showing considerable improvement to disease prediction using deep learning methods [43–45]. Some possible explanations include the small number of predictor variables and linear interactions between variables [46]. In addition to that, the small sample size of our study and the small number of outcomes might account for the limited performance of the machine learning models. Indeed, as machine learning models work better if non-linearity and interactions exist in the data, hence they are not expected to perform better than the Cox model in disease prediction. We also applied variable selection techniques before using our Cox model, which has been shown to improve performance comparable to most machine learning algorithms. Nevertheless, as these machine learning algorithms will continue to advance in the future, the results will need to be revisited.

Strength and limitations

This study offers several strengths. It is the first study in Malaysia to investigate the risk factors of incident CKD in people with T2D. We also developed a model to assess the decline in eGFR. This has significant impact for the healthcare system, due to the large number of people with diabetes in Malaysia who subsequently develop CKD [7, 47]. Our model used basic demographics and clinical data which are easily available from routine medical records. Our study also conducted a benchmark of traditional Cox model against the state-of-art machine learning models for disease prediction. We have conducted rigorous cross-validation and hyperparameter tunings for the robust and stable performance of the machine learning models.

Our analysis had some limitations. First, the study follow-up was relatively short with a median of 1.63 years and we had a small sample size. The study was carried out in tertiary setting, where most people were at an advanced stage of diabetes with comorbidities. We did not perform an external validation with a separate cohort, which limits the generalizability of the results. Our model did not include socioeconomic and lifestyle factors such as physical activity, alcohol or traditional

Table 5: Comparison of the performance of the prediction models using final variables.

Outcome	Variables	Dataset	Model	C-index	Integrated Graf Score	van Houwelingen's Alpha	D-calibration P-value
Incident CKD	Final variables	Training	PC-Hazard	0.635	0.347	0.301	.000
			Cox-Time	0.680	0.311	1.870	.527
			DeepHit	0.601	0.282	0.683	.000
			Nnet-Survival	0.552	0.335	0.540	.266
			DeepSurv	0.795	0.145	1.009	.717
			Kaplan–Meier Estimator	0.500	0.174	1.021	.693
			CoxPH	0.822	0.124	1.022	.887
			Random survival forest	0.812	0.135	1.122	.897
		Test	PC-Hazard	0.800	0.408	0.218	.003
			Cox-Time	0.561	0.169	1.072	.999
			DeepHit	0.448	0.339	0.338	.130
			Nnet-Survival	0.375	0.192	0.699	.516
			DeepSurv	0.814	0.144	1.007	.990
			Kaplan–Meier Estimator	0.500	0.171	0.997	.999
			CoxPH	0.870	0.113	1.030	.980
			Random survival forest	0.850	0.121	1.087	.988
CKD progression	Final variables	Training	PC-Hazard	0.504	0.260	0.595	.136
			Cox-Time	0.566	0.187	1.030	.821
			DeepHit	0.532	0.339	Inf	.053
			Nnet-Survival	0.522	0.341	0.270	.000
			DeepSurv	0.591	0.169	1.017	.640
			Kaplan–Meier Estimator	0.500	0.187	1.000	.976
			CoxPH	0.615	0.162	1.000	.666
			Random survival forest	0.619	0.177	1.072	.984
		Test	PC-Hazard	0.615	0.315	0.382	.000
			Cox-Time	0.576	0.199	1.015	.940
			DeepHit	0.480	0.361	0.654	.372
			Nnet-Survival	0.481	0.310	0.418	.009
			DeepSurv	0.658	0.181	1.056	.928
			Kaplan–Meier Estimator	0.500	0.197	1.033	.941
			CoxPH	0.666	0.177	1.018	.757
			Random survival forest	0.666	0.189	1.146	.669

C-statistics in the training and test set for different prediction models.

herbal medications consumption. As our models use few variables and routinely assessable laboratory data, which are particularly suited for resource-limited settings, the models can be generalizable to Asian regions where resources are limited in terms of diagnosis and treatment of CKD. Most currently available models were developed in the predominantly white populations in high-income countries in North America and Europe, whereby well-conducted clinical trials or well-established registries were placed [5, 6]. Additionally, the final variables in our models have a strong theoretical foundation from existing studies which means they are the common predictors for CKD disease and progression, increasing their generalizability. However, caution should be practiced while using the models in nations other than Asia due to the different demographic makeup, diagnosis and treatment approaches for CKD. Third, we relied on data from electronic medical records. Data on urine albumin–creatinine ratio, urine creatinine and urine albumin were limited as they were not routinely collected. Future studies that include these variables are warranted. Fourth, a fair and systematic comparison of the machine learning models is difficult due to small sample size, methodological differences used to develop prediction models and the lack of external validation. We did not include death as competing risk in the models due to data unavailability and limitation in the R package mlr3proba. Finally, the impact of the model on patient care is unclear. As such, future studies could assess whether the use of this model reduces the total number of patients developing CKD or delay worsening of CKD. Indeed, in view of the increasing prevalence of diabetes and CKD in the country, this study could pave way for future studies with a larger sample size and a more representative sample. Further prospective study can examine the possibility of implementing this CKD risk calculator in primary care settings with better external validation of the model.

The Cox regression model outperformed current machine learning techniques in predicting risk of incident CKD and CKD progression for a Malaysian cohort. As our model uses routinely available laboratory data, this simple yet practical method can be used to estimate the risk of incident CKD or CKD progression. This affords clinicians the tools to screen people in routine clinical setting, and helps people to recognize their potential risk profile, and prevent or delay the future incidence of CKD. However, external validation in multiple diverse cohorts of people with diabetes in clinical trials are needed.

SUPPLEMENTARY DATA

Supplementary data are available at *ckj* online.

ACKNOWLEDGEMENTS

We would like to thank the Director General of Health Malaysia for his permission to publish this article.

CONFLICT OF INTEREST STATEMENT

None declared.

AUTHORS' CONTRIBUTIONS

S.W.H.L. conceptualized the idea. R.S. and S.W.H.L. planned and carried out data curation and formal analysis. The project was supervised by S.W.H.L. and C.W.C. R.S. and S.W.H.L. wrote the first draft and all authors (R.S., C.W.C., N.K.L., N.L.A., Z.H., S.W.H.L.) provided critical input for important intellectual content.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

DATA AVAILABILITY STATEMENT

Data are available upon reasonable request to the authors.

REFERENCES

1. Magliano DJ, Boyko EJ. IDF Diabetes Atlas 10th edition scientific committee. *IDF DIABETES ATLAS* [Internet]. 10th edition. Brussels: International Diabetes Federation, 2021. https://www.ncbi.nlm.nih.gov/books/NBK581934/ (10 October 2022, date last accessed).
2. Saeedi P, Petersohn I, Salpea P *et al.* Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res Clin Pract* 2019;**157**:107843. https://doi.org/10.1016/j.diabres.2019.107843.
3. National Diabetes Registry Report 2013-2019. *Ministry of Health Malaysia*. 2020. https://www.moh.gov.my/moh/resources/Penerbitan/Rujukan/NCD/Diabetes/National_Diabetes_Registry_Report_2013-2019_26082021.pdf (10 October 2022, date last accessed).
4. National Diabetes Registry Report 2020. Ministry of Health Malaysia. 2021. https://www.moh.gov.my/moh/resources/Penerbitan/Rujukan/NCD/Diabetes/National_Diabetes_Registry_Report_2020.pdf (10 October 2022, date last accessed).
5. 26th Report of the Malaysian Dialysis & Transplant Registry 2018. Malaysian Society of Nephrology, National Kidney Foundation. 2018.
6. Ismail H, Abdul Manaf MR, Abdul Gafor AH *et al.* Economic burden of ESRD to the Malaysian Health Care System. *Kidney Int Rep* 2019;**4**:1261–70. https://doi.org/10.1016/j.ekir.2019.05.016.
7. Saminathan TA, Hooi LS, Mohd Yusoff MF *et al.* Prevalence of chronic kidney disease and its associated factors in Malaysia; findings from a nationwide population-based cross-sectional study. *BMC Nephrol* 2020;**21**:344. https://doi.org/10.1186/s12882-020-01966-8.
8. Tangri N, Kitsios GD, Inker LA *et al.* Risk prediction models for patients with chronic kidney disease: a systematic review. *Ann Intern Med* 2013;**158**:596–603. https://doi.org/10.7326/0003-4819-158-8-201304160-00004.
9. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS Med* 2012;**9**:e1001344. https://doi.org/10.1371/journal.pmed.1001344.
10. Slieker RC, van der Heijden AAWA, Siddiqui MK *et al.* Performance of prediction models for nephropathy in people with type 2 diabetes: systematic review and external validation study. *BMJ* 2021;**374**:n2134. https://doi.org/10.1136/bmj.n2134.
11. Alicic RZ, Rooney MT, Tuttle KR. Diabetic kidney disease. Challenges, progress, and possibilities. *Clin J Am Soc Nephrol* 2017;**12**:2032–45. https://doi.org/10.2215/cjn.11491116.
12. Hussain S, Chand Jamali M, Habib A *et al.* Diabetic kidney disease: an overview of prevalence, risk factors, and biomarkers. *Clin Epidemiol Global Health* 2021;**9**:2–6. https://doi.org/10.1016/j.cegh.2020.05.016.
13. Low S, Lim SC, Zhang X *et al.* Development and validation of a predictive model for chronic kidney disease progression in type 2 diabetes mellitus based on a 13-year study in Singapore. *Diabetes Res Clin Pract* 2017;**123**:49–54. https://doi.org/10.1016/j.diabres.2016.11.008.
14. Tuntayothin W, Kerr SJ, Boonyakrai C *et al.* Development and validation of a chronic kidney disease prediction model for type 2 diabetes mellitus in Thailand. *Value Health Reg Issues* 2021;**24**:157–66. https://doi.org/10.1016/j.vhri.2020.10.006.
15. Fregoso-Aparicio L, Noguez J, Montesinos L *et al.* Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetol Metab Syndr* 2021;**13**:148. https://doi.org/10.1186/s13098-021-00767-9.
16. Malaysia Ministry of Health. *Clinical Practice Guidelines Management of Type 2 Diabetes Mellitus*, 6th edition. 2020. https://mems.my/wp-content/uploads/2021/01/CPG_T2DM_6thEdition_2020.pdf (10 October 2022, date last accessed).
17. Nelson RG, Grams ME, Ballew SH *et al.* Development of risk prediction equations for incident chronic kidney disease. *JAMA* 2019;**322**:2104–14. https://doi.org/10.1001/jama.2019.17379.
18. De Cosmo S, Viazzi F, Pacilli A *et al.* Predictors of chronic kidney disease in type 2 diabetes: a longitudinal study from the AMD Annals initiative. *Medicine (Baltimore)* 2016;**95**:e4007. https://doi.org/10.1097/MD.0000000000004007.
19. Kidney Disease. Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney inter Suppl* 2013;**3**:1–150.
20. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994;**81**:515–26. https://doi.org/10.1093/biomet/81.3.515.
21. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of Martingale-based residuals. *Biometrika* 1993;**80**:557–72. https://doi.org/10.2307/2337177.
22. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017.
23. Sonabend R, Király FJ, Bender A *et al.* mlr3proba: an R package for machine learning in survival analysis. *Bioinformatics* 2021;**37**:2789–91. https://doi.org/10.1093/bioinformatics/btab039.
24. Sonabend R. survivalmodels: models for survival analysis 2022. Available from: https://cran.r-project.org/web/packages/survivalmodels/index.html.

25. D'Agostino R, Nam B. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of Statistics* 2003; **23**:1–25.
26. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. *Stat Med* 2015;**34**:1659–80. https://doi.org/10.1002/sim.6428.
27. Afghahi H, Cederholm J, Eliasson B *et al*. Risk factors for the development of albuminuria and renal impairment in type 2 diabetes—the Swedish National Diabetes Register (NDR). *Nephrol Dial Transplant* 2010;**26**:1236–43. https://doi.org/10.1093/ndt/gfq535.
28. Dagliati A, Marini S, Sacchi L *et al*. Machine learning methods to predict diabetes complications. *J Diabetes Sci Technol* 2018;**12**:295–302. https://doi.org/10.1177/1932296817706375.
29. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Statist Assoc* 1958;**53**:457–81. https://doi.org/10.2307/2281868.
30. Cox DR. Regression models and life-tables. *J R Stat Soc B* 1972;**34**:187–220.
31. Breiman L. Random forests. *Machine Learning* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324.
32. Ishwaran H, Kogalur UB, Blackstone EH *et al*. Random survival forests. *Ann Appl Stat* 2008;**2**:841–60, 20.
33. Katzman JL, Shaham U, Cloninger A *et al*. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Method* 2018;**18**:24. https://doi.org/10.1186/s12874-018-0482-1.
34. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. *PeerJ* 2019;**7**:e6257. https://doi.org/10.7717/peerj.6257.
35. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *arXiv* 2019. https://doi.org/10.48550/ARXIV.1907.00825.
36. Lee C., Zame WR, Yoon J *et al*., editors. DeepHit: A deep learning approach to survival analysis with competing risks, *Proceedings of the AAAI Conference on Artificial Intelligence* 2018;**32**:1. https://doi.org/10.1609/aaai.v32i1.11842 (10 October 2022, date last accessed).
37. Kvamme H, Borgan Ø. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal* 2021;**27**:710–36. https://doi.org/10.1007/s10985-021-09532-6.
38. Harrell FE, Jr, Califf RM, Pryor DB *et al*. Evaluating the yield of medical tests. *JAMA* 1982;**247**:2543–6.
39. Graf E, Schmoor C, Sauerbrei W *et al*. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;**18**:2529–45. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18⟨2529::aid-sim274⟩3.0.co;2-5.
40. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* 2000;**19**:3401–15. https://doi.org/10.1002/1097-0258(20001230)19:24⟨3401::aid-sim554⟩3.0.co;2-2.
41. Haider H, Hoehn B, Davis S *et al*. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*. 2020. https://www.jmlr.org/papers/volume21/18-772/18-772.pdf (10 October 2022, date last accessed).
42. Jardine MJ, Hata J, Woodward M *et al*. Prediction of kidney-related outcomes in patients with type 2 diabetes. *Am J Kidney Dis* 2012;**60**:770–8. https://doi.org/10.1053/j.ajkd.2012.04.025.
43. Moncada-Torres A, van Maaren MC, Hendriks MP *et al*. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci Rep* 2021;**11**:6968. https://doi.org/10.1038/s41598-021-86327-7.
44. Adeoye J, Koohi-Moghadam M, Lo AWI *et al*. Deep learning predicts the malignant-transformation-free survival of oral potentially malignant disorders. *Cancers (Basel)* 2021;**13**:6054. https://doi.org/10.3390/cancers13236054.
45. Moradmand H, Aghamiri SMR, Ghaderi R *et al*. The role of deep learning-based survival model in improving survival prediction of patients with glioblastoma. *Cancer Med* 2021;**10**:7048–59. https://doi.org/10.1002/cam4.4230.
46. Du M, Haag DG, Lynch JW *et al*. Comparison of the tree-based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: analyses based on SEER database. *Cancers (Basel)* 2020;**12**:2802. https://doi.org/10.3390/cancers12102802.
47. Bikbov B, Purcell CA, Levey AS *et al*. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet North Am Ed* 2020;**395**:709–33. https://doi.org/10.1016/S0140-6736(20)30045-3.