BMC Nephrology

RESEARCH

Open Access

Enhancing individual glomerular filtration rate assessment: can we trust the equation? Development and validation of machine learning models to assess the trustworthiness of estimated GFR compared to measured GFR

Antoine Lanot[1,2,3*], Anna Akesson[4,5], Felipe Kenji Nakano[6,7], Celine Vens[6,7], Jonas Björk[8], Ulf Nyman[9], Anders Grubb[10], Per-Ola Sundin[11], Björn O. Eriksen[12], Toralf Melsom[12], Andrew D. Rule[13], Ulla Berg[14], Karin Littmann[15], Kajsa Åsling-Monemi[16], Magnus Hansson[17], Anders Larsson[18], Marie Courbebaisse[19], Laurence Dubourg[20], Lionel Couzi[21], Francois Gaillard[22], Cyril Garrouste[23], Lola Jacquemont[24], Nassim Kamar[25], Christophe Legendre[26], Lionel Rostaing[27], Natalie Ebert[28], Elke Schaeffner[28], Arend Bökenkamp[29], Christophe Mariat[30], Hans Pottel[6†] and Pierre Delanaye[31,32†]

Abstract

Background Creatinine-based estimated glomerular filtration rate (eGFR) equations are widely used in clinical practice but exhibit inherent limitations. On the other side, measuring GFR is time consuming and not available in routine clinical practice. We developed and validated machine learning models to assess the trustworthiness (i.e. the ability of equations to estimate measured GFR (mGFR) within 10%, 20% or 30%) of the European Kidney Function Consortium (EKFC) equation at the individual level.

Methods This observational study used data from European and US cohorts, comprising 22,343 participants of all ages with available mGFR results. Four machine learning and two traditional logistic regression models were trained on a cohort of 9,202 participants to predict the likelihood of the EKFC creatinine-derived eGFR falling within 30% (p30), 20% (p20) or 10% (p10) of the mGFR value. The algorithms were internally and then externally validated on cohorts of respectively 3,034 and 10,107 participants. The predictors included in the models were creatinine, age, sex, height, weight, and EKFC.

Results The random forest model was the most robust model. In the external validation cohort, the model achieved an area under the curve of 0.675 (95%CI 0.660;0.690) and an accuracy of 0.716 (95%CI 0.707;0.725) for the P30

†Hans Pottel and Pierre Delanaye contributed equally to this work as last senior authors.

*Correspondence:
Antoine Lanot
antoine.lanot@gadz.org
Full list of author information is available at the end of the article

criterion. Sensitivity was 0.756 (95%CI 0.747;0.765) and specificity was 0.485 (95%CI 0.460; 0.511) at the 80% probability level that EKFC falls within 30% of mGFR. At the population level, the PPV of this machine learning model was 89.5%, higher than the EKFC P30 of 85.2%. A free web-application was developed to allow the physician to assess the trustworthiness of EKFC at the individual level.

Conclusions A strategy using machine learning model marginally improves the trustworthiness of GFR estimation at the population level. An additional value of this approach lies in its ability to provide assessments at the individual level.

Keywords Glomerular filtration rate, Chronic kidney disease, Creatinine, Machine learning, Random forest

Background

The evaluation of kidney function through glomerular filtration rate (GFR) stands as a cornerstone of routine care across numerous clinical cases. The GFR result is necessary to adjust the dose of drugs cleared by the kidney. It defines and assesses the severity of chronic kidney disease (CKD). Additionally, it serves as a guiding metric for nephrologists in determining the need for renal transplantation or dialysis [1].

Measured GFR (mGFR) using an exogenous marker such as iohexol is the gold standard for evaluating GFR. However, these techniques are more time consuming and more expensive than using biomarkers. Moreover, these techniques are not universally available.

Alternatively, estimated GFR (eGFR) is commonly used in clinical practice, derived from equations incorporating a single measurement of serum creatinine. Over the last three decades, various equations have been developed. The European Kidney Function Consortium (EKFC) equation was shown to be one of the most accurate for European and US populations [2, 3]. Nevertheless, no single equation offers a perfect estimation of GFR, often displaying a systematic bias and more importantly, imprecision at the individual level. It is noteworthy that a key benchmark for the validation of eGFR equations is achieving a high accuracy within 30% (P30). P30 indicates the percentage of eGFR results within 30% of mGFR.

Serum creatinine concentration depends on muscle mass, and accordingly equations for estimation of GFR incorporate age and sex. However, these variables may not consistently capture the impact of all non-GFR determinants of creatinine. This is reflected by the fact that the best equations achieve a P30 of 80 to 85%, meaning that in more than one patient out of ten, creatinine-based equations are inaccurate to estimate GFR [4]. Consequently, physicians are frequently faced with the question of whether to rely on eGFR results or consider direct GFR measurement when interpreting patient results.

Artificial intelligence (AI) represents a transformative paradigm in healthcare, offering personalized solutions to clinical challenges, in contrast to traditional evidence derived from population-level trials [5].

We aimed at developing a machine learning model to aid clinicians in assessing the trustworthiness of eGFR obtained by the EKFC creatinine-based equation, individualizing this estimate according to patients' characteristics.

Methods

Study design and setting

We used data from cohorts of European participants of all ages and US adult participants. Details about the cohorts have been published previously and are detailed in supplementary material (Supplementary Table S1A, Supplementary Table S1B, Supplementary Table S1C) [2, 3]. There was no patient or public involvement in the design or drafting of this study. All data were anonymized from the source cohorts. the original study was approved by the Ethical Board at Lund University (Sweden) with amendment approved by the Swedish Ethical Review Agency. Procedures involving humans and data were realized in agreement with the ethical principles for medical research involving human subjects established in the World Medical Association's Declaration of Helsinki. Written consent had been obtained from the participants of MDRD, ALTOLD, CRISP, GENOA/ECAC and PERL studies. A waiver of consent was obtained from the Mayo Clinic IRB to study the patients from the Mayo Clinic Renal Studies Unit database due to the retrospective nature of these clinical data.

Study population

European data were those used for the development, internal and external validation of the EKFC creatinine-based equation [2]. US data were extracted among those used to assess the performance of the EKFC creatinine-based equation in the US population, notably data available from the National Institute of Diabetes and Digestive and Kidney Diseases [3]. Height and weight were not available from four of the US cohorts, and therefore these cohorts were not included in the present work.

Data set splitting

The population was separated into a training and an internal validation datasets from the same cohorts, and an external validation dataset from different cohorts.

The training cohort consisted of 9,202 participants: 1245 from the USA, 500 from Germany, 3,096 from France, 3,158 from Sweden, and 1203 from Norway. There were 3,034 participants in the internal validation cohort: 1,001 from France, 157 from Germany, 424 from Norway, 1,037 from Sweden and 415 from the USA. The external validation cohort encompass 10,107 participants: 442 from Belgium, 2,572 from France, 447 from Netherland, 3,281 from Sweden, 394 from the UK and 2,971 from the USA. It should be pointed out that none of the patients in the external validation dataset were selected from the cohorts used for the development of the EKFC equation. The distribution of the cohorts is detailed in Supplementary Material (Supplementary Table S1).

Covariates and outcomes

We used the EKFC creatinine-derived eGFR, which is calculated based on a normalized serum creatinine value. This value is obtained by dividing the serum creatinine by the Q-value, which represents the sex- and age-specific median creatinine value in healthy individuals [2]. Race-free Q-values were used for the participants from the US cohorts [3]. Details of the computation of EKFC are presented in supplementary material (Supplementary Methods S1). The measured glomerular filtration rate (GFR) value was determined for each participant utilizing plasma or urinary clearance of an exogenous filtration marker such as iohexol, inulin, ^{51}Cr-EDTA, or iothalamate.

The biomarker used for GFR estimation was serum creatinine, measured with assays traceable to the gold standard method of isotope dilution mass spectrometry. Age, sex, height, and weight were the variables available. In the European cohorts, all subjects were considered as White as race is frequently unavailable. Because the sample size of both Black Americans and Black Europeans was low, we decided not to include them in this work.

For each participant, the probability of GFR assessment falling within an acceptable margin of error was computed. Three margins of errors were assessed: 30%, 20% and 10% respectively referred to as P30 or p30, P20 or p20 and P10 or p10. For instance, if the relative difference between the mGFR and the EKFC creatinine based eGFR was less than 30% of the mGFR, then EKFC would be considered the recommended GFR assessment method within the acceptable margin of error. In order to clarify the distinction between the proportion of patients meeting the margin at the population level and the probability of eGFR being within 30% at the individual level, we used different notations: "P30" referring to the population-level measure and "p30" referring to the individual level. Main analysis focused on P30 (p30), but P20 (p20) and P10 (p10) criteria were explored in secondary analyzes.

Missing data

Globally there were 248 subjects with missing data, concerning only weight and height. This corresponded to fewer than 1% of missing data, and therefore we chose to exclude the corresponding 248 persons and to perform complete case analysis.

Statistical analysis

Continuous variables were described using mean and standard deviation (SD) or median and interquartile range (IQR), according to their distribution. Reporting of the machine learning models was realized according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [6] statement (Supplementary Methods S2).

Machine learning model development and validation

We employed supervised machine learning techniques to train models aimed at predicting the likelihood of the EKFC-derived eGFR falling within p30, p20 and p10. The models developed were random forest, extreme gradient boosting, k-nearest neighbors, and support vector machine. The features included in the models were age, sex, height, weight, creatinine, and EKFC creatinine eGFR. Features were scaled and centered during preprocessing. For each model, we performed a tenfold inner cross-validation repeated 3 times in the training dataset, and the hyperparameters were optimized with a grid search algorithm (Supplementary Method S3).

Traditional statistical prediction models

We compared the performances of the four machine learning models with those of a traditional logistic regression model. Some of the features included in the machine learning models presented multicollinearity, as attested by the evaluation of the variance inflation factors (VIF) in a "full feature" model. In order to make fair comparisons, we first computed a logistic regression model using all the features included in the machine learning algorithms, called logistic regression "inflated" model. We then built a second logistic regression model without EKFC, called "reduced" logistic regression model. VIF values are shown in supplementary material (Supplementary Table 2).

Performances of the models

The performances of the models were evaluated in the internal validation dataset in terms of area under the curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). For each model, the metrics were assessed with the threshold probability optimizing the sensitivity–specificity balance, meaning that the recommended method predicted by a model was EKFC if the probability that the EKFC-based creatinine eGFR fell within p30 exceeded the threshold. The choice of the best performing model was made considering the metrics obtained on the internal validation cohort, and with a priority given to AUC. AUC values were compared using the Delong's test [7]. The p-values were adjusted using a Bonferroni correction to address concern related to multiple comparisons.

Features importance

In order to get an insight into the mechanisms driving the machine learning models, features importance was computed for the best performing model. Features importance were computed by measuring the average decrease in impurity contributed by each variable, calculated using the Gini impurity. In a set of items with N possible classes, the Gini impurity is a measure used to determine how often a randomly chosen item would be incorrectly labeled if it was randomly labeled. For example, the Gini impurity of a set of elements that all belong to the same class, i.e. perfectly pure, is zero. Its maximum value is 0.5 for a set of elements of two classes equally distributed [8].

All analyses were performed using R (version 4.2.0; R Foundation for Statistical Computing, Vienna, Austria). Machine learning models were computed with the CRAN package *caret*.

Results

The study encompassed data from 22,343 participants, with 4,631 hailing from the USA and 17,712 from Europe. The participants ranged in age from 2 to 97 years, with a median age of 53 years (IQR 23–65), and 10,709 (47.9%) participants were female. Median serum creatinine was 76.6 µmol/l (IQR 61.0–106.2), while the median value for EKFC creatinine eGFR and mGFR were 82 ml/min/1,73m^2 (IQR 57–97), and 80 (IQR 54–99) respectively.

Patient's characteristics categorized by dataset are detailed in Table 1. Notably, patients were younger in the training dataset, with higher height and heavier weight than in the external validation dataset. Creatinine mean value was lower, and accordingly eGFR and mGFR were higher in the training and in the internal validation datasets than in the external validation dataset.

Outcome evaluation

EKFC creatinine eGFR was correctly predicted as within 30% of the actual mGFR value (P30) in 19,217 (86.01%, 95% confidence interval (CI) 85.55 to 86.46) participants in the whole population. Similar values of P30 were observed in the training dataset: 7,972 (86.63%, 95% confidence interval 85.92 to 87.31), in the internal validation dataset: 86.88%, 95%CI 85.63 to 88.04, and in the external validation dataset: 85.18%, 95%CI 84.47 to 85.86.

Fifteen thousand and 968 (71.47%, 95%CI 70.87 to 72.06) subjects had an EKFC creatinine eGFR within 20% of the mGFR (P20 criteria). Finally, EKFC creatinine eGFR was within 10% of mGFR for 9,454 (42.31%, 95%CI 41.67 to 42.96) participants.

Internal validation

P30

The receiver operating characteristics (ROC) curves of the six models developed are presented in Fig. 1A. Random forest and extreme gradient boosting models showed the highest AUC (respectively 0.695, 95%CI 0.667;0.724 and 0.668, 95%CI 0.639;0.698), followed by the k-nearest neighbor model (AUC 0.644, 95%CI 0.616;0.672). AUCs of the full variables logistic regression model and the reduced logistic regression model were respectively 0.615, 95%CI 0.584;0.645, and 0.623, 95%CI 0.593;0.653), both superior to support vector machines model (AUC 0.583, 95%CI 0.551;0.614) (Fig. 1B). The AUC of the Random forest model was significantly superior to AUC of the five others models. P-value for the difference in random forest's and extreme gradient boosting's models was 0.02 according to DeLong's test. The accuracy of the random forest model was 0.770, (95%CI 0.754; 0.784) with a sensitivity of 0.819 (95%CI 0.804; 0.834) and a specificity of 0.442 (95%CI 0.393; 0.493). The full logistic model had the highest accuracy: 0.868 (95%CI 0.856; 0.880) at the price of an imbalance between sensitivity of 0.999 (95%CI 0.997; 1) and specificity of 0.003 (95%CI 0; 0.014). Metrics of the six algorithms for the P30 criteria on the internal validation dataset are shown in Fig. 1B.

P20 and P10

Regarding the P20 criterion, the random forest achieved the highest AUC, at 0.655 (95%CI 0.634; 0.677). However, this was not significantly different from the AUC of the extreme gradient boosting model, which was 0.644 (95%CI 0.622; 0.666) ($p=0.73$ according to DeLong's test).

Table 1 Patients' characteristics in the different datasets

	Training (*N* = 9202)	Internal validation (*N* = 3034)	External validation (*N* = 10107)	Overall (*N* = 22343)
Country				
France	3096 (33.6%)	1001 (33.0%)	2572 (25.4%)	6669 (29.8%)
Germany	500 (5.4%)	157 (5.2%)	0	657 (2.9%)
Norway	1203 (13.1%)	424 (14.0%)	0	1627 (7.3%)
Sweden	3158 (34.3%)	1037 (34.2%)	3281 (32.5%)	7476 (33.5%)
USA	1245 (13.5%)	415 (13.7%)	2971 (29.4%)	4631 (20.7%)
Belgium	0	0	442 (4.4%)	442 (2.0%)
Netherland	0	0	447 (4.4%)	447 (2.0%)
UK	0	0	394 (3.9%)	394 (1.8%)
Age				
Median [IQR]	50.9 [18.7, 64.2]	50.8 [18.5.0, 63.6]	55.0 [39.0, 66.0]	53.0 [23.0, 65.0]
Sex				
Female	4257 (46.3%)	1406 (46.3%)	5046 (49.9%)	10,709 (47.9%)
Male	4945 (53.7%)	1628 (53.7%)	5061 (50.1%)	11,634 (52.1%)
Weight (kg)				
Median [IQR]	69.3 [54.8, 83.0]	69.0 [55.0, 83.0]	73 [60.0, 85.5]	71 [57.5, 84.1]
Height (cm)				
Median [IQR]	167 [158.0, 175.0]	167.0 [158.1, 175.5]	168 [160.0, 175.0]	167.4 [159.7, 175.0]
Creatinine (µmol/l)				
Median [IQR]	74.0 [59.0, 99.0]	73.7 [59.2, 97.6]	79.6 [62.8, 118.0]	76.6 [61.0, 107.0]
EKFC eGFR (ml/min/1.73m^2)				
Median [IQR]	84.5 [61.7, 99.1]	84.9 [63.7, 99.1]	77.7 [50.6, 94.3]	81.6 [57.0, 97.1]
mGFR (ml/min/1.73m^2)				
Median [IQR]	82.9 [56.4, 101.6]	84.0 [46.3, 102.3]	76.4 [47.7, 96.0]	80.0 [54.0, 99.0]

eGFR estimated glomerular filtration rate, *mGFR* measured glomerular filtration rate, *IQR* Interquartile range

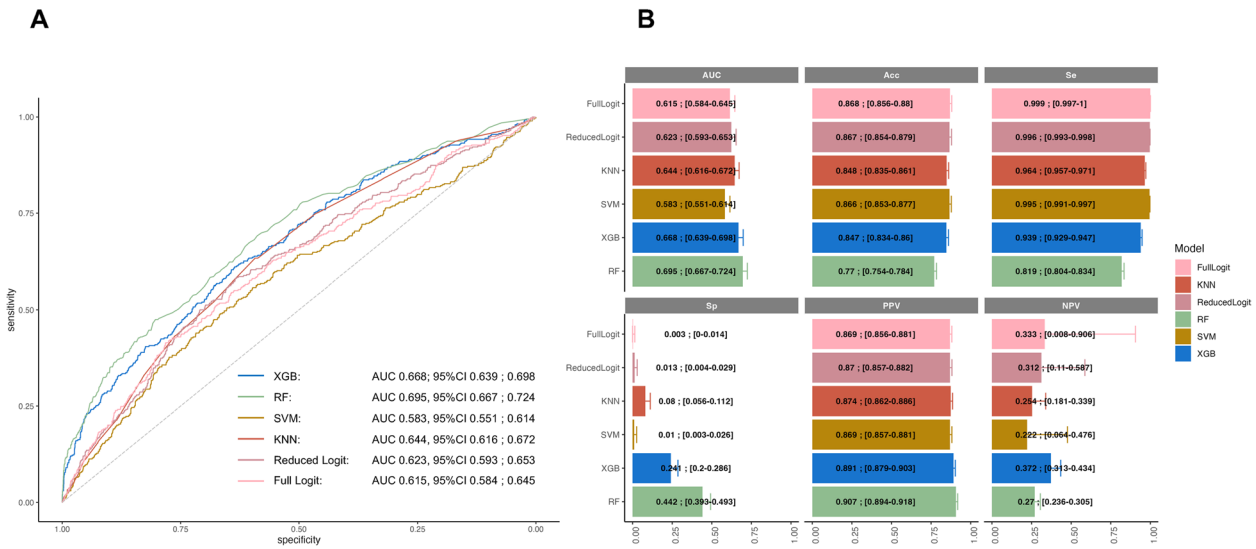

Fig. 1 Performance of the six algorithms in internal validation for the P30 criteria. **A** Receiver Operating Curve for the 6 models. **B** Performance metrics for the 6 models. AUC: area under the curve; Acc: accuracy; Se: sensitivity; Sp: specificity; PPV: Predictive positive value; NPV: Negative predictive value; Logit: Logistic regression model; KNN: k-nearest neighbors; SVM: Support vector machine; XGB: extreme gradient boosting; RF: random forest.

Considering the P10 criterion, the random forest model had the highest AUC (0.621, 95%CI 0.601;0.641), significantly superior to the extreme gradient boosting which ranked second with an AUC of 0.569 (95%CI 0.549; 0.590) ($p < 0.001$). The metrics' performances of the models for the P20 and P10 criteria within the internal validation dataset are reported in supplementary material (Supplementary Table S3).

Choice of the best performing model and external validation

Our aim was to answer a classification problem with a good ability to discriminate between two classes (EKFC-based creatinine eGFR vs mGFR) so we prioritized the AUC to rank the performance of the models built, while also considering the other metrics. The results obtained on the P30 criteria were examined in priority.

P30

The random forest model was the best performing model according to the AUC and accuracy values; therefore, this model was used for the external validation.

Considering the P30 criteria, the AUC of the random forest model was 0.675 (95%CI 0.660;0.690). Its accuracy was 0.716, (95%CI 0.707; 0.725), with a sensitivity of 0.756 (95%CI 0.747;0.765) and a specificity of 0.485 (95%CI 0.460; 0.511) at the 80% probability threshold. The metrics performance for the different criteria are shown in Fig. 2.

Figure 3 illustrates for the participants of the external validation dataset, the probability that EKFC eGFR is within 30% of mGFR according to the 'trustworthiness' prediction of the random forest model (Fig. 3). There were 8,609 (85.2%) of the EKFC predictions within 30% of mGFR in the external validation dataset. In 6,538 of these cases, the random forest model gave a probability $\geq 80\%$ that EKFC is within 30% of mGFR, thus confirming the trustworthiness of the predicted EKFC result. However, in 2071 of these 8,609 cases, the random forest model gave a probability $< 80\%$ that EKFC is within 30% of mGFR, thus incorrectly warning against the trustworthiness of EKFC. On the other hand, there were 1,498 (14.8%) cases with an EKFC prediction deviating more than 30% of mGFR. In 727 out of 1,498 of these cases the random forest model correctly warns against the use of EKFC as a trustworthy GFR assessment within p30, while the random forest model incorrectly gives a probability $> 80\%$ in 771 out of 1,498 of these cases, thus indicating that the EKFC was dependable, while in fact, it was not.

P20 and P10

When tested on the external validation dataset, the random forest trained with aim of predicting the p20 criteria had an AUC of 0.661 (95%CI 0.649; 0.672) and an accuracy of 0.641 (95%CI 0.632; 0.650), with a sensitivity of 0.698 (95%CI 0.687; 0.709) and a specificity of 0.504 (95%CI 0.486; 0.522).

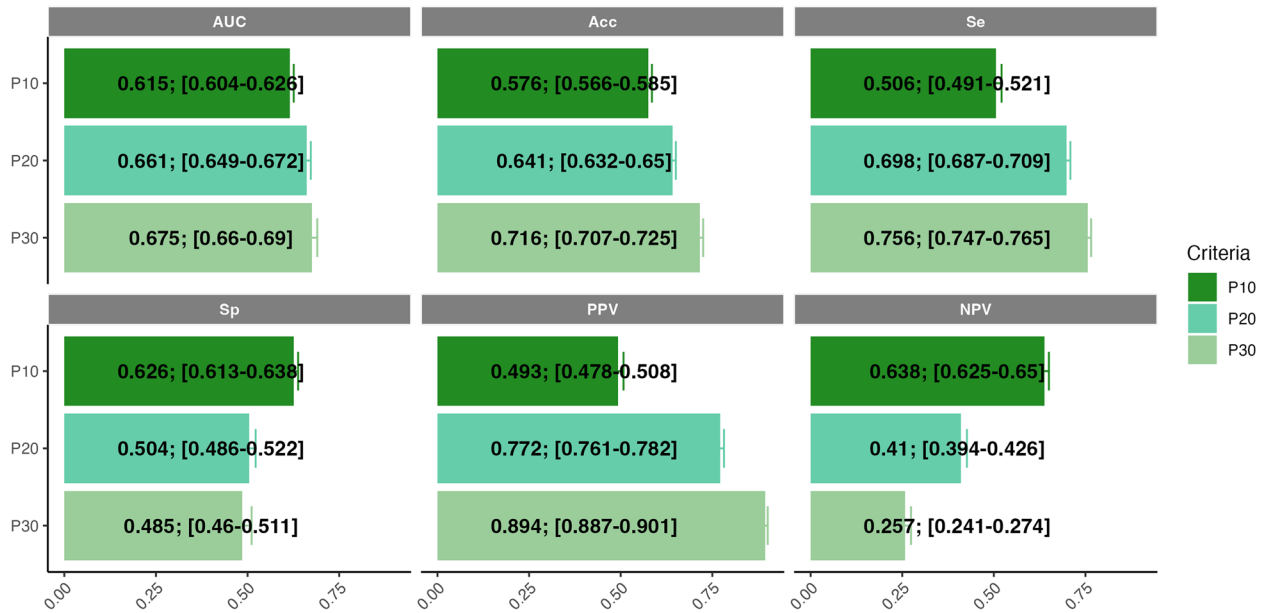

Fig. 2 Performance of the random forest model in external validation for the P30, P20 and P10 criteria. The probability thresholds used to evaluate the model's performance metrics were 0.80 for the P30 criteria, 0.65 for the P20 criteria, and 0.45 for the P10 criteria. AUC: area under the curve; Acc: accuracy; Se: sensitivity; Sp: specificity; PPV: Predictive positive value; NPV: Negative predictive value

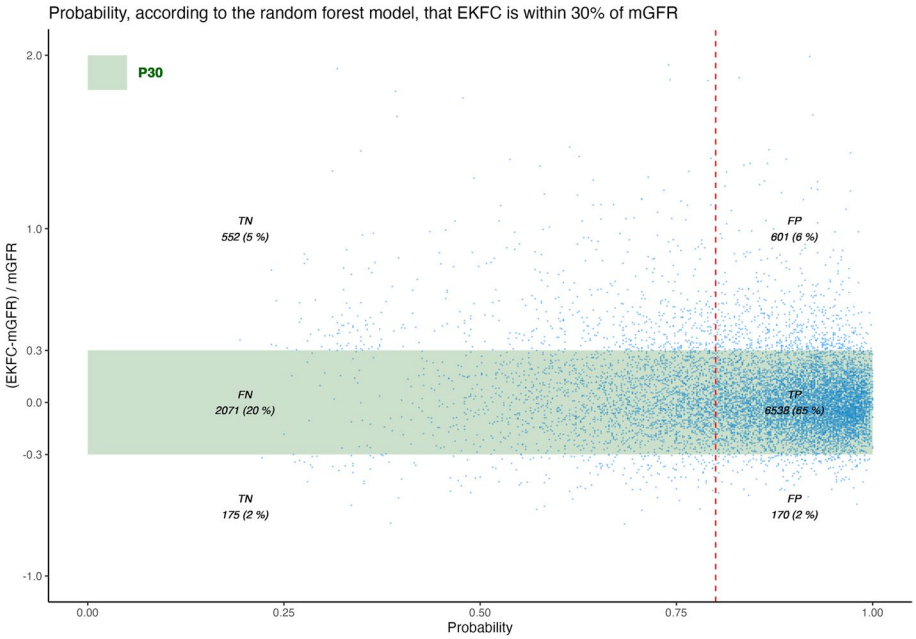

Fig. 3 Probability that EKFC eGFR is within 30% of mGFR according to the random forest model. Each dot represents a participant from the external validation dataset. The probability threshold of 0.80 was selected to evaluate the model's performance metrics. EKFC: European kidney function consortium formula for estimation of GFR; mGFR: measured glomerular filtration rate. P30: EKFC result within 30% of mGFR value. TP: true positive, meaning that the random forest model predicts accurately that EKFC is within P30, TN: true negative, meaning that the random forest model predicts accurately that EKFC is not within P30, FP: false positive, meaning that the random forest model falsely predicts that EKFC is within P30

The model trained for the P10 criteria had an AUC of 0.615 (95%CI 0.604; 0.626), an accuracy of 0.576 (95%CI 0.566; 0.585), and a sensitivity of 0.506 (95%CI 0.491; 0.521) with a specificity of 0.626 (95%CI 0.613; 0.638).

The numbers and probabilities of EKFC eGFR being within p20 or p10 of mGFR according the prediction of the random forest are shown in Supplementary material (Supplementary Figures S1 and S2).

Variable importance

EKFC equation was the most important features in the random forest model with P30 criteria. All features except sex were of similar importance in the model, according to the mean decrease in Gini impurity values, going from 363 for height to 454 for EKFC. The importance of the different features in the random forest models are shown in supplementary material (Supplementary figure S3).

Implementation in a web application

We implemented the random forest algorithm in a free web application (available at https://trustekfcegfr.shinyapps.io/GFR_shiny_Lanot/). This app allows the user to evaluate the probability of EKFC creatinine eGFR being within p30, p20 or p10 for a given patient, knowing his age, sex, height, weight and serum creatinine. The 95%CI for the probabilities, calculated using a bootstrap approach, are presented. As an illustration in ten patients, Table 2 displays the characteristics, mGFR values and the machine learning algorithm's predictions for the likelihood of EKFC eGFR being within 30% of mGFR (Table 2). A screenshot of the application is shown in supplementary material (Supplementary figure S4).

Added-value of a strategy including the machine learning algorithm for GFR assessment

When using EKFC-based creatinine equation in a given patient, the probability that the eGFR is within p30 is 85.2% according to our results in the external validation cohort, in line with the figures previously described at a population level [2, 3].

Global assessment of GFR using our random forest algorithm is presented in Fig. 4. The first step is to assess the random forest prediction according to the subject's characteristics. A threshold of 80% may be chosen by default. If the 'trustworthiness' probability of EKFC eGFR is inferior to the threshold, then a GFR measurement should be performed, with a p30 of 100% (mGFR being the gold standard). In the other case, if the probability predicted by the random forest model is superior or equal to the threshold, then the p30 is the PPV of our model (i.e. the probability of EKFC eGFR being within

Table 2 Examples of prediction of the random forest model to patients from the external validation dataset

Sex	Age (years)	Height (cm)	Weight (kg)	Creatinine (µmol/l)	EKFC (ml/min/1.73m^2)	mGFR (ml/min/1.73m^2)	EKFC within P30	EKFC within P20	EKFC within P10	Chance for EKFC to be within P30 (%) (RF prediction)	Chance for EKFC to be within P20 (%) (RF prediction)	Chance for EKFC to be within P10 (%) (RF prediction
F	67	167	88	98.7	48	51	Yes	Yes	Yes	79.6	51.0	20.4
F	66	155	56	78	64	63	Yes	Yes	Yes	97.2	92.4	68.8
M	39	164	68	218.3	34	29.9	Yes	No	No	70.0	37.8	30.0
M	9	139	36	42.3	108	135	Yes	Yes	No	94.8	89.6	51.2
F	59	164	98	235.1	20	16	Yes	No	No	52.0	27.6	20.4
M	49	186	111	95.0	81	47	No	No	No	76.0	58.6	28.4
F	83	157	52	61.4	70	24	No	No	No	67.4	56.6	39.4
F	20	170	55	84.9	74	63	Yes	Yes	No	98.0	91.4	4.2
M	73	169	68	66.3	82	95	Yes	Yes	No	97.2	88.4	28.8
M	17	175	70	78.5	94	95	Yes	Yes	Yes	99.6	68.6	31.2

BMI Body mass index, *EKFC* European Kidney Function Consortium, *mGFR* measured glomerular filtration rate, *P30* Chance for estimated GFR to be within 30% of mGFR value, *F* Female, *M* Male

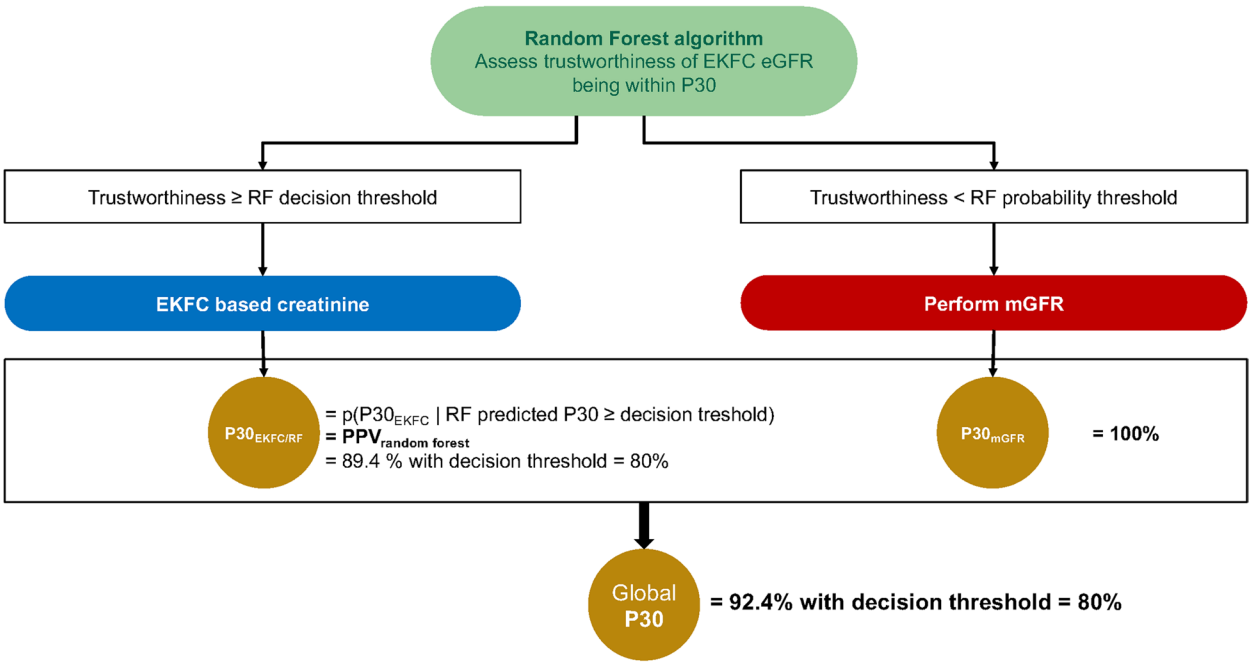

Fig. 4 Use of the random forest algorithm for GFR assesment. Under this approach, if the probability of EKFC eGFR falling within 30% of mGFR is less than 80%, mGFR should be measured. The Global P30 metric is then calculated as the ratio of True Positives to the sum of True Negatives and False Negatives. For global P30 with different probability tresholds values, see Table 3. RF: random forest; PPV: positive predictive value; RF: random forest. EKFC: european kidney function consortium; $P30_{EKFC/RF}$: Probability that EKFC eGFR is within P30 knowing that this probability is superior or equal to the decision treshold, according to the random forest model

p30, known that the random forest model predicted an EKFC trustworthiness superior or equal to the threshold), which is equal to 89.4% with a threshold of 80%. The global P30 of this strategy calculated on the external validation dataset is 92.4% for a threshold of 80%. Global P30 with other threshold values are reported in Table 3, along with the proportion of GFR measurement which would be performed even though EKFC eGFR was within p30 (called "a posteriori un-necessary mGFR") (Table 3). The term "a posteriori un-necessary mGFR" describes the cases in which the random forest strategy would have led to a measurement of GFR while EKFC was indeed within 30% of mGFR, in a clinical context where the clinician would target an assessment within this range.

Table 3 Global P30 with the random forest strategy to assess GFR, according to the chosen threshold

Threshold	PPV of the Random Forest model	Global P30	A posteriori un-necessary mGFR performed
50%	86.4%	86.8%	42.4% (120/283)
60%	86.9%	87.7%	59.0% (364/617)
70%	87.8%	89.4%	68.4% (921/1346)
75%	88.6%	88.5%	70.8% (1377/1945)
80%	89.5%	92.4%	74.0% (2071/2798)
85%	90.4%	94.1%	77.0% (3029/3933)
90%	91.6%	96.1%	79.6% (4320/5425)
95%	93.5%	98.4%	82.6% (6346/7687)

Threshold is the probability computed by the random forest model that is chosen as sufficient to use the EKFC equation (see Fig. 4). PPV is the positive predictive value (i.e. the proportion of individuals whose eGFR falls within 30% of mGFR, given that the random forest model predicted this outcome. Global P30 is the probability of GFR being within 30% of mGFR when the strategy presented in Fig. 4 is used. *A* posteriori un-necessary mGFR performed is the probability of performing mGFR in patients whose eGFR calculated with the EKFC equation was already within 30% of mGFR

BMI Body mass index, *EKFC* European Kidney function consortium, *mGFR* measured glomerular filtration rate, *P30* Chance for estimated GFR to be within 30% of mGFR value, *F* Female, *M* Male

Discussion

In this study, we focused on assessing the trustworthiness of estimated eGFR calculated using the EKFC creatinine equation. We utilized supervised machine learning techniques to develop predictive models aimed at determining the likelihood of eGFR being within a predetermined acceptable margin of error compared to mGFR. Six algorithms were trained on a cohort of 9,202 participants, and internally validated on cohorts of 3,034 subjects. The best performing model, namely random forest algorithm was externally validated on a dataset of 10,107 subjects. This random forest model was implemented in a free web application aimed at aiding clinicians in evaluating eGFR trustworthiness for individual patients.

Artificial intelligence and machine learning tools are steadily gaining prominence in healthcare, presenting opportunities to enhance clinicians' diagnostic abilities, early detection and preventive measures, prognostic accuracy, and treatment strategies [9, 10] Yet it was pointed out that machine learning has been less used in research works in the field of nephrology than in other specialties [11]. Some authors have used artificial intelligence models in works related to kidney function assessment. Some of them have built models to estimate GFR comparing their performance with those of validated equations [12–15]. Others studies aimed at detecting or predicting CKD [16–18]. A systematic review gathered 55 works using artificial intelligence algorithms to predict CKD [19].

To our knowledge, only one study was designed to develop a machine learning model for selecting the optimal GFR assessment among several modalities. In a monocentric Chinese study, data were collected from 518 subjects who underwent ^{99m}Tc-DTPA renal dynamic imaging to detect GFR. A decision tree model was trained to choose the most accurate equation from BIS-2, CKD-EPI with cystatin C, CKD-EPI with creatinine and cystatin C, and Ruijin. Features included in the model were body surface area, BMI, 24-h urine protein, presence of diabetic nephropathy, age and prescription of a RAS inhibitor [20]. There are important differences between the design and objective of this study and our own. Fan et al. tried to determine the best equation to use for a given subject knowing some of his/her characteristics, and finally propose an eGFR value. This approach is finally similar to using machine learning to predict GFR. Instead, our algorithm focuses only on creatinine and one equation (EKFC, which is supposed to be the most accurate to date) to estimate the probability that eGFR is within a given margin of error compared to mGFR.

Despite the enthusiastic promise of high performance, machine learning algorithms sometimes fall short of surpassing well-constructed traditional models [21]. We built two logistic regression models, one of them including all available covariates despite high variance inflation factors, in order to have fair comparisons with the machine learning models which take advantages of all these covariates. Our analysis revealed that machine learning models, particularly random forest and extreme gradient boosting algorithms, outperformed traditional logistic regression models in predicting the recommended method for GFR assessment within acceptable error margins (P30, P20 and P10).

One of the criticisms levelled at machine learning algorithms is the lack of transparency and interpretability of the models, known as the "black box" issue. We assessed variable importance to help understanding the influence of each variable on the results rendered by the algorithm. Analysis revealed that EKFC, age, height, weight and creatinine were of comparable importance while sex had virtually no impact on the model (Supplementary Figure S3).

The use of p30 is a subject of debate, as it represents a relatively broad criterion for clinical decision-making at the individual level. However, current equations for eGFR do not allow for narrower accuracy thresholds. The 2002 KDOQI guidelines deemed P30 satisfactory for clinical interpretation in many scenarios [22]. This position was reaffirmed in the 2024 KDIGO guidelines on evaluation and management of CKD [23]. These guidelines, along with the broader literature, emphasize that when greater accuracy is required, mGFR should be used, even if mGFR is also subject to inherent within-subject variability.

Several strengths of our work may be underlined. We used a very large cohort of participants from several countries in Europe and USA. Participants of all ages, from childhood to old age were included, with renal function ranging from normal to CKD stage 5 in non-dialyzed persons with native kidneys and kidney graft. We have been able to carry out external validation of the models that we developed, which is often lacking in studies presenting prognostic models. Finally, we propose a free web-application made available to physicians for their clinical practice. We warn readers against using this model exclusively. The choice of GFR assessment method should account for several dimensions related to the patient as well as the indication for assessment, and the clinician's experience must prevail. The model should be viewed as a complementary tool for selecting the most appropriate methods for GFR assessment.

Practically, if the probability of EKFC eGFR falling within P30 for an individual is low, a measured GFR should be obtained. If this is not feasible, measuring cystatin C could be used to calculate the mean of creatinine-based and cystatin C-based EKFC estimates, as this approach improves accuracy compared to relying on either biomarker equation alone [24].

Our study presents some limitations. The protocol was not prespecified. We focused on creatinine as the only biomarker included in the EKFC equation whereas cystatin C may be considered to enhance the accuracy of the eGFR with EKFC [24]. Creatinine is the most widely available biomarker worldwide, and the recent KDIGO guidelines on evaluation and management of CKD recommend its use as first choice for GFR evaluation. In these recommendations, cystatin C measurement is advised as second choice and measurement of GFR is considered in cases of potential sources of error

in eGFR with creatinine and cystatin C, and if a more accurate assessment is needed [1]. Moreover, discrepancies may exist between creatinine-based eGFR and cystatin C-based eGFR [25]. Generalizability may be limited to certain populations, because we did not include non-White participants, and the population included contained only European and US subjects.

The difficulties in enhancing the performance of the EKFC equation underscore that EKFC P30 is already operating at a very high level. However, the value of our approach lies in its ability to provide assessments at the individual level. Specifically, it may help to identify outliers where creatinine-based EKFC eGFR estimates may not be reliable, suggesting the need for alternative methods in such cases. The performance of our models may have been constrained by the limited number of features available for training the algorithms. Further research on the topic of eGFR trustworthiness should consider the possibility of including cystatin C in the assessment. The inclusion of more features in the algorithm may improve the performance.

Conclusions

In conclusion, our study demonstrates a marginal potential for machine learning models to improve the trustworthiness of GFR estimation at the population level. The random forest algorithm that we developed enables modest improvement to support clinical decision-making in kidney function assessment. Our results highlight the challenges of using machine learning to address complex clinical problems. Simply increasing the sample size is unlikely to significantly enhance model performance. Instead, we suggest that further studies should focus on incorporating a broader range of variables to optimize model accuracy and applicability.

Abbreviations

GFR	Glomerular filtration rate
eGFR	Estimated glomerular filtration rate
mGFR	Measured glomerular filtration rate
CKD	Chronic kidney disease
P30	Percentage of eGFR results within 30% of mGFR
P20	Percentage of eGFR results within 20% of mGFR
P10	Percentage of eGFR results within 10% of mGFR
AI	Artificial intelligence
EKFC	European kidney function consortium
SD	Standard deviation
IQR	Interquartile range
VIF	Variance inflation factor
BMI	Body mass index
AUC	Area under the curve
PPV	Positive predictive value
NPV	Negative predictive value
95% CI	95% Confidence interval
ROC	Receiver operating curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12882-025-03972-0.

Supplementary Material 1.

Acknowledgements

The Assessing Long Term Outcomes in Living Kidney Donors, Chronic Renal Insufficiency Cohort, Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease, Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications.Preventing Early Renal Loss in Diabetes, African American Study of Kidney Disease and Hypertension, and Modified Diet in Renal Disease studies were performed by respective investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data from these studies reported here were supplied by the NIDDK Central Repository. This manuscript was not prepared in collaboration with the investigators of the different studies and does not necessarily reflect the opinions or views of these studies, the NIDDK Central Repository, or the NIDDK.

Clinical trial number

Not applicable.

Authors' contributions

Research idea and study design: AL, HP, PD; data acquisition: AG, POS, BOE, TM, ADR, UB, KL, KAM, AL, MC, LD, LC, FG, CG, LJ, NK, CL, LR, NE, ES, AB, EJL, CM; data analysis/interpretation: AL, HP, PD, AA, FKN, CV, JB, UN; statistical analysis: AL; supervision or mentorship: HP, PD. Each author contributed important intellectual content during manuscript drafting or revision and agrees to be personally accountable for the individual's own contributions and to ensure that questions pertaining to the accuracy or integrity of any portion of the work, even one in which the author was not directly involved, are appropriately investigated and resolved, including with documentation in the literature if appropriate.

Funding

None.

Data availability

Study protocol and statistical code: available from Antoine Lanot (e-mail, antoine.lanot@gadz.org). Data set: The EKFC data set used in the present study is hosted by the Lund University Population Research Platform. Legal and ethical restrictions prevent public sharing of the data set. Data may be made available to interested researchers for collaboration upon request but would generally require a new ethical permission and the permission of each of the data owners. Contact information for the data host may be found at www.lupop.lu.se.

Declarations

Ethics approval and consent to participate

The original study was approved by the Ethical Board at Lund University (Sweden) with amendment approved by the Swedish Ethical Review Agency. Procedures involving humans and data were realized in agreement with the ethical principles for medical research involving human subjects established in the World Medical Association's Declaration of Helsinki. Written consent had been obtained from the participants of MDRD, ALTOLD, CRISP, GENOA/ECAC and PERL studies. A waiver of consent was obtained from the Mayo Clinic IRB to study the patients from the Mayo Clinic Renal Studies Unit database due to the retrospective nature of these clinical data.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details
[1]Normandie Univ, UNICAEN, CHU de Caen Normandie, Néphrologie, Caen, France. [2]Normandie Université, Unicaen, UFR de Médecine, 2 Rue Des Rochambelles, Caen, France. [3]ANTICIPE" U1086 INSERM-UCN, Centre François Baclesse, Caen, France. [4]Skane University Hospital, Clinical Studies Sweden Forum South, Remissgatan 4, Lund 22185, Sweden. [5]Lund University, Malmö, Sweden. [6]Department of Public Health and Primary Care, KU Leuven, Campus Kulak, Kortrijk, Belgium. [7]Itec, Imec Research Group, KU Leuven, Kortrijk, Belgium. [8]Lund University, Box 117, 221 00 Lund, Sweden. [9]Östra Vallgatan 41, 223 61 Lund, Sweden. [10]Department of Clinical Chemistry and Pharmacology, Laboratory Lund University, Lund 22185, Sweden. [11]Karla Healthcare Centre, Faculty of Medicine and Health, Örebro University, Örebro 701 85, Sweden. [12]University Hospital of North Norway (UNN), 9038 Breivika, Troms, Norway. [13]Division of Nephrology and Hypertension, Mayo Clinic, Rochester, MN, USA. [14]Department of Clinical Science, Intervention and Technology, Division of Pediatrics, Karolinska Institutet, Karolinska University. Hospital Huddinge, 14186 Stockholm, Sweden. [15]Department of Medicine Huddinge, Karolinska Institutet, C2:91 Karolinska University Hospital, Huddinge SE-141 52, Sweden. [16]Barnnjursektionen K 88, Astrid Lindgrens Barnsjukhus, Karolinska University Hospital, Stockholm 141 86, Sweden. [17]Department of Clinical Chemistry, C1:74 Huddinge, Karolinska University Hospital, Stockholm SE-141 86, Sweden. [18]Clinical Chemistry and Pharmacology, Entrance 61, 2Nd Floor, Akademiska Hospital, 751 85 Uppsala, Sweden. [19]Service de Physiologie-Explorations, Fonctionnelles Renales Hopital Europeen Georges Pompidou, 20 Rue Leblanc, Paris 75015, France. [20]Exploration Fonctionnelle Renale Pavillon P, Hopital Edouard Herriot, 5 Place d'Arsonval, 69437, Lyon Cedex 03, France. [21]CHU de Bordeaux, Nephrologie-Transplantation-Dialyse, Hopital Pellegrin, Universite de Bordeaux, Place Amelie Raba Leon, Bordeaux 33076, France. [22]Renal Transplantation Department, Assistance Publique–Hopitaux de Paris (AP-HP), Hopital Bichat, 46 Rue Henri Huchard, Paris 75018, France. [23]Department of Nephrology, Clermont-Ferrand University Hospital, Clermont-Ferrand, France. [24]Service de Nephrologie Et Immunologie Clinique, CHU de Nantes, 30 Boulevard Jean Monnet, 44093, Nantes Cedex 1, France. [25]Department of Nephrology and Organ Transplantation, CHU Rangueil, 1 Avenue J.Poulhes, TSA 50032, 31059, Toulouse Cedex 9, France. [26]Transplantation Renale, Hopital Necker, 145 Rue de Sevres, Paris 75015, France. [27]Service de Nephrologie, Hemodialyse, Aphereses Et Transplantation Renale, Hopital Michallon, Centre Hospitalier Universitaire Grenoble-Alpes, Boulevard de La Chantourne, La Tronche 38700, France. [28]Institute of Public Health, Charité. Universitätsmedizin Berlin, Luisenstrasse 57, Berlin 10117, Germany. [29]Amsterdam UMC, Vrije Universiteit, De Boelelaan 1112, Amsterdam 1081 HV, the Netherlands. [30]Service de Nephrologie, Dialyse Et Transplantation Renale, Hopital Nord, CHU de Saint-Etienne, 25 Boulevard Pasteur, 42055, Saint-Etienne Cedex 2, France. [31]Department of Nephrology-Dialysis-Transplantation, University of Liège, CHU Sart Tilman, Liège, Belgium. [32]Department of Nephrology-Dialysis-Apheresis, Hôpital Universitaire Carémeau, Nîmes, France.

Received: 2 November 2024 Accepted: 21 January 2025
Published online: 30 January 2025

References

1. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int. 2024;105: Suppl 4S:117–314.
2. Pottel H, Björk J, Courbebaisse M, Couzi L, Ebert N, Eriksen BO, Dalton RN, Dubourg L, Gaillard F, Garrouste C, Grubb A, Jacquemont L, Hansson M, Kamar N, Lamb EJ, Legendre C, Littmann K, Mariat C, Melsom T, Rostaing L, Rule AD, Schaeffner E, Sundin PO, Turner S, Bökenkamp A, Berg U, Åsling-Monemi K, Selistre L, Åkesson A, Larsson A, Nyman U, Delanaye P. Development and Validation of a Modified Full Age Spectrum Creatinine-Based Equation to Estimate Glomerular Filtration Rate : A Cross-sectional Analysis of Pooled Data. Ann Intern Med. 2021;174:183–91.
3. Delanaye P, Rule AD, Schaeffner E, Cavalier E, Shi J, Hoofnagle AN, Nyman U, Björk J, Pottel H. Performance of the European Kidney Function Consortium (EKFC) creatinine-based equation in United States cohorts. Kidney Int. 2024;105(3):629–37. https://doi.org/10.1016/j.kint.2023.11.024. Epub 2023 Dec 13 PMID: 38101514.
4. Delanaye P, Cavalier E, Stehlé T, Pottel H. Glomerular Filtration Rate Estimation in Adults : Myths and Promises. Nephron. 2024;148:408–14.
5. Burlacu A, Iftene A, Busoiu E, Cogean D, Covic A. Challenging the supremacy of evidence-based medicine through artificial intelligence: the time has come for a change of paradigms. Nephrol Dial Transplant. 2020;35:191–4.
6. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
7. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(837):845.
8. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34:3711–8.
9. Matheny ME, Whicher D, Thadaney IS. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. JAMA. 2020;323:509–10.
10. Glassock RJ. Artificial intelligence in medicine and nephrology: the hope, the hype, and reality. Clin Kidney J. 2024;17:sfae074.
11. Verma A, Chitalia VC, Waikar SS, Kolachalama VB. Machine Learning Applications in Nephrology: A Bibliometric Analysis Comparing Kidney Studies to Other Medicine Subspecialities. Kidney Med. 2021;3:762–7.
12. Liu X, Li N-S, Lv L-S, Huang J-H, Tang H, Chen J-X, et al. A comparison of the performances of an artificial neural network and a regression model for GFR estimation. Am J Kidney Dis. 2013;62:1109–15.
13. Wang H, Bowe B, Cui Z, Yang H, Swamidass SJ, Xie Y, et al. A Deep Learning Approach for the Estimation of Glomerular Filtration Rate. IEEE Trans Nanobiosci. 2022;21:560–9.
14. Jiang S, Li Y, Jiao Y, Zhang D, Wang Y, Li W. A back propagation neural network approach to estimate the glomerular filtration rate in an older population. BMC Geriatr. 2023;24(23):322.
15. Nakano FK, Åkesson A, de Boer J, Dedja K, D'hondt R, Haredasht FN, Björk J, Courbebaisse M, Couzi L, Ebert N, Eriksen BO, Dalton RN, Derain-Dubourg L, Gaillard F, Garrouste C, Grubb A, Jacquemont L, Hansson M, Kamar N, Legendre C, Littmann K, Mariat C, Melsom T, Rostaing L, Rule AD, Schaeffner E, Sundin PO, Bökenkamp A, Berg U, Åsling-Monemi K, Selistre L, Larsson A, Nyman U, Lanot A, Pottel H, Delanaye P, Vens C. Comparison between the EKFC-equation and machine learning models to predict Glomerular Filtration Rate. Sci Rep. 2024;14:26383.
16. Singh V, Asari VK, Rajasekaran R. A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. Diagnostics (Basel). 2022;12:116.
17. Salekin A, Stankovic J. Detection of chronic kidney disease and selecting important predictive attributes. In: Salekin A, editor. 2016 IEEE International Conference on Healthcare Informatics (ICHI). Piscataway: IEEE; 2016. p. 262–270.
18. Wang W, Chakraborty G, Chakraborty B. Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm. Appl Sci. 2020;11:202.
19. Schena FP, Anelli VW, Abbrescia DI, Di Noia T. Prediction of chronic kidney disease and its progression by artificial intelligence algorithms. J Nephrol. 2022;35:1953–71.
20. Fan Z, Yang Q, Xu Z, Sun K, Yang M, Yin R, Zhao D, Fan J, Ma H, Shen Y, Xia H. Construct a classification decision tree model to select the optimal equation for estimating glomerular filtration rate and estimate it more accurately. Sci Rep. 2022;12:14877.
21. Truchot A, Raynaud M, Kamar N, Naesens M, Legendre C, Delahousse M, Thaunat O, Buchler M, Crespo M, Linhares K, Orandi BJ, Akalin E, Pujol GS, Silva HT Jr, Gupta G, Segev DL, Jouven X, Bentall AJ, Stegall MD, Lefaucheur C, Aubert O, Loupy A. Machine learning does not outperform traditional statistical modelling for kidney allograft failure prediction. Kidney Int. 2023;103:936–48.
22. Levey AS et al. K/DOQI clinical practice guidelines for chronic kidney disease: Evaluation, classification, and stratification. Am J Kidney Dis. 2002;39:i-ii+S1-S266.

23. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO. Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney Int. 2024;2024(105):S117–314.
24. Pottel H, Rule AD, Ebert N, Eriksen BO, Dubourg L, Vidal-Petiot E, Grubb A, Hansson M, Lamb EJ, Littman K, Mariat C, Melsom T, Schaeffner E, Sundin PO, Akesson A, Larsson A, Cavalier E, Bukabau JB, Sumaili EK, Yayo E, Monnet D, Flamant M, Nyman U, Delanaye P. Cystatin C-based equation to estimate GFR without the insclusion of race and sex. New Engl J Med. 2023;388(333):343.
25. Carrero JJ, Fu EL, Sang Biostat Y, Ballew S, Evans M, Elinder CG, et al. Discordances Between Creatinine and Cystatin C-Based Estimated GFR and Adverse Clinical Outcomes in Routine Clinical Practice. Am J Kidney Dis. 2023;82(5):534–42.

Publisher's Note