Original Article

Kidney Res Clin Pract 2025;44(5):802-813
pISSN: 2211-9132 • eISSN: 2211-9140
https://doi.org/10.23876/j.krcp.23.212

KIDNEY RESEARCH AND CLINICAL PRACTICE

Conventional machine learning-based prediction models did not outperform the International IgA Nephropathy Prediction Tool

Sehoon Park[1,*], Yisak Kim[2,3,*], Chung Hee Baek[4], Hyunjeong Cho[5], Ji In Park[6], Eun Sil Koh[7], Jung Pyo Lee[8], Sun-Hee Park[9], Hyung Woo Kim[10], Seung Hyeok Han[10], Ho Jun Chin[11,12], Dong Ki Kim[1,12], Kyung Chul Moon[13], Young-Gon Kim[14,15,†], Hajeong Lee[1,12,†]

*For further information on the authors' affiliations, see **Additional information**.*

Background: Immunoglobulin A nephropathy (IgAN) is a major cause of end-stage kidney disease (ESKD). The International IgA Nephropathy Prediction Tool (IIgAN-PT) predicts IgAN prognosis, but improvement in the prediction performance using machine learning (ML)-based methods is needed.

Methods: We analyzed 4,425 biopsy-confirmed patients with IgAN and ≥6 months of follow-up from nine tertiary university hospitals in Korea. The study population was divided into development and validation cohorts. Using the collected 87 clinicodemographic and pathological variables, ML-based prediction models for ESKD or estimated glomerular filtration rate decline (50% reduction or <15 mL/min/1.73 m^2) were constructed: 1) the conventional CatBoost model, 2) the optimized CatBoost model with Cox proportional hazards, 3) the deep Cox proportional hazards model, and 4) the deep Cox mixture model. The area under the curve (AUC) and calibration plots were used to investigate the discriminative and calibration performance of the models, which were then compared with those of the IIgAN-PT full model.

Results: The full model showed excellent performance (AUC [95% confidence interval] for 5-year outcome, 0.896 [0.853–0.940]), with acceptable calibration results. The ML-based models showed good performance in predicting adverse kidney outcomes and revealed acceptable discrimination performance in the external validation (AUC [95% confidence interval] for the 5-year outcome: 1) 0.829 [0.791–0.866]; 2) 0.847 [0.804–0.890]; 3) 0.823 [0.784–0.862]; and 4) 0.832 [0.794–0.870]), although the models showed underestimation in calibration analysis of the external validation cohort. With the validation data, the overall performance of the IIgAN-PT was non-inferior to that of the ML-based model.

Conclusions: Our ML-based models showed good performance in predicting adverse kidney outcomes in patients with IgAN but they did not outperform the IIgAN-PT.

Keywords: Disease progression, IGA glomerulonephritis, Machine learning, Prognosis

Received: August 10, 2023; **Revised:** March 2, 2024; **Accepted:** March 21, 2024

Correspondence: Young-Gon Kim
Department of Transdisciplinary Medicine, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea.
E-mail: younggon2.kim@gmail.com
ORCID: https://orcid.org/0000-0003-2148-1299

Hajeong Lee
Department of Internal Medicine, Seoul National University Hospital, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea. E-mail: mdhjlee9@snu.ac.kr
ORCID: https://orcid.org/0000-0002-1873-1587

*Sehoon Park and Yisak Kim contributed equally to this study as co-first authors.
[†]Young-Gon Kim and Hajeong Lee contributed equally to this study as co-corresponding authors.

Introduction

Immunoglobulin A nephropathy (IgAN) is the most prevalent primary glomerulonephritis worldwide [1]. The clinical presentation and overall prognosis of IgAN are extremely heterogeneous. IgAN may be worsened by high blood pressure, significant proteinuria, the presence of kidney dysfunction, or unfavorable pathologic characteristics. Approximately one-third of patients with IgAN progress to end-stage kidney disease (ESKD) in their middle age, ranking the disease as one of the important causes of socioeconomic burden related to kidney failure, especially in Asian countries [2–4]. However, a certain portion of patients with IgAN exhibit a benign course without notable deterioration of kidney function. Therefore, the current KDIGO (Kidney Disease: Improving Global Outcomes) guideline for glomerular diseases recommends stratifying the kidney progression risk of patients with IgAN based on clinical and histologic data and quantifying progression risk at diagnosis using the International IgA Nephropathy Prediction Tool (IIgAN-PT) [5,6]. The IIgAN-PT is the prediction model that includes the largest number of patients with IgAN from various regions of the world [5]. The prognostic performance of the IIgAN-PT has been also validated in certain external cohorts, including children, supporting the validity of the model [7–10].

Artificial intelligence (AI) provides an emerging opportunity to develop automatic clinical/pathological image annotations, construct clinical decision support systems, and build robust prediction models. However, the ability of AI to handle complex high-dimensional data without being affected by characteristics of parameters or statistical assumptions remains to be determined [11]. Machine learning (ML)-based methods, a subfield of AI that teaches machines to learn from past data without explicit programming, have also been trialed for the prognostic IgAN model [12–14], yet, a widely validated deep learning (DL)-based model has not been established. Additional studies implementing the AI approach to integrate the complex clinicopathological information of patients with IgAN may improve the performance of prognostic strategies for the disease.

This study aimed to develop ML-based models to predict the prognosis of IgAN. We trained and validated ML- and DL-based models using a comprehensive collection of 87 demographic, clinical, and pathologic variables from a large-scale multicenter cohort in South Korea. We also validated the full IIgAN-PT model in the Korean population and compared the performance of the AI models with that of the IIgAN-PT model derived from the conventional Cox proportional hazards model.

Methods

Ethics considerations

The study was approved by the Institutional Review Board of Seoul National University Hospital/Seoul National University Bundang Hospital/SMG-SNU Boramae Medical Center (No. H-2103-091-1205), Severance Hospital (No. 4-2021-0376), The Catholic University of Korea, Yeouido St. Mary's Hospital (No. SC21RIDI0090), Asan Medical Center (No. 2021-1333), Kyungpook National University Hospital (2021-04-036), Chungbuk National University Hospital (No. 2021-09-004), and Gangwon National University Hospital (No. KNUH-A-2021-08-012-001). Data on all study participants were collected from each hospital and sent to the central analysis laboratory after anonymization using the standard protocol. The requirement for informed consent was waived because this was a retrospective observational study without medical intervention. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Study setting

This multicenter study included biopsy-confirmed IgAN cases from nine tertiary hospitals throughout Korea. We first collected the diverse demographic, clinical, and pathological characteristics of patients with IgAN by reviewing their electronic health records. Next, we implemented a multiple ML-based approach to construct a prediction model for kidney disease progression in IgAN. Finally, we compared the discriminative and calibration performances of the models with those of IIgAN-PT.

Study population

We included all available biopsy-confirmed native IgAN cases from the electronic medical records of the study hospitals (Fig. 1). Patients who progressed to the adverse kidney outcome within 6 months were excluded because such acute aggravation is not the target of the current study. The development cohort for the ML-based model included patients with IgAN from Seoul National University Hospital, Seoul National University Bundang Hospital, and SMG-SNU Boramae Medical Center. The three hospitals are all affiliated with the Seoul National University College of Medicine and may share a distinct medical environment; thus, combining the data from other hospitals as the validation cohort strengthened the external validation cohort.

Study outcome

The study outcome included a decrease in estimated glomerular filtration rate (eGFR) of less than half of the baseline or ESKD, defined as kidney replacement therapy or eGFR of <15 mL/min/1.73 m^2. The study population was censored at the time of outcome or loss to follow-up.

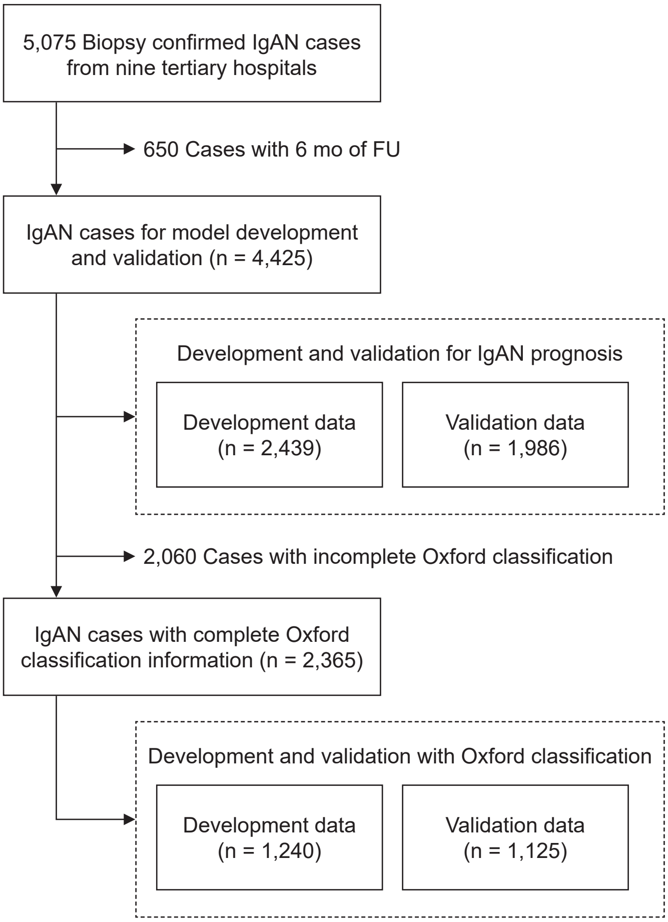
5,075 Biopsy confirmed IgAN cases from nine tertiary hospitals
650 Cases with 6 mo of FU
IgAN cases for model development and validation (n = 4,425)
Development and validation for IgAN prognosis
Development data (n = 2,439)
Validation data (n = 1,986)
2,060 Cases with incomplete Oxford classification
IgAN cases with complete Oxford classification information (n = 2,365)
Development and validation with Oxford classification
Development data (n = 1,240)
Validation data (n = 1,125)

Figure 1. Study flow diagram.
FU, follow-up; IgAN, immunoglobulin A nephropathy.

Data collection for model variables

A total of 87 demographic, clinical, and pathological variables were collected and included in the model. For instance, we reviewed all variables included in the full IIgAN-PT model; these were social habits (e.g., smoking), various laboratory test results (e.g., serum electrolyte levels, serum protein/albumin levels, and complete blood counts including white blood cells, hemoglobin, and platelets), anthropometric measures (e.g., body mass index), pathological features, including light microscopy findings (e.g., global sclerosis, segmental sclerosis, or cellular or fibrocellular crescent) and electron or immunofluorescence microscopy. Because our longitudinal cohort covered a long period, some patients with IgAN were diagnosed before their institution adopted the Oxford classification for the pathologic diagnosis of IgAN. The pathologic parameters were assessed by each pathologist in the study hospitals and we retrospectively collected the pathology reports. Supplementary Table 1 (available online) provides a complete list of the collected variables.

Machine learning-based model construction

We used two ML-based and two DL-based models to construct a prognostic prediction model for IgAN. For the ML-based model, the conventional CatBoost [15] and optimized CatBoost with the Cox proportional hazards were trained using the collected data. CatBoost is a gradient-boosted decision tree model [16] with ordered target statistics and boosting and is a powerful tool for classification and regression. As a decision tree-based algorithm, it is well-suited to ML tasks involving categorical, heterogeneous data and can also compute feature importance [17]. CatBoost with the Cox proportional hazards is a model with a modified loss function for survival regression. Unlike common supervised tasks in which the target variable is known and observed during the entire period in the training dataset, survival regression can handle partially observed or censored target variables. Therefore, unlike the CatBoost method, which requires the development of a separate model for each time section to handle censored data, CatBoost with the Cox proportional hazards can handle various time sections using a single model that optimizes the log partial likelihood derived from the hazard function for Cox proportional hazards.

For the DL-based model, deep logistic hazards [18] and deep Cox mixture [19] were used for survival regression. Deep logistic hazard is a discrete-time survival prediction method with neural networks that parameterizes discrete hazards and optimizes the survival likelihood. We use a multilayer perceptron with two hidden layers to implement deep logistic hazards. The deep Cox mixture is another survival prediction method that generalizes the proportional hazards assumption via a mixture model by assuming that there are latent groups and that within each group, the proportional hazards assumption holds. This method is not restricted by the strong assumption of proportional hazards, which allows the model to choose these latent groups and build a more expressive survival prediction model.

The variables that contributed to the prognostic ability of the models were weighted by feature importance analysis in the ML-based models, including the CatBoost model for 5-year adverse kidney outcomes and the CatBoost model with Cox proportional hazards.

During the model production, we preprocessed the original data, including missing value filling, data standardization, and data normalization. Both CatBoost and the CatBoost with the Cox proportional hazards were implemented using PyCaret [20] and the official CatBoost Python package. Deep logistic hazards and deep Cox mixtures were implemented using the pycox python [18] package and the official deep Cox mixture repository. Missing values were masked (categorical) or averaged (numerical) according to the data type to maintain simplified method for further application in external datasets. The training/validation ratio for the DL methods was 9:1, and performance stability was assessed by bootstrapping. The assessment of Cox assumption of the Cox-based DL models used visualization of the Kaplan-Meier survival curves and checked whether the survival curves crossed in follow-up duration which may indicate violation of the assumption (Supplementary Fig. 1, available online).

Statistical analysis

For validation, the prediction scores extracted from the ML-based models were used to inspect the discriminative and calibration performance of the validation set. For the CatBoost model with the Cox proportional hazards function, the predictor for survival analysis was available as the IIgAN-PT, allowing calculation of the c-index. All four models provided prediction scores at specific time points of the outcomes, and we extracted the prediction scores to calculate the receiver-operating characteristic area under the curve (ROC-AUC) values at the 1-, 3-, 5-, and 10-year points to assess discriminative power. Calibration was performed using a calibration plot to assess the true and expected risks for 5-year adverse outcomes. The ROC-AUC values were directly compared to those of the IIgAN-PT calculated by the full model, within those with complete information on the variables required to apply the IIgAN-PT (e.g., Oxford classification) using the Delong test. As a sensitivity analysis, we additionally constructed ML-based models within those with complete information for the IIgAN-PT application and again compared the results in the validation set with the available data. Statistical significance was set at $p < 0.05$ significance. Clinical statistical analysis was performed using R software (version 3.6.2; R Foundation for Statistical Computing). Censoring of the data was considered to occur in a random manner.

Results

Baseline characteristics

A total of 5,075 biopsy-confirmed IgAN cases were screened in this study. Supplementary Table 2 (available online) summarizes the characteristics of the cohort. The overall characteristics differed between the study hospitals, and the median age of patients ranged from 32 to 44 years. Approximately 5% and 30%–40% of the study participants had diabetes mellitus and hypertension, respectively. The treatment history of immunosuppressive drugs at the time of biopsy was mostly less than 10%, while the proportion of those treated with renin-angiotensin-aldosterone blockade ranged from 24% to 58%.

After excluding patients with IgAN and a follow-up of less than 6 months, we constructed development and validation datasets comprising 2,439 and 1,986 patients with IgAN, respectively; Table 1 summarized the characteristics of these patients. The median follow-up duration was 5.8 years (interquartile range [IQR], 2.6–10.1 years) with a median biopsy date of March 2011 (IQR, April 2004–March 2016). The median follow-up duration in the development cohort and that in the validation cohort was 3.8 years (IQR, 1.5–7.3 years) with a median biopsy date of August 2015 (IQR, January 2011–June 2018). Among them, 1,240 and 1,125 patients with IgAN had complete information on the Oxford classifications, respectively; thus, they were included in the additional analysis with model development within the full Oxford classification information (Supplementary Table 3, available online).

Table 1. Baseline characteristics of the discovery and validation cohorts

Characteristic	Development cohort (n = 2,439)	Validation cohort (n = 1,986)
Clinical characteristics		
Age (yr)	36.0 (22.0–49.0)	39.0 (29.0–49.0)
Sex		
Female	1,207 (49.5)	1,068 (53.8)
Male	1,232 (50.5)	918 (46.2)
Diabetes mellitus	118 (4.9)	68 (3.5)
Hypertension	958 (39.4)	584 (29.9)
Systolic BP (mmHg)	120.0 (110.0–130.0)	121.0 (111.0–134.0)
Diastolic BP (mmHg)	73.0 (67.0–80.0)	80.0 (70.0–86.0)
eGFR (mL/min/1.73 m^2)	93.7 (62.6–121.0)	89.7 (64.2–109.7)
Proteinuria (g/g or g/24 hr)	1.1 (0.5–2.1)	0.8 (0.4–1.7)
ISD	150 (6.4)	125 (6.3)
RAASB	1,111 (47.5)	912 (46.0)
Pathologic characteristics		
M1	829 (64.5)	278 (21.8)
E1	212 (16.5)	357 (28.0)
S1	844 (65.7)	859 (67.2)
T1	334 (26.0)	192 (15.0)
T2	35 (2.7)	67 (5.2)
C1	425 (17.8)	226 (21.2)
C2	22 (0.9)	8 (0.8)
Adverse kidney outcome (during total follow-up duration)	515 (21.1)	308 (15.5)

Data are expressed as median (interquartile range) or number (%). The total numbers of composite, 50% decline of eGFR, end-stage kidney disease outcome events were 99/4,199, 60/4,201, 37/4,215 in 1-year, 286/3,303, 241/3,307, 153/3,325 in 3-year, 438/2,679, 394/2,684, 236/2,692, in 5-year, and 689/1,602, 659/1,607, 369/1,594 in 10-year follow-up period, with censoring the cases that not reach the follow-up endpoint.

BP, blood pressure; eGFR, estimated glomerular filtration rate; ISD, immunosuppressive drug; RAASB, renin-angiotensin-aldosterone blockades.

Performance of the IIgAN-PT

We first applied the IIgAN-PT full model to the collected dataset, which contained the complete Oxford classification information (n = 2,178). In study subjects with complete information for IIgAN-PT, IIgAN-PT showed acceptable performance, with AUC values of 0.836 (95% CI, 0.752–0.920), 0.873 (95% CI, 0.840–0.906), 0.857 (95% CI, 0.828–0.885), and 0.799 (95% CI, 0.757–0.840) for 1-, 3-, 5-, and 10-year outcomes, respectively. The overall calibration was acceptable when inspected using a calibration plot (Supplementary Fig. 2, available online).

Performance of the machine learning-based models

We then developed an ML-based model for 2,439 patients with or without Oxford classification information, and its performance was tested in the validation set (n = 1,717). In the validation set, the conventional CatBoost, optimized CatBoost with the Cox proportional hazards, deep logistic hazard, and deep Cox mixture models provided AUC values mostly ranging from 0.7 to 0.8 (Table 2, Fig. 2). The result of a single model did not show prominent superiority over the others, although the conventional CatBoost model showed low discriminative power (AUC, 0.512) toward the 10-year outcome data. When assessing the calibration of the developed models, the four models showed generally acceptable calibration results, as no significant deviation was identified in the calibration plots. However, a slight underestimation of the risk of adverse kidney outcomes was identified in the models developed using these four methods. When the composite outcome was divided into ESKD or eGFR 50% reduction, the performance was better towards ESKD outcome than the eGFR 50% reduction (Supplementary Table 4, available online).

Feature importance

We inspected the feature importance, which refers to the

Table 2. Discriminative performances of the artificial intelligence-based models

Model	Time point of outcome (yr)	AUC (95% CI)
CatBoost	1	0.741 (0.628–0.854)
	3	0.848 (0.803–0.893)
	5	0.829 (0.791–0.866)
	10	0.561 (0.512–0.610)
CatBoost with the Cox proportional hazards	1	0.775 (0.661–0.889)
	3	0.847 (0.804–0.890)
	5	0.850 (0.817–0.883)
	10	0.810 (0.772–0.848)
Deep Cox proportional hazards	1	0.779 (0.666–0.893)
	3	0.826 (0.775–0.876)
	5	0.823 (0.784–0.862)
	10	0.806 (0.768–0.844)
Deep Cox mixture	1	0.779 (0.667–0.891)
	3	0.832 (0.783–0.882)
	5	0.832 (0.794–0.870)
	10	0.825 (0.789–0.861)

AUC, area under the curve; CI, confidence interval.

variables that the constructed models mostly referred to for their prediction (Fig. 3) in the models constructed using ML-based methods. In the CatBoost model for 5-year outcomes and the CatBoost model with Cox proportional hazards, when we tested the results in the discovery cohort and the cohort with complete information for the IIgAN-PT, the notable variables included serum creatinine, global sclerosis (%), eGFR, and proteinuria levels as the variables ranked among the top five variables. The number of glomeruli in the entire biopsy specimen, serum uric acid, blood urea nitrogen, serum albumin, and segmental sclerosis (%) were the variables that appeared in the top 20 variables in all four models.

Performance comparison between the IIgAN-PT and machine learning-based models

We compared the model performance within the validation dataset (n = 1,125) with the complete information for

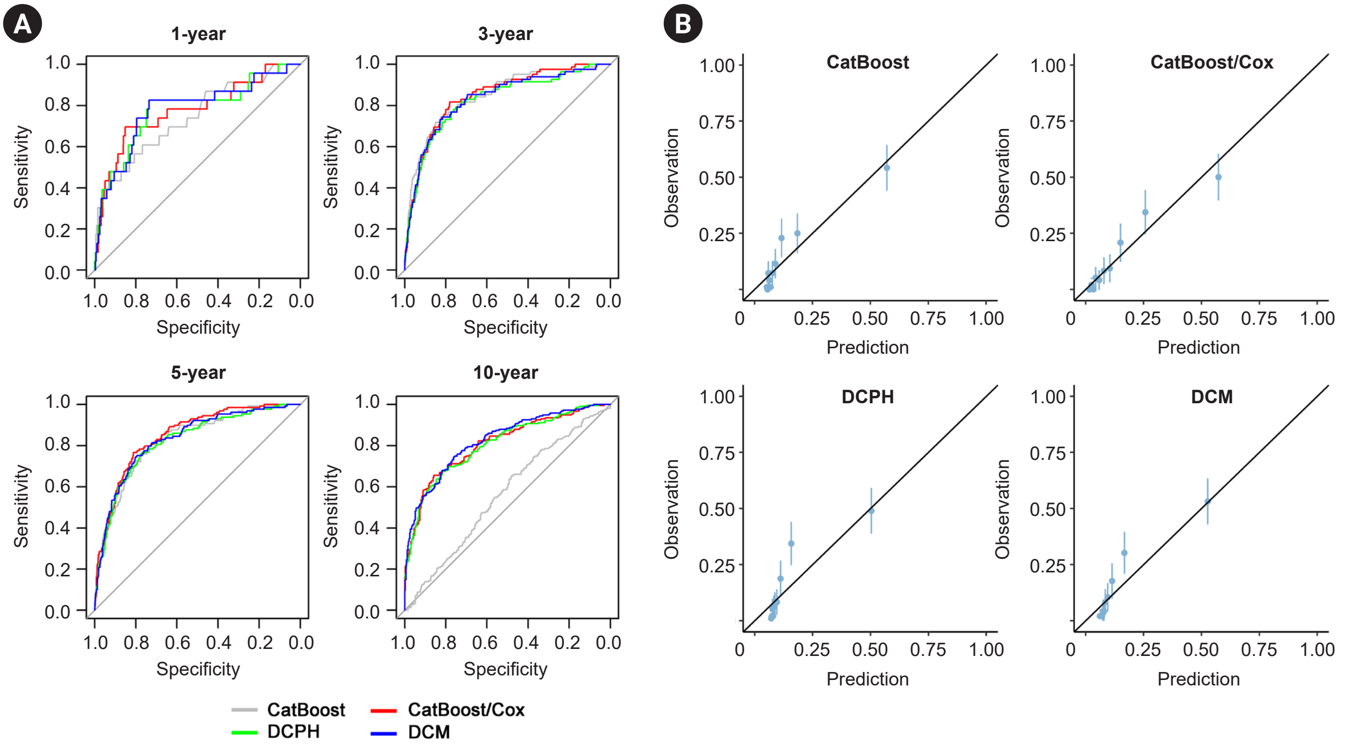

Figure 2. Artificial intelligence-based model performances. (A) The graphs show the discriminative performance and area under the curve value of receiver-operating characteristics curve towards the adverse kidney outcomes censored at four time points. The grey lines indicate the results by the CatBoost model, the red by CatBoost with the Cox proportional hazards, the green by deep Cox proportional hazards (DCPH), and the blue by deep Cox mixture (DCM), respectively. (B) The graphs are the calibration plots showing the predicted (x-axis) and observed (y-axis) risk for the 5-year adverse kidney outcomes by the models constructed by the according methods.

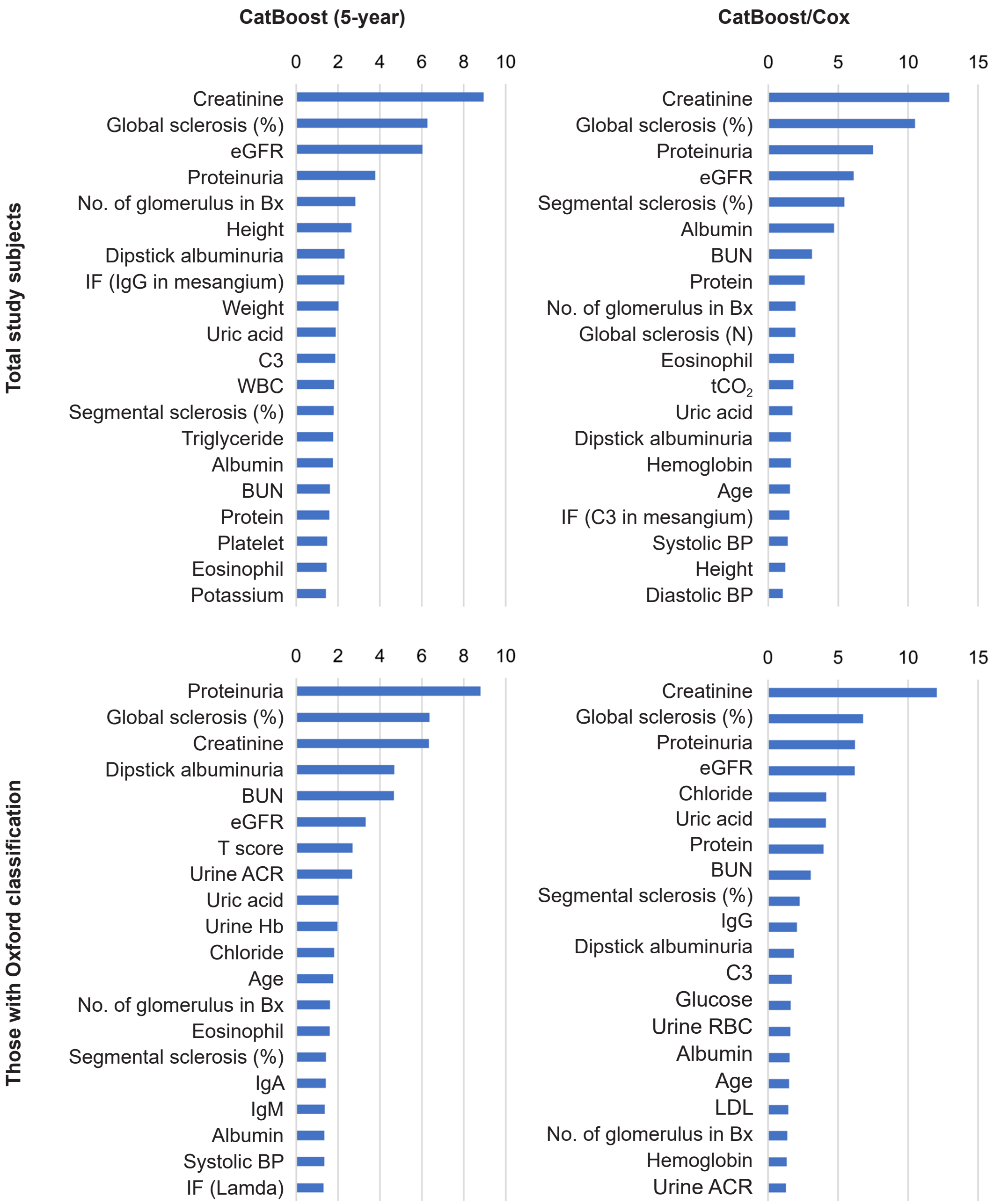

Figure 3. Feature importance of the constructed CatBoost models. The left graphs show the results by CatBoost model for 5-year adverse kidney outcomes and the right graphs show the results by CatBoost model with Cox proportional hazards. The upper two graphs show the results in total study samples with >6 months of follow-up, and the lower two graphs show the results within the population with complete information of the Oxford classification and thus included in the analysis for the comparison of performance with that of the IIgAN-PT. The relative importance value indicates the weighted importance in the prediction models contributing to the overall performance of the constructed model. Results from the top 20 variables are listed in the figure.

ACR, albumin-to-creatinine ratio; BP, blood pressure; BUN, blood urea nitrogen; Bx, biopsy; eGFR, estimated glomerular filtration rate; IF, immunofluorescence; IIgAN-PT, International IgA Nephropathy Prediction Tool; IgA, immunoglobulin A; IgG, immunoglobulin G; LDL, low-density lipoprotein; RBC, red blood cell; WBC, white blood cell.

the IIgAN-PT (Table 3, Fig. 4). The IIgAN-PT again showed acceptable discriminative performance within the validation cohort, as the AUC values ranged from 0.834 to 0.896 for adverse outcomes at 1, 3, 5, and 10 years. The performances were similar to those of the ML-based methods, although no modeling results were statistically superior to the discriminative performance of the IIgAN-PT. Similarly, the calibration results were acceptable for both the IIgAN-PT and ML-based models. However, some underestimation of the risks of adverse kidney outcomes was noted in all tested ML-driven models.

Discussion

In this study, we developed ML-driven prediction models for the prognosis of IgAN kidneys by incorporating various clinicopathological variables. The constructed models demonstrated good discrimination and calibration performance in the external validation. As a reference, the full IIgAN-PT model showed excellent performance in our large-scale cohort study. The overall performance of the IIgAN-PT was non-inferior to that of ML-based models, additionally supporting the clinical utility of the IIgAN-PT in patients with IgAN.

Accurate prediction of IgAN kidney prognosis is crucial for appropriate risk stratification, scheduling follow-up visits, determining treatment strategies, and counseling patients. The IIgAN-PT is the most widely validated prognostic model for IgAN, and the full model includes age, blood pressure, baseline eGFR, proteinuria amounts, treatment history by renin-angiotensin-aldosterone blockades or by immunosuppressive drugs, the Oxford classification and with or without ethnicity [5]. The IIgAN-PT has been well validated in various cohorts [7–9]. However, the Korean population was not included in the development of the data, and some underestimation of kidney risk was suspected in a previous report [8]. Herein, we demonstrated the IIgAN-PT also showed acceptable predictive performances in this multicenter Korean IgAN cohort.

There is a relevant question regarding whether AI can

Table 3. Comparisons between the discriminative performance of the IIgAN-PT and the artificial intelligence-based models

Model	Time point of outcome (yr)	AUC (95% CI)	p-value vs. IIgAN-PT
IIgAN-PT	1	0.834 (0.710–0.958)	NA
	3	0.885 (0.830–0.941)	NA
	5	0.896 (0.853–0.940)	NA
	10	0.850 (0.787–0.913)	NA
CatBoost	1	0.700 (0.543–0.858)	0.04
	3	0.857 (0.793–0.922)	0.23
	5	0.890 (0.851–0.930)	0.71
	10	0.662 (0.573–0.751)	<0.001
CatBoost with the Cox proportional hazards	1	0.848 (0.719–0.978)	0.59
	3	0.877 (0.811–0.944)	0.67
	5	0.854 (0.803–0.904)	0.02
	10	0.846 (0.783–0.908)	0.87
Deep Cox proportional hazards	1	0.841 (0.732–0.951)	0.70
	3	0.883 (0.817–0.949)	0.93
	5	0.860 (0.806–0.913)	0.06
	10	0.859 (0.799–0.919)	0.06
Deep Cox mixture	1	0.841 (0.728–0.955)	0.74
	3	0.876 (0.806–0.946)	0.72
	5	0.860 (0.805–0.914)	0.09
	10	0.890 (0.839–0.941)	0.07

The numbers of study subjects with the available outcome data until the designated follow-up period were 874 (13 outcomes), 577 (34 outcomes), 388 (63 outcomes), and 149 (90 outcomes) for the 1-, 3-, 5-, and 10-year outcomes, respectively, within the validation dataset with complete information for the IIgAN-PT.

AUC, area under the curve; CI, confidence interval; IIgAN-PT, International IgA Nephropathy Prediction Tool; NA, not applicable.

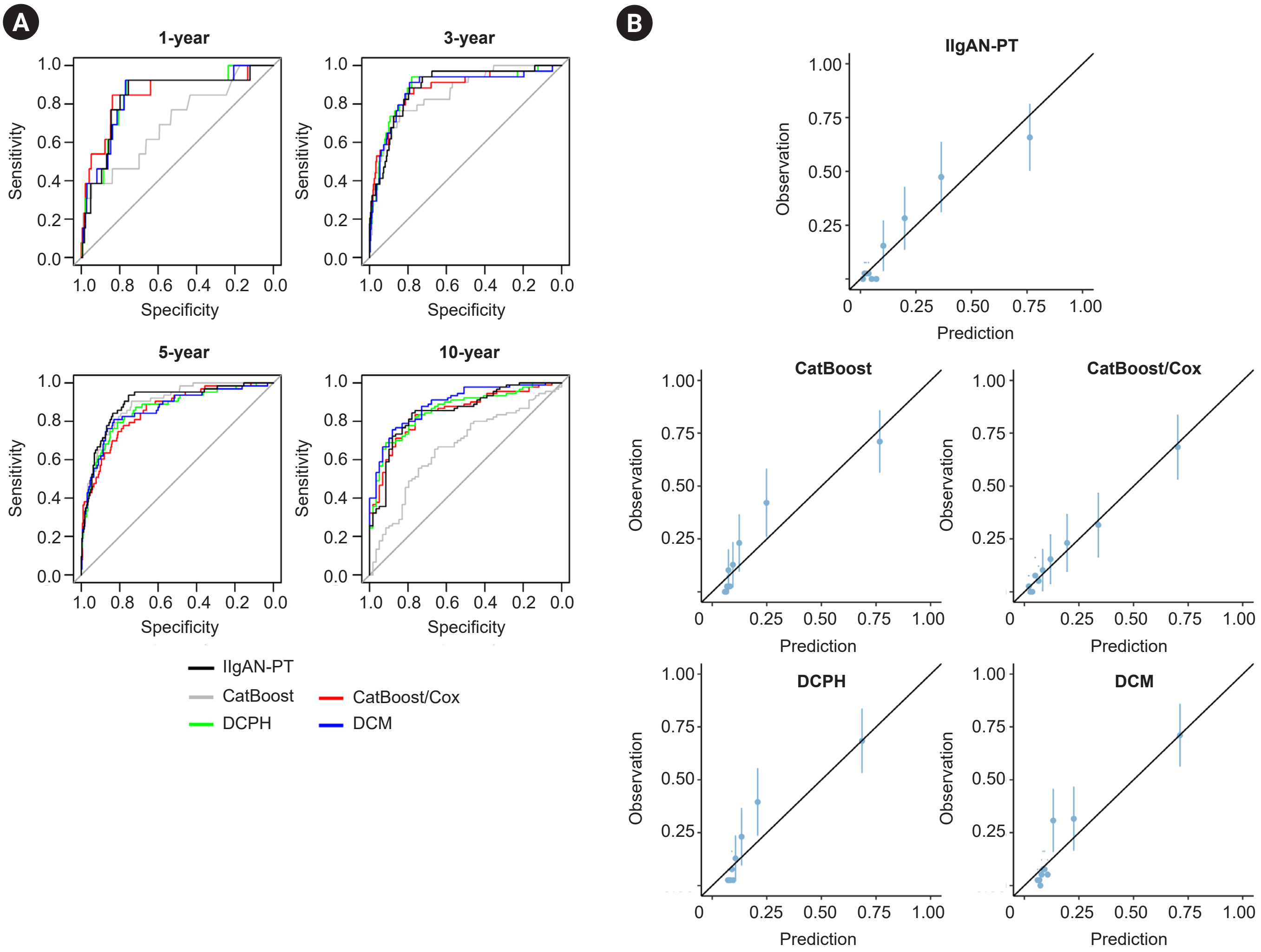

Figure 4. Comparison between the performance by the artificial intelligence-based models and that by the IIgAN-PT within the dataset with complete information for the Oxford classification. (A) The graphs show the discriminative performance and area under the curve value of receiver-operating characteristics curve towards the adverse kidney outcomes censored at four time points. The grey lines indicate the results by CatBoost model, the red by CatBoost with the Cox proportional hazards, the green by deep Cox proportional hazards (DCPH), the blue by deep Cox mixture (DCM), and the black by the IIgAN-PT, respectively. (B) The graphs are the calibration plots showing the predicted (x-axis) and observed (y-axis) risk for the 5-year adverse kidney outcomes by the models constructed by the according methods.

IIgAN-PT, International IgA Nephropathy Prediction Tool.

develop a more advanced prediction model for IgAN, as this approach has recently proliferated and opened a new field of clinical prediction modeling. The ML-based approach is now actively used in the clinical image reading systems [21,22] and has shown excellent performance in risk stratification, combining hundreds of complex clinical features [23,24]. As the prediction of IgAN kidney prognosis may be improved from additional medical information, the AI-based approach is a promising method for developing a model with better prognostic performance. A previous deep learning-based model showed a non-inferior predictive performance to that of the IIgAN-PT; however, a superior finding has not yet been reported [11]. In the current study, we developed multiple ML-based models, enhanced by deep learning-related approaches, including a wide range of variables of IgAN patients at the time of diagnosis. These models generally demonstrated acceptable performance for the prognosis of IgAN. However, the clinical utility of IIgAN-PT was well validated in our cohort, and its performance was non-inferior to that of the models despite trialing multiple ML- and DL-based methods. In addition, the results showing the validity of the IIgAN-PT for 10-year kidney outcomes support that the model can be useful in predicting the long-term prognosis of IgAN patients [9]. Considering the generalizability, interpretability, and accessibility that had been demonstrated in the IIgAN-PT model, our current AI models seem to be unable to beat the IIgAN-PT model without securing outperformance in predicting kidney prognosis of IgAN. Therefore, the current study supports the clinical utility of the IIgAN-PT, as the model is easy to use without collecting extensive medical information, unlike AI-based models.

The ML-based models failed to show superior performance compared with the IIgAN-PT, despite the inclusion of a wide range of medical information, which can be explained by several factors. First, the variables included in the full IIgAN-PT model are not mere predictors but have significant causal effects on kidney prognosis or directly reflect kidney health. Elevated blood pressure or high amounts of proteinuria are not only common in chronic kidney disease but also directly damage the kidney [25,26], and the baseline eGFR reflects the underlying kidney function impairment. The Oxford classification includes mesangial proliferation, subsequent glomerular alteration, or active inflammation such as crescent formation, and final tubulointerstitial pathology; thus, it reflects the overall pathophysiologic aspect of IgAN progression from the initial stages to late pathologic consequences [27]. Constructed from these very important clinicopathologic features, variables not included in the IIgAN-PT may have only a minor impact on the prognosis of IgAN; thus, combining the effects of the variables by ML-based methods may have only a small advantage. Next, IgAN cohorts are relatively small compared to big data, which are widely used when applying AI-based methods. Although some AI-based methods are targeted at constructing prediction models for middle-to small-sized data, the superiority of ML- or DL-based methods may be weakened in datasets with a few thousand samples. A larger dataset may be required to develop a superior prognostic model; however, collecting standardized medical information from multiple cohorts and countries is challenging.

This study had some limitations that should be addressed in future research. First, as noted above, the study sample size may not have been sufficient to develop a superior model for an ML-based approach, even though we included >3,000 patients with IgAN from multiple hospitals. A multinational consortium may collect a wide range of clinical information to develop an ML-based prediction model for IgAN using a larger sample size. Second, AI can deal with additional complex data, such as digital pathologic images, combined multiomics data, and time-sequenced information [11]. Rather than the current analysis using cross-sectional baseline information, additional studies may include datasets having multiple dimensions with complex features for which the AI-based approach has superiority. Third, this study included a population with a single ethnic background. Similar to the original multinational cohort used for IIgAN-PT development, the AI-based approach may also be trialed for those of various ethnicities. Lastly, some heterogeneity in collection of the study variables (e.g., pathology parameters) might have existed because we retrospectively collated the information from the study hospitals.

In conclusion, the IIgAN-PT performance was validated in the current large-scale IgAN cohort in Korea. Although ML-based prediction models may provide acceptable prediction performance for IgAN prognosis, a prediction model combining diverse baseline features may not be sufficient to develop an advanced model that is superior to IIgAN-PT. Future efforts, including large-scale and high-level data on IgAN, are warranted to improve the performance of the IgAN prognostic prediction models.

Additional information

[1]*Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea*
[2]*Interdisciplinary Program in Bioengineering, Seoul National University Graduate School, Seoul, Republic of Korea*
[3]*Integrated Major in Innovative Medical Science, Seoul National University Graduate School, Seoul, Republic of Korea*
[4]*Department of Internal Medicine, Asan Medical Center, Seoul, Republic of Korea*
[5]*Department of Internal Medicine, Chungbuk National University Hospital, Cheongju, Republic of Korea*
[6]*Department of Internal Medicine, Kangwon National University Hospital, Kangwon National University School of Medicine, Chuncheon, Republic of Korea*
[7]*Department of Internal Medicine, The Catholic University of Korea, Yeouido St. Mary's Hospital, Seoul, Republic of Korea*
[8]*Department of Internal Medicine, SMG-SNU Boramae Medical*

Center, Seoul, Republic of Korea
[9]*Department of Internal Medicine, School of Medicine, Kyungpook National University, Kyungpook National University Hospital, Daegu, Republic of Korea*
[10]*Department of Internal Medicine, Severance Hospital, Seoul, Republic of Korea*
[11]*Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea*
[12]*Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea*
[13]*Department of Pathology, Seoul National University Hospital, Seoul, Republic of Korea*
[14]*Department of Transdisciplinary Medicine and Innovative Medical Technology Research Institute, Seoul National University Hospital, Seoul, Republic of Korea*
[15]*Department of Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea*

Conflicts of interest

All authors have no conflicts of interest to declare.

Funding

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00219548) and cooperative research funding from the Korean Nephrology Research Foundation 2020.

Data sharing statement

The data presented in this study are available from the corresponding author upon reasonable request.

Authors' contributions

Conceptualization, Formal analysis, Methodology: SP, YK, KCM, YGK, HL
Data curation: CHB, HC, JIP, ESK, JPL, SHP, HWK, SSH, HJC, DKK
Funding acquisition: SP, KCM, YGK, HL
Investigation: SP, CHB, HC, JIP, ESK, JPL, SHP, HWK, SHH, HJC, DKK
Writing–original draft: SP, YK, KCM, YGK, HL
Writing–review & editing: All authors
All authors read and approved the final manuscript.

ORCID

Sehoon Park, https://orcid.org/0000-0002-4221-2453
Yisak Kim, https://orcid.org/0000-0002-5420-1436
Chung Hee Baek, https://orcid.org/0000-0001-7611-2373
Hyunjeong Cho, https://orcid.org/0000-0002-6005-3484
Ji In Park, https://orcid.org/0000-0003-4662-3759
Eun Sil Koh, https://orcid.org/0000-0003-1282-7876
Jung Pyo Lee, https://orcid.org/0000-0002-4714-1260
Sun-Hee Park, https://orcid.org/0000-0002-0953-3343
Hyung Woo Kim, https://orcid.org/0000-0002-6305-452X
Seung Hyeok Han, https://orcid.org/0000-0001-7923-5635
Ho Jun Chin, https://orcid.org/0000-0002-3710-0190
Dong Ki Kim, https://orcid.org/0000-0002-5195-7852
Kyung Chul Moon, https://orcid.org/0000-0002-1969-8360
Young-Gon Kim, https://orcid.org/0000-0003-2148-1299
Hajeong Lee, https://orcid.org/0000-0002-1873-1587

References

1. Schena FP, Nistor I. Epidemiology of IgA nephropathy: a global perspective. *Semin Nephrol* 2018;38:435–442.
2. Cai GY, Chen XM. Immunoglobulin A nephropathy in China: progress and challenges. *Am J Nephrol* 2009;30:268–273.
3. Lee H, Kim DK, Oh KH, et al. Mortality of IgA nephropathy patients: a single center experience over 30 years. *PLoS One* 2012;7:e51225.
4. Hamano T, Imaizumi T, Hasegawa T, et al. Biopsy-proven CKD etiology and outcomes: the Chronic Kidney Disease Japan Cohort (CKD-JAC) study. *Nephrol Dial Transplant* 2023;38:384–395.
5. Barbour SJ, Coppo R, Zhang H, et al. Evaluating a new international risk-prediction tool in IgA nephropathy. *JAMA Intern Med* 2019;179:942–952.
6. Kidney Disease: Improving Global Outcomes (KDIGO) Glomerular Diseases Work Group. KDIGO 2021 clinical practice guideline for the management of glomerular diseases. *Kidney Int* 2021;100:S1–S276.
7. Zhang J, Huang B, Liu Z, et al. External validation of the International IgA Nephropathy Prediction Tool. *Clin J Am Soc Nephrol* 2020;15:1112–1120.
8. Joo YS, Kim HW, Baek CH, et al. External validation of the international prediction tool in Korean patients with immunoglobulin A nephropathy. *Kidney Res Clin Pract* 2022;41:556–566.
9. Haaskjold YL, Lura NG, Bjørneklett R, Bostad L, Bostad LS,

Knoop T. Validation of two IgA nephropathy risk-prediction tools using a cohort with a long follow-up. *Nephrol Dial Transplant* 2023;38:1183–1191.

10. Barbour SJ, Coppo R, Er L, et al. Updating the International IgA Nephropathy Prediction Tool for use in children. *Kidney Int* 2021;99:1439–1450.
11. Testa F, Fontana F, Pollastri F, et al. Automated prediction of kidney failure in IgA nephropathy with deep learning from biopsy images. *Clin J Am Soc Nephrol* 2022;17:1316–1324.
12. Chen T, Li X, Li Y, et al. Prediction and risk stratification of kidney outcomes in IgA nephropathy. *Am J Kidney Dis* 2019;74:300–309.
13. Li Y, Chen T, Chen T, et al. An interpretable machine learning survival model for predicting long-term kidney outcomes in IgA nephropathy. *AMIA Annu Symp Proc* 2020;2020:737–746.
14. Liu Y, Zhang Y, Liu D, et al. Prediction of ESRD in IgA nephropathy patients from an Asian cohort: a random forest model. *Kidney Blood Press Res* 2018;43:1852–1864.
15. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. Advances in Neural Information Processing Systems 31 (NeurIPS 2018). Proceedings of the 32nd International Conference on Neural Information Processing Systems; December 3-8, 2018; Montréal, Canada. Neural Information Processing Systems Foundation, Inc; 2018. p. 6639–6649.
16. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001;29:1189–123.
17. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data* 2020;7:94.
18. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *J Mach Learn Res* 2019;20:1–30.
19. Nagpal C, Yadlowsky S, Rostamzadeh N, Heller K. Deep Cox mixtures for survival regression. *Proc Mach Learn Res* 2021;149:674–708.
20. Ali M. PyCaret: an open source, low-code machine learning library in Python [Internet]. PyCaret, c2020 [cited 2022 Dec 22]. Available from: https://www.pycaret.org
21. Milea D, Najjar RP, Zhubo J, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med* 2020;382:1687–1695.
22. Noseworthy PA, Attia ZI, Behnken EM, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. *Lancet* 2022;400:1206–1212.
23. Rim TH, Lee CJ, Tham YC, et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *Lancet Digit Health* 2021;3:e306–e316.
24. Rank N, Pfahringer B, Kempfert J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *NPJ Digit Med* 2020;3:139.
25. Thompson A, Carroll K, Inker LA, et al. Proteinuria reduction as a surrogate end point in trials of IgA nephropathy. *Clin J Am Soc Nephrol* 2019;14:469–481.
26. Zheng Y, Wang Y, Liu S, et al. Potential blood pressure goals in IgA nephropathy: prevalence, awareness, and treatment rates in chronic kidney disease among patients with hypertension in China (PATRIOTIC) study. *Kidney Blood Press Res* 2018;43:1786–1795.
27. Trimarchi H, Barratt J, Cattran DC, et al. Oxford Classification of IgA nephropathy 2016: an update from the IgA Nephropathy Classification Working Group. *Kidney Int* 2017;91:1014–1021.