



BMJ Open Development of a machine learning-based prediction model for extremely rapid decline in estimated glomerular filtration rate in patients with chronic kidney disease: a retrospective cohort study using a large data set from a hospital in Japan

Daijo Inaguma ¹, Hiroki Hayashi,² Ryosuke Yanagiya,³ Akira Koseki,⁴ Toshiya Iwamori,⁴ Michiharu Kudo,⁴ Shingo Fukuma ⁵, Yukio Yuzawa²

To cite: Inaguma D, Hayashi H, Yanagiya R, *et al*. Development of a machine learning-based prediction model for extremely rapid decline in estimated glomerular filtration rate in patients with chronic kidney disease: a retrospective cohort study using a large data set from a hospital in Japan. *BMJ Open* 2022;**12**:e058833. doi:10.1136/bmjopen-2021-058833

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-058833>).

Received 31 October 2021
Accepted 23 May 2022



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Daijo Inaguma;
daijo@fujita-hu.ac.jp

ABSTRACT

Objectives Trajectories of estimated glomerular filtration rate (eGFR) decline vary highly among patients with chronic kidney disease (CKD). It is clinically important to identify patients who have high risk for eGFR decline. We aimed to identify clusters of patients with extremely rapid eGFR decline and develop a prediction model using a machine learning approach.

Design Retrospective single-centre cohort study.

Settings Tertiary referral university hospital in Toyooka city, Japan.

Participants A total of 5657 patients with CKD with baseline eGFR of 30 mL/min/1.73 m² and eGFR decline of ≥30% within 2 years.

Primary outcome Our main outcome was extremely rapid eGFR decline. To study-complicated eGFR behaviours, we first applied a variation of group-based trajectory model, which can find trajectory clusters according to the slope of eGFR decline. Our model identified high-level trajectory groups according to baseline eGFR values and simultaneous trajectory clusters. For each group, we developed prediction models that classified the steepest eGFR decline, defined as extremely rapid eGFR decline compared with others in the same group, where we used the random forest algorithm with clinical parameters.

Results Our clustering model first identified three high-level groups according to the baseline eGFR (G1, high GFR, 99.7±19.0; G2, intermediate GFR, 62.9±10.3 and G3, low GFR, 43.7±7.8); our model simultaneously found three eGFR trajectory clusters for each group, resulting in nine clusters with different slopes of eGFR decline. The areas under the curve for classifying the extremely rapid eGFR declines in the G1, G2 and G3 groups were 0.69 (95% CI, 0.63 to 0.76), 0.71 (95% CI 0.69 to 0.74) and 0.79 (95% CI 0.75 to 0.83), respectively. The random forest model identified haemoglobin, albumin and C reactive protein as important characteristics.

Conclusions The random forest model could be useful in identifying patients with extremely rapid eGFR decline.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ We also adapted a unique and novel approach for clustering using the hierarchy of curve steepness and trajectory, which makes the algorithm efficient and fast.
- ⇒ A key limitation is that training of our model only relies on electronic health record data we have, and, thus, even though our data were collected from a large hospital and we use fair algorithms, our resultant models may incur inherent bias in the data if existing.
- ⇒ We adopted an intriguing design of prediction model using machine learning, which differentiates the rapid decline group from the other groups using subject's covariates with such groups, which are clustered by the shape of estimated glomerular filtration rate declines using a novel and practical automatic trajectory clustering algorithm.

Trial registration UMIN 000037476; This study was registered with the UMIN Clinical Trials Registry.

INTRODUCTION

The number of patients with chronic kidney disease (CKD) is increasing worldwide, resulting in an increased number of patients requiring dialysis and kidney transplantation and suffering from cardiovascular (CV) events.^{1–4} According to the Global Burden Disease study data, the incidence, prevalence and mortality rate of CKD increased by 89%, 87% and 98%, respectively, between 1990 and 2016.⁵ However, many patients with CKD are asymptomatic until the kidney function deteriorates. Therefore, a diagnosis of CKD at an

earlier stage is required. However, trajectories of glomerular filtration rate (GFR) vary between patients and are reported to depend on the primary kidney disease, blood pressure and proteinuria.^{6–8} We believe that it would be useful if we could accurately predict the extremely rapid estimated GFR (eGFR) decline before the deterioration begins to determine the causes of deterioration, avoid renotoxicities and provide earlier treatment with renoprotective drugs. In addition, there may be crucial causes such as incidence of rapidly progressing glomerulonephritis accompanying original kidney diseases especially in patients with extremely rapid eGFR decline. If causes of kidney function deterioration are identified earlier, we could treat the patient appropriately.

Artificial intelligence (AI) has been in development since the 1980s, and investigations using machine learning have been progressing in various fields, including medicine.^{9–13} Machine learning can identify irregularities in data and deal with large data sets with complex variables and relationships (ie, ‘big data’). Therefore, machine learning can often be used for the prediction of associated phenomena from big data in healthcare.^{14–16} In nephrology, for instance, AI has enabled a more precise equation for eGFR.^{17,18}

Our hospital has maintained a large database of more than 900 000 patients who were treated and followed up for various diseases since 2004. We previously demonstrated a prediction model for patients with an eGFR decline of $\geq 30\%$ within 2 years using machine learning.¹⁹ However, we realised that the patterns of eGFR decline varied even among those patients. A more detailed prediction is crucial in the real-world clinical settings because of higher risks of progression to end-stage kidney disease and the incidence of CV disease. Therefore, we focused on extremely rapid eGFR declines and analysed a much larger population data of 914 280 patients from the single-centre database than in the previous model. In addition, the purpose of the present study was to adopt an intriguing design of prediction model using a novel and practical automatic trajectory clustering algorithm. To the best of our knowledge, no AI-based methods have been proposed to create a prediction model related to the trajectories of eGFR, especially among patients with extremely rapid eGFR decline. Therefore, we aimed to create a model for extremely rapid eGFR decline among patients with CKD with an eGFR decline of $\geq 30\%$ within 2 years based on a large database and using machine learning.

METHODS

Data set and samples

We used a database of 914 280 patients from the Fujita Health University Hospital between June 2004 and July 2019. Medical data were available for 286 494 patients with eGFR, of which the findings in 29 466 patients included the following CKD criteria: an eGFR < 60 mL/min/1.73 m² and/or urine protein $\geq 1+$ on a dipstick for > 90 days.

Patients of < 20 years of age and those who had undergone kidney transplantation were excluded. When measuring eGFR, we used the average eGFR measurements over the preceding 90 days to avoid temporal spikes in measurements. On detection of different and distinct spans of GFR decline in the same patient, we included the first value in the analyses. Patients with CKD with an eGFR decline of $\geq 30\%$ within 2 years, defined as rapid eGFR decline according to previous reports, were enrolled.^{20–22} Overall, there were 7315 such samples in the study. Of these, we only included samples with an initial eGFR of ≥ 30 mL/min/1.73 m² because we aimed to detect patients with extremely rapid eGFR decline at an earlier stage. In addition, many patients with eGFR < 30 mL/min/1.73 m² have a clinical course of extremely rapid eGFR decline within a short period because the lower the initial eGFR, the more rapid is the eGFR decline, in general. Hence, we excluded patients with advanced kidney dysfunction. Finally, 5657 unique samples of GFR decline were analysed.

Clustering eGFR decline curves

To automatically cluster the eGFR decline curves, we used an expansion of the group-based trajectory model posited by Nagin.²³ The original algorithm was modified as follows: the curves were hierarchically grouped by the initial eGFR values and curve steepness. We used a single response variable for the eGFR. We sampled 10 points with equal time intervals after applying linear interpolation to the response variables to obtain 10 points with equal time intervals for each patient. To obtain such 10 points by sampling for the entire population, we applied linear interpolation of the response variables on all of them. We also did not assume that those points were on the specific function of the elapsed times.

To modify the existing method for hierarchical grouping, we used the following equation:

$$f(y) = \sum_{m=1}^M \sum_{k=1}^K p_m p_{m,k} f_y(y | C_1 = m, C_2 = k),$$

where m indicates the groups of initial eGFR values and k indicates the groups of curve steepness. The notation y is the response variable consisting of 10 points on the eGFR curve, while the probability that y belongs to class m is denoted by p_m , and to class m and k is denoted by $p_{m,k}$. Note that we hierarchically grouped the curves using p_m and $p_{m,k}$, which are both estimated from the data as model parameters. The notations M and K are the total number of those classes. The function $f_y(y | C_1 = m, C_2 = k)$ is the conditional density of the observed data, m is the initial eGFR value of class C_1 and k is the curve steepness of class C_2 .

To form the conditional density function, $f_y(y | C_1 = m, C_2 = k)$, we assumed that each point in y is generated from the Gaussian distribution:

$$\mu_{m,k,d,t_d} = \beta_{m,k,d} t_d, \text{ for } d = 2, \dots, T,$$

$$\mu_m = \beta_m$$

$$y^d | C_1 = m, C_2 = k \sim N(\mu_{m,k,d,t_d}, \sigma_d^2), \text{ for } d = 2, \dots, T,$$

$$\mu_{m,k,d,t_d} = \beta_{m,k,d} t_d, \text{ for } d = 2, \dots, T,$$

where β is the estimated distribution parameter. In this study, we did not estimate the variance in the distribution, σ . In our model, y^1 was determined only based on the initial eGFR class of m , whereas the latter points of y were determined by m and k and the timestamp t . Note that, unlike the original model, we assumed that each point is independently generated from the Gaussian distribution.

In our experiments, we used 3 for m and 3 for k , which are set to give a similar number of trajectory groups as Nagin reported.²³

Samples for classification

For each type of high-eGFR, middle-eGFR and low-eGFR decline curve, we classified the most acute curves. In the low-GFR decline group, which included 2437 samples, curves were categorised in detail as mild, moderate and acute; we identified 222 positive samples of acute curves and 2215 negative samples of mild and moderate curves. In the middle-GFR decline group, which included 2652 samples, the curves were categorised as mild and moderate ($n=2139$) and acute ($n=513$). In the high-GFR decline group, which included 568 samples, acute curves included 103 samples, while moderate and severe curves included 465 samples.

Features used for the classification model

With the aforementioned positive and negative samples of eGFR decline, we extracted the laboratory values, including 15 longitudinal data as well as 5 static data. The former includes transferrin saturation, blood urea nitrogen, serum uric acid, haemoglobin, haemoglobin A1c, ferritin, eGFR, systolic and diastolic blood pressures, C reactive protein (CRP); body mass index, serum total cholesterol, serum creatinine, serum albumin and urine protein. The latter includes sex, age, comorbidity of diabetes mellitus, history of acute kidney injury and prescription of renin–angiotensin system inhibitors. Summarising these longitudinal data to form explanatory variables in the prediction model, we used the average and SD for each variable over 90, 180, 360 days to the beginning of the eGFR decline and the exponentially smoothed average (ESA) of all available past data. The methods are shown in online supplemental table 1). These nine summarisation methods are then applied to all longitudinal data, respectively, and, thus, we have 135 longitudinal features. By adding static data, we use 140 features in total as inputs to our classification models.

Model solving

We used logistic regression and random forest algorithm for classification. To evaluate the proposed model, we tuned hyperparameters, including the number of trees in the forest, the minimum number of samples required at a leaf node and the minimum number of samples required to split an internal node, for example, random forest. To identify the best parameters in the inner four-fold cross-validation, we evaluated the random forest and logistic

regression models using outer five-fold cross-validation. Note that we found the best hyper parameters only using the training data of each fold of the cross-validation, not using the test data. By solving the model, we can compute the probability of a patient to be classified in the group of the most acute curves. The following is the formula to compute the probability on using logistic regression:

$$\text{probability} = \frac{1}{1 + e^{-\sum_{i=0}^n \alpha_i (x_i - \text{mean}_i) / \text{std}_i}}$$

where α_i is trained parameters, x_i is a feature value indexed by i ,

mean_i , std_i are the mean and SD for i th feature values, and n is the number of features. Actual values for α_i , mean_i and std_i are shown in online supplemental table 2, online supplemental table 3 and online supplemental table 4 to compute the probability. We omitted the computation while using random forest to avoid complexity but the probability can be computed by averaging the results of each decision tree used in the algorithm. We then used the area under the curve (AUC) of the receiver operating characteristic (ROC) curve as representative performance metrics, which is computed as the mean of the results of fivefold. In our experiment setting, validation data sets were created from the data that were not used in training for each fold, which is a common and practical method when evaluating the model even using one cohort. Other statistical parameters (such as sensitivity and specificity) were computed using the fold, which has a median AUC out of five folds. The best cut-off was found at a point of the ROC curve at the minimum distance from the top left corner, which is commonly used when determining the cut-off as well as Youden index, taking the AUC into account.²⁴ We should note that such derivation of ‘best cut-off’ could lead to bias^{25 26} where bias reduction such as using smoothed ROC curve and others is discussed.²⁵ In this study, we mostly use AUC for comparing and evaluating prediction performance as suggested in a previous report.²⁶ For examining the importance of features, the contributions to the eGFR decline were examined according to the weight of each variable in logistic regression and by Gini impurity in random forest model.

In this study, we applied logistic regression and random forest using the Python code with scikit-learn library (<https://scikit-learn.org/>) as well as for classification model solving. For solving the clustering model, we used PyStan (<https://pystan.readthedocs.io/en/latest/>), which is the Python interface of Stan (<https://mc-stan.org/>). Stan is a general statistical modelling platform where we declaratively depicted our probabilistic models as we explained before. Note that we used the Markov chain Monte Carlo algorithm to estimate model parameters including p_m , $p_{m,k}$, β_m and $\beta_{m,k,d}$.

Statistical analysis

We compared the all-cause mortality among the nine subgroups. The data on the outcome were obtained

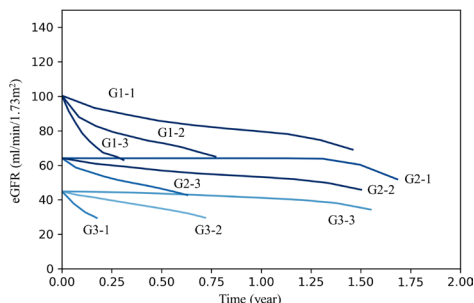


Figure 1 Patterns of eGFR decline classified using machine learning. eGFR, estimated glomerular filtration rate.

from the medical records. All-cause mortality rates were compared using log rank test with Kaplan-Meier curves.

Patient and public involvement

Patients were not involved at any stage of this research.

RESULTS

Patterns of eGFR decline classified by machine learning

The patients were automatically classified into three groups according to eGFR at the reference points (G1: high; G2: middle and G3: low) and further divided into nine subgroups according to the rate of eGFR decline using machine learning (G1-1, G2-1, G3-1: low-rate decline in each group; G1-2, G2-2, G3-2: intermediate-rate decline in each group and G1-3, G2-3, G3-3: high-rate decline in each group) (figure 1). The G1-3, G2-3 and G3-3 subgroups were defined as those with extremely rapid eGFR decline.

Comparison of patient characteristics and laboratory data at reference points between the groups

Table 1 summarises the comparisons of patient characteristics and laboratory data at the reference points between the subgroups with high and other rates of eGFR decline. Patients with extremely rapid eGFR decline were significantly older than those in the other groups in G1 and G2. Blood haemoglobin and serum total cholesterol and albumin levels were significantly lower in the extremely rapid eGFR decline subgroups in each group. Serum CRP levels were significantly higher in the extremely rapid eGFR decline subgroups in each group.

Comparison of cumulative all-cause survival rate between the subgroups

We compared the cumulative survival rate among the nine groups. Significant differences were observed between them (log-rank test: $p < 0.001$).

AUC and calibration plot in each group

Figure 2 illustrates the ROC curves and calibration plots with its slope and intercept for prediction of extremely rapid eGFR decline in each group according to the random forest-based model. Table 2 summarises the AUC of the logistic regression and the random forest models for prediction of extremely rapid eGFR decline.

The AUCs of the G1, G2 and G3 groups according to the logistic regression model were 0.682, 0.647 and 0.754, respectively, and those according to the random forest-based model were 0.694, 0.712 and 0.788, respectively. We conducted the same analysis without including the serum creatinine level because we considered multicollinearity between eGFR and serum creatinine level. Subsequently, the AUCs of the G1, G2 and G3 groups according to the random forest-based model were 0.687, 0.705 and 0.789, respectively. The calibration plot of G1 indicated that the predicted probabilities of the machine learning model were close to the actual probabilities. Meanwhile, in G2 and G3, the higher the predicted probabilities, the higher were the actual probabilities compared with the predicted probabilities.

Features affecting the prediction in the three groups

Online supplemental figure 3 illustrates the heatmap for features that affected the random forest-based model. The redder a column, the higher is its effect on extremely rapid eGFR decline; in contrast, the greener a column, the lesser is its effect. Kidney function including eGFR; age; diastolic blood pressure and albumin, cholesterol, haemoglobin and uric acid levels were demonstrated as features potentially useful for distinguishing G3 from G1 and G2. Figure 3 summarises the ranking of the top 10 features in the random forest model. Except for the features of kidney function, including eGFR and creatinine levels, the features related to haemoglobin and cholesterol were ranked high in G1, those related to albumin and haemoglobin were ranked high in G2, and those related to CRP and albumin were ranked high in G3.

DISCUSSION

The prediction model that we created could detect patients with CKD with extremely rapid decline in eGFR using machine learning. The results of the present study have three features. First, the patient data of the present study were obtained from a large-scale database. There have been some reports concerning CKD by analysing big data.^{27–29} However, most of the studies in this field have included the general population. We believe that the present study is significant, as it examined a large number of patients with various diseases. Second, we used AI to analyse and create different prediction models, including random forests. AI-based disease prediction is progressing in many fields. Electronic medical record systems have been in use for >10 years in many hospitals in Japan. Therefore, a large amount of information, including laboratory data, can be analysed using AI. Machine learning enables the addition of an exponential smoothing average for different variables from a large amount of data. Third, we found that the variables that affect the pattern of eGFR decline vary according to the kidney function at baseline. We were able to classify into three groups automatically according to eGFR at the reference points.

Table 1 Baseline characteristics and laboratory findings

Variables	G1-1, 2 (N=465)	G1-3 (n=103) Extremely rapid eGFR decline	P value	G2-1, 2 (N=2139)	G2-3 (N=513) Extremely rapid eGFR decline	P value	G3-1, 2 (N=2215)	G3-3 (N=222) Extremely rapid eGFR decline	P value
Age (years) [*]	67(51, 76)	74(65.5, 80)	<0.001	73(65, 79)	75(67, 80)	0.002	74(67, 81)	76(65, 83)	0.402
Female sex (%)	54.6	48.5	0.264	44.0	46.8	0.253	39.9	42.8	0.396
Comorbidity of DM (%)	30.5	24.3	0.207	25.2	24.2	0.614	26.0	19.8	0.042
History of AKI (%)	2.1	1.9	0.895	1.6	2.3	0.243	2.6	5.0	0.045
BMI (kg/m ²)	22.4, 4.4	21.9, 4.2	0.236	23.0, 3.9	22.3, 3.9	<0.001	22.9, 3.8	22.1, 3.6	0.013
SBP (mm Hg)	131, 27	134, 29	0.479	134, 26	132, 26	0.107	134, 26	132, 28	0.424
DBP (mm Hg)	76, 17	73, 18	0.059	75, 14	74, 16	0.226	73, 15	73, 16	0.995
Use of RASB (%)	43.4	46.6	0.559	52.0	52.2	0.933	59.3	52.2	0.041
eGFR (mL/min/1.73 m ²)	99.8, 17.8	99.4, 23.9	0.251	62.5, 10.2	64.2, 10.7	<0.001	43.2, 7.3	48.7, 10.3	<0.001
Serum creatinine (mg/dL)	0.56, 0.11	0.56, 0.13	0.696	0.84, 0.16	0.81, 0.17	<0.001	1.21, 0.37	1.09, 0.38	<0.001
BUN (mg/dL)	14.0, 4.9	14.8, 6.5	0.785	17.3, 5.3	17.7, 6.2	0.610	22.4, 7.2	21.1, 8.6	<0.001
Haemoglobin (mg/dL)	12.3, 2.1	11.3, 2.3	<0.001	12.7, 1.9	11.9, 2.2	<0.001	12.0, 2.0	11.2, 2.2	<0.001
Serum T-chol (mg/dL)	192, 54	163, 44	<0.001	189, 44	181, 45	<0.001	188, 45	181, 60	0.003
Serum uric acid (mg/dL)	4.7, 1.6	4.7, 1.9	0.819	5.7, 1.5	5.5, 1.6	0.077	6.4, 1.6	6.2, 2.0	0.047
Serum albumin (g/dL)	3.55, 0.74	3.17, 0.78	<0.001	3.91, 0.55	3.64, 0.69	<0.001	3.86, 0.56	3.41, 0.69	<0.001
CRP (mg/dL) [‡]	0.3(0.3, 1.6)	1.2(0.3, 4.1)	<0.001	0.3(0.3, 0.6)	0.3(0.3, 1.5)	<0.001	0.3(0.3, 0.6)	0.7(0.3, 2.8)	<0.001
Transferrin saturation (%)	24.0, 14.6	25.3, 13.6	0.512	24.8, 13.4	24.1, 13.5	0.560	26.5, 13.3	23.8, 14.5	0.024
Ferritin (ng/mL) [*]	118(35, 307)	152(74, 580)	0.189	111(44, 271)	208(61, 381)	0.038	164(63, 328)	190(87, 423)	0.397
Haemoglobin A1c (%)	6.8, 1.8	6.8, 2.1	0.660	6.4, 1.4	6.2, 1.2	0.007	6.3, 1.2	6.1, 1.0	0.137
Urine protein [†]	1.6, 1.5	1.3, 1.3	0.053	1.2, 1.4	1.0, 1.3	0.029	1.4, 1.5	1.4, 1.5	0.983
Urine protein [‡]	2(0, 3)	1(0, 2)	<0.001	0(0, 2)	0(0, 2)	0.029	1(0, 3)	1(0, 3)	0.983

Data are shown as mean (SD), value (%).

*Median (first quartile, third quartile), as appropriate.

†Continuous value of urine protein test by dipstick.

‡Semi-quantitative test of urine protein by dipstick 50% (25%, 75%).

AKI, acute kidney injury; BMI, body mass index; BUN, blood urea nitrogen; CRP, C reactive protein; DBP, diastolic blood pressure; DM, diabetes mellitus; eGFR, estimated glomerular filtration rate; RASB, renin-angiotensin system blocker; SBP, systolic blood pressure; T-chol, total cholesterol.

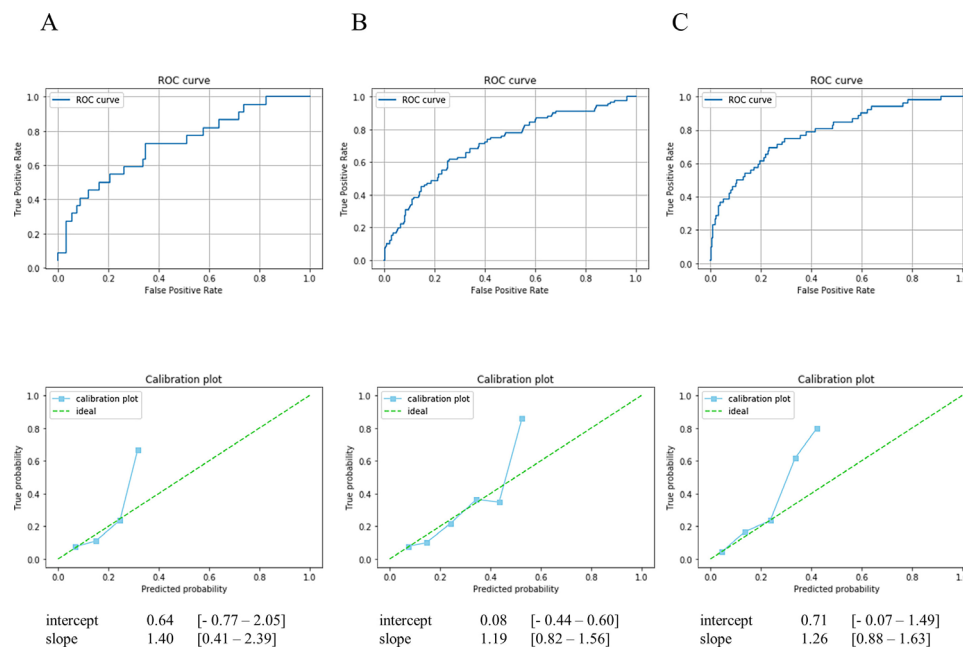


Figure 2 Receiver-operating characteristic (ROC) curves and calibration plots in each group using random forest-based model. (A) G1, (B) G2, (C) G3.

In general, kidney function in patients with CKD gradually worsens. However, patterns of eGFR decline, which are called trajectories of eGFR, vary between patients.³⁰ A report from six large-scale, randomised controlled trials revealed that more cases presented with non-linear eGFR decline among diabetes patients.³¹ Many reports have demonstrated risk factors related to the decline in eGFR. The Chronic Renal Insufficiency Cohort study, which was conducted in the USA, clarified many factors associated with CKD progression in patients with pre-dialysis CKD.^{32–35} These risk factors included proteinuria, the presence of inflammatory cytokines and elevated serum uric acid levels. In Japan, the Chronic Kidney Disease Japan Cohort revealed that anaemia, blood pressure and albuminuria were independent risk factors of CKD

progression.³⁶ The eGFR of the patients enrolled in the two representative cohorts ranged from 10 mL/min/1.73 m² to 70 mL/min/1.73 m². Meanwhile, we decided that an eGFR ≥ 30 mL/min/m² at baseline was the cut-off in the present study because the period before initiating renal replacement therapy was extremely short to classify the patterns of eGFR decline. Furthermore, patients with an eGFR of ≥ 90 mL/min/m² at baseline were enrolled. We created a prediction model for identifying patients with rapid eGFR decline among those with CKD in our previous study.¹⁹ However, we could not adjust the models and stratify them according to eGFR. Machine learning enables the classification of trajectories of eGFR decline into nine patterns using eGFR at baseline and the rate of eGFR decline. Interestingly, we found that different

Table 2 Ability of prediction models (with serum creatinine)

Outcome	Group	AUC	Sensitivity	Specificity	PPV	NPV	PLR	NLR
Logistic regression	G1	0.76 (0.72–0.80)	0.62 (0.48–0.76)	0.75 (0.70 to 0.79)	0.20 (0.13–0.26)	0.95 (0.93–0.97)	2.44 (1.85–3.22)	0.51 (0.35–0.74)
	G2	0.65 (0.62–0.68)	0.52 (0.42–0.62)	0.70 (0.66 to 0.74)	0.29 (0.23–0.36)	0.86 (0.82–0.90)	1.72 (1.36–2.18)	0.69 (0.56–0.85)
	G3	0.67 (0.61–0.73)	0.67 (0.47–0.87)	0.73 (0.64 to 0.82)	0.36 (0.21–0.51)	0.91 (0.84–0.97)	2.48 (1.58–3.90)	0.46 (0.25–0.85)
Random forest	G1	0.69 (0.63–0.76)	0.68 (0.49–0.88)	0.65 (0.56 to 0.75)	0.32 (0.19–0.45)	0.90 (0.82–0.97)	1.96 (1.31–2.92)	0.49 (0.26–0.92)
	G2	0.71 (0.69–0.74)	0.60 (0.50–0.71)	0.74 (0.70 to 0.78)	0.33 (0.26–0.40)	0.90 (0.87–0.93)	2.33 (1.85–2.93)	0.53 (0.41–0.69)
	G3	0.79 (0.75–0.83)	0.67 (0.55–0.80)	0.77 (0.73 to 0.81)	0.26 (0.18–0.33)	0.95 (0.93–0.97)	2.90 (2.25–3.74)	0.43 (0.29–0.63)

AUC, area under curve; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio.

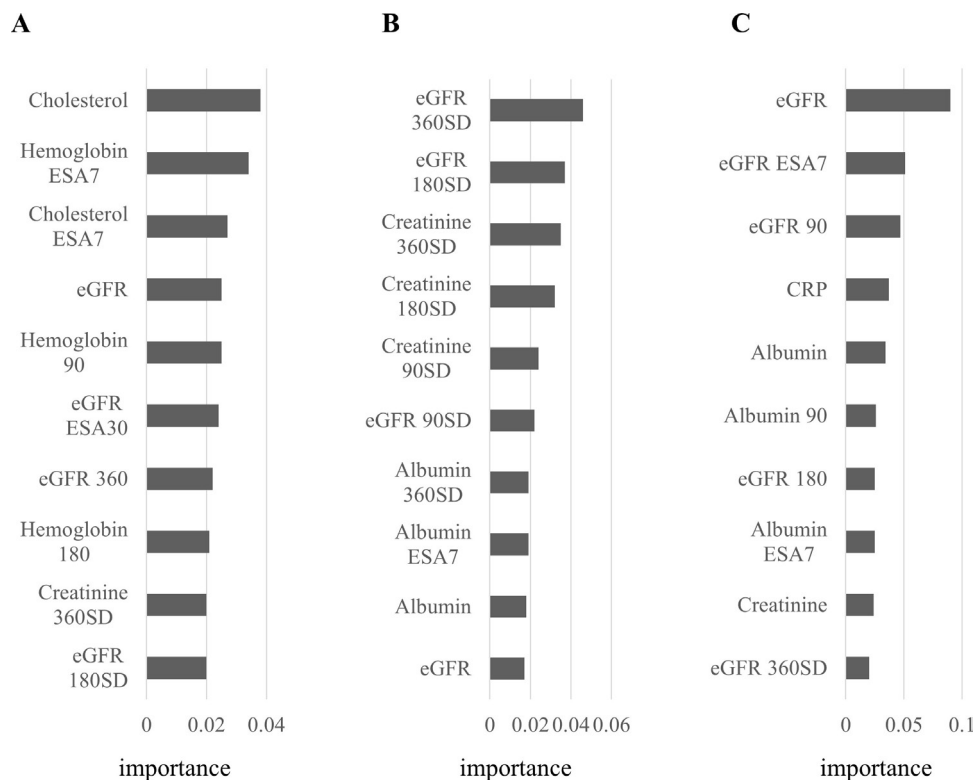


Figure 3 Variable importance for random forest model in each group. (A) G1, (B) G2, (C) G3. eGFR, estimated glomerular filtration rate; ESA, exponentially smoothed average.

clinical parameters, including haemoglobin and CRP levels, were more important in predicting an extremely rapid eGFR decline according to the baseline kidney function. Some prediction models for eGFR decline based on AI have been reported.²⁷ Raynaud *et al* recently demonstrated that the donor age, eGFR, proteinuria and pathological findings of the transplanted kidney predicted progression to end-stage kidney disease in kidney transplanted patients who were classified into eight groups.³⁷ Meanwhile, our model included features that we used as variables, including haemoglobin and CRP levels, which were not used in other models. Factors related to anaemia, such as haemoglobin ESA7 and the 90-day and 180-day averages, were associated with trajectories of eGFR in patients with an eGFR of approximately 90 mL/min/m². Anaemia is often accompanied by CKD because the reduction of functional kidney mass leads to a decrease in the production and secretion of erythropoietin. However, renal anaemia usually develops at an eGFR of <30 mL/min/1.73 m². In the present study, we defined anaemia as a haemoglobin level below its lower normal limit. In other words, anaemia might be detected more strictly in the study than in a clinical setting. Therefore, we found that the management of renal anaemia may be vital in patients in earlier stages of CKD. In contrast, the factors related to serum albumin were associated with the trajectories of eGFR in patients with an eGFR of approximately 60 mL/min/m² and 40 mL/min/m². We believe that the progression of CKD causes not only an increase in the amount of proteinuria but also

malnutrition. Interestingly, the serum CRP level ranked high in patients with an eGFR of only 40 mL/min/m², which suggested that inflammation might be more significant in patients in advanced CKD stages. Additionally, eGFR starts decreasing closer to critical points, and larger changes in variables can be observed. It is possible that the predicted probability increases by adding an exponential smoothing average for different variables. We showed good calibration for the G1 group and lesser for the G2 and G3 groups. If nephrologists estimate higher risk in patients in G2 and G3 from the present prediction model, practices for renal protection such as use of renin-angiotensin system blockers, sodium glucose cotransporter-2 inhibitors, erythropoiesis-stimulating agents and protein restriction can be changed. Many studies have indicated that proteinuria or albuminuria is crucial risk factors for a decline in kidney function, incidence of CV disease and all-cause mortality. Contrary to expectations, proteinuria did not significantly influence the prediction, as opposed to our previous report on a prediction model for rapid eGFR decline.¹⁹ We believe that this was because patients with extremely rapid eGFR decline who were at a higher risk were enrolled. A relatively high amount of proteinuria was already recognised in the patients in this study. Some clinical variables have been reported to be risk factors of eGFR decline.^{38–43} To create a more precise prediction model for trajectory of eGFR decline, more variables related to kidney function should be considered. To accomplish that, it is necessary to conduct a prospective study in the future.

The present study has some limitations. First, the participants might have suffered from not only kidney diseases but also other diseases. Therefore, the results may not always apply to the general population. Unfortunately, we could not get detailed information about the diseases that the patients suffered from because items included in the data were huge and complicated. We considered that underlying diseases might be an independent risk factor. Second, the intervals between the frequency of examinations, including blood tests, differed between patients. Therefore, we used the average values over periods of 90, 180 and 360 days prior to the reference points. Third, the results of the calibration plots were different among the three groups. This could be because the criteria for CKD in G1 differed from that in G2 and G3. Most patients in G1 were defined to have CKD with only proteinuria because the eGFR values at the start of the decline were approximately 90 mL/min/1.73 m²; therefore, eGFR values of <60 mL/min/1.73 m², which is one of the conditions for CKD, were rare. However, in both G2 and G3, most patients were diagnosed with CKD using both eGFR and proteinuria because the eGFR values at the start of the rapid decline were approximately 60 and 40 mL/min/1.73 m², respectively. Additionally, in both G2 and G3, the actual observation probabilities were higher than the predicted probabilities by approximately 0.3. Therefore, we may need to estimate the risk of decline in eGFR than the predicted probabilities in patients with relatively high predicted probabilities.

CONCLUSION

We have created a prediction model for extremely rapid eGFR decline using machine learning to identify patients at a high risk. Further statistical research includes exploiting recent deep learning algorithms, which could enable to use more complex and multimodal features and to provide higher classification and calibration performance. Recent generative approaches using deep learning also has the potential to directly predict the shape of eGFR decline curves even in a non-parametric approach. To use this model in a real-world clinical setting in the future, we could intervene by preventing eGFR decline. A typical usage is, thus, by inputting patient's features to the model, showing whether a patient is in the group of the extremely rapid eGFR decline, and also exhibiting the typical shape of eGFR decline. Such an application can be provided by being connected to the electronic health record system of a hospital to obtain features and showing the model results via a personal computer or mobile interfaces.

To this end, validations using external datasets are needed because our models are created from one hospital data as well as prospective studies to confirm the accuracy of the present results.

Author affiliations

¹Internal Medicine, Fujita Health University Bantane Hospital, Nagoya, Japan

²Nephrology, Fujita Health University, Toyoake, Japan

³Medical Information Systems, Fujita Health University, Toyoake, Japan

⁴IBM Research, Tokyo, Japan

⁵Human Health Science, Kyoto University, Kyoto, Japan

Contributors DI, RY and YY were involved in study design and data interpretation. DI and HH contributed to the writing of the manuscript. AK, TI and MK analysed the data. SF overviewed and criticized the manuscript. All authors were involved in drafting, reviewing, and approving the final manuscript. AK and TI had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. DI acts in as the guarantor.

Funding This study was supported by research funds with no restrictions on publication from Kyowa Kirin Co., Ltd. IBM Research provided support for this study in the form of salaries for A. Koseki, T. Iwamori and M. Kudo. There are no patents, products in development or marketed products associated with this research.

Competing interests Yes, there are competing interests for one or more authors and I have provided a Competing Interests statement in my manuscript and in the box below.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Ethics Committee: Center for research promotion and support, Fujita Health University number ID: approval number: HM19-157. We used only medical information, which was not included identified personal information such as name, ID number of medical records. However, informed consent was obtained in the form of opt-out on the website.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Daijo Inaguma <http://orcid.org/0000-0002-3977-5933>

Shingo Fukuma <http://orcid.org/0000-0002-8379-8761>

REFERENCES

- 1 Global, GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 2020;395:709-33.
- 2 Hill NR, Fatoba ST, Oke JL, *et al*. Global prevalence of chronic kidney disease - a systematic review and meta-analysis. *PLoS One* 2016;11:e0158765.
- 3 International society of nephrology Ason, European Renal Association - European Dialysis and Transplant Association. The hidden epidemic: worldwide, over 850 million people suffer from kidney diseases; PRESS RELEASE. 2018.
- 4 Imai E, Horio M, Watanabe T, *et al*. Prevalence of chronic kidney disease in the Japanese general population. *Clin Exp Nephrol* 2009;13:621-30.
- 5 Xie Y, Bowe B, Mokdad AH, *et al*. Analysis of the Global Burden of Disease study highlights the global, regional, and national trends of

- chronic kidney disease epidemiology from 1990 to 2016. *Kidney Int* 2018;94:567–81.
- 6 Li L, Astor BC, Lewis J, et al. Longitudinal progression trajectory of GFR among patients with CKD. *Am J Kidney Dis* 2012;59:504–12.
 - 7 Li T-C, Yu T-Y, Li C-I, et al. Three-Year renal function trajectory and its association with adverse renal event in patients with type 2 diabetes. *J Diabetes Complications* 2018;32:784–90.
 - 8 Soohoo M, Streja E, Obi Y, et al. Predialysis kidney function and its rate of decline predict mortality and hospitalizations after starting dialysis. *Mayo Clin Proc* 2018;93:1074–85.
 - 9 Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur Urol* 2015;67:1142–51.
 - 10 Niel O, Bastard P. Artificial intelligence in nephrology: core concepts, clinical applications, and perspectives. *American Journal of Kidney Diseases* 2019;74:803–10.
 - 11 Hamet P, Tremblay J. Artificial intelligence in medicine. metabolism. *clinical and experimental*. 2017:69s–S36.
 - 12 Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial intelligence in cardiology. *J Am Coll Cardiol* 2018;71:2668–79.
 - 13 Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;2:719–31.
 - 14 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
 - 15 Damen JAAG, Hooft L, Schuit E. Prediction models for cardiovascular disease risk in the general population: systematic review. *Bmj clinical research* 2016;353:i2416.
 - 16 Collins GS, Mallett S, Omar O, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103.
 - 17 Liu X, Li N, Lv L, et al. Improving precision of glomerular filtration rate estimating model by ensemble learning. *J Transl Med* 2017;15:231.
 - 18 Li N, Huang H, Qian H-Z, et al. Improving accuracy of estimating glomerular filtration rate using artificial neural network: model development and validation. *J Transl Med* 2020;18:120.
 - 19 Inaguma D, Kitagawa A, Yanagiya R, et al. Increasing tendency of urine protein is a risk factor for rapid eGFR decline in patients with CKD: a machine learning-based prediction model by using a big database. *PLoS One* 2020;15:e0239262.
 - 20 Coresh J, Turin TC, Matsushita K, et al. Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. *JAMA* 2014;311:2518–31.
 - 21 Levey AS, Inker LA, Matsushita K, et al. GFR decline as an end point for clinical trials in CKD: a scientific workshop sponsored by the National kidney Foundation and the US food and drug administration. *Am J Kidney Dis* 2014;64:821–35.
 - 22 Matsushita K, Chen J, Sang Y, et al. Risk of end-stage renal disease in Japanese patients with chronic kidney disease increases proportionately to decline in estimated glomerular filtration rate. *Kidney Int* 2016;90:1109–14.
 - 23 Nagin DS. Group-Based trajectory modeling: an overview. *Ann Nutr Metab* 2014;65:205–10.
 - 24 Akobeng AK. Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatr* 2007;96:644–7.
 - 25 Leeftang MMG, Moons KGM, Reitsma JB, et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clin Chem* 2008;54:729–37.
 - 26 Wynants L, van Smeden M, McLernon DJ, et al. Three myths about risk thresholds for prediction models. *BMC Med* 2019;17:192.
 - 27 Niel O, Boussard C, Bastard P. Artificial intelligence can predict GFR decline during the course of ADPKD. *Am J Kidney Dis* 2018;71:911–2.
 - 28 Kanda E, Kanno Y, Katsukawa F. Identifying progressive CKD from healthy population using Bayesian network and artificial intelligence: a worksite-based cohort study. *Sci Rep* 2019;9:5082.
 - 29 Norouzi J, Yadollahpour A, Mirbagheri SA, et al. Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system. *Comput Math Methods Med* 2016;2016.
 - 30 Kerschbaum J, Rudnicki M, Dzien A, et al. Intra-individual variability of eGFR trajectories in early diabetic kidney disease and lack of performance of prognostic biomarkers. *Sci Rep* 2020;10:19743.
 - 31 Weldegiorgis M, de Zeeuw D, Li L, et al. Longitudinal estimated GFR trajectories in patients with and without type 2 diabetes and nephropathy. *Am J Kidney Dis* 2018;71:91–101.
 - 32 Yang W, Xie D, Anderson AH, et al. Association of kidney disease outcomes with risk factors for CKD: findings from the chronic renal insufficiency cohort (CRIC) study. *Am J Kidney Dis* 2014;63:236–43.
 - 33 Townsend RR, Anderson AH, Chirinos JA, et al. Association of pulse wave velocity with chronic kidney disease progression and mortality: findings from the CRIC study (chronic renal insufficiency cohort). *Hypertension* 2018;71:1101–7.
 - 34 Koye DN, Magliano DJ, Reid CM, et al. Risk of progression of nonalbuminuric CKD to end-stage kidney disease in people with diabetes: The CRIC (Chronic Renal Insufficiency Cohort) study. *Am J Kidney Dis* 2018;72:653–61.
 - 35 Kwan B, Fuhrer T, Zhang J, et al. Metabolomic markers of kidney function decline in patients with diabetes: evidence from the chronic renal insufficiency cohort (CRIC) study. *Am J Kidney Dis* 2020;76:511–20.
 - 36 Inaguma D, Imai E, Takeuchi A, et al. Risk factors for CKD progression in Japanese patients: findings from the chronic kidney disease Japan cohort (CKD-JAC) study. *Clin Exp Nephrol* 2017;21:446–56.
 - 37 Raynaud M, Aubert O, Reese PP, et al. Trajectories of glomerular filtration rate and progression to end stage kidney disease after kidney transplantation. *Kidney Int* 2021;99:186–97.
 - 38 Voormolen N, Noordzij M, Grootendorst DC, et al. High plasma phosphate as a risk factor for decline in renal function and mortality in pre-dialysis patients. *Nephrol Dial Transplant* 2007;22:2909–16.
 - 39 Chang W-xiu, Arai S, Tamura Y, et al. Time-dependent risk factors associated with the decline of estimated GFR in CKD patients. *Clin Exp Nephrol* 2016;20:58–70.
 - 40 Chang AR, Grams ME, Ballew SH, et al. Adiposity and risk of decline in glomerular filtration rate: meta-analysis of individual participant data in a global consortium. *BMJ* 2019;144:k5301.
 - 41 Eriksen BO, Stefansson VTN, Jenssen TG, et al. High ambulatory arterial stiffness index is an independent risk factor for rapid age-related glomerular filtration rate decline in the general middle-aged population. *Hypertension* 2017;69:651–9.
 - 42 Perkins RM, Bucaloiu ID, Kirchner HL, et al. Gfr decline and mortality risk among patients with chronic kidney disease. *Clin J Am Soc Nephrol* 2011;6:1879–86.
 - 43 Portale AA, Wolf MS, Messenger S, et al. Fibroblast growth factor 23 and risk of CKD progression in children. *Clin J Am Soc Nephrol* 2016;11:1989–98.