1、Data preprocessing

The coding of binary variable *sex* was 0, 1. The continuous variables *age*, *BMI*, *ALB*, *UA*, and *HGB* were scaled to mean 0 and standard deviation 1, while *Scr*, *Scys* were log-transformed, and followed by linearly normalization to mean 0 and standard deviation 1. The dependent variable measure GFR was log transformed.

2、Sensitivity analysis of data splitting method

We performed sensitivity analysis using random data splitting method, that is, the development and internal validation datasets used for glomerular filtration rate (GFR) estimation models development were derived by randomly splitting the whole dataset rather than splitting the data based on time participant recruitment. The results from both methods are similar.

3、Details of recursive feature elimination

The recursive feature elimination procedure was implemented using scikit-learn’s RFECV method (scikit-learn version 0.22.2). The details of model perimeters are shown below:

|  |  |
| --- | --- |
| Model Perimeters | Value |
| estimator | RandomForest |
| step | 1 |
| cv | 5 |
| scoring | 'neg\_mean\_squared\_error' |

4、Details of random forest model of GFR estimation

The hyper perimeters were tuned using grid search method with 5-fold cross validation. Grid search is implemented by using scikit-learn’s GridsearchCV method (scikit-learn version 0.22.2). The details of model perimeters are shown below:

|  |  |
| --- | --- |
| Model Perimeters | Value |
| bootstrap | True |
| min\_sample\_split | np.arange(5,50,step=1) |
| n\_estimators | [500,1000] |
| max\_features | np.arange(3,20, step=1) |
| max\_depth | np.arange(2,10,step=1) |
| min\_samples\_leaf | np.arange(2,10,step=1) |

5、Model illustration of 4-variable revised CKD-EPI model (Age, Sex, Scr and Scys)

|  |  |  |  |
| --- | --- | --- | --- |
| Gender | CREAT | CYC | Model |
| Female | <1.6 | <0.6 |  |
| Female | <1.6 | >=0.6 |  |
| Female | >= 1.6 | <0.6 |  |
| Female | >= 1.6 | >=0.6 |  |
| Male | <1.6 | <0.6 |  |
| Male | <1.6 | >=0.6 |  |
| Male | >= 1.6 | <0.6 |  |
| Male | >= 1.6 | >=0.6 |  |

6、Model illustration of 9-variable spline regression (Age, Weight, BMI, BUN, UA, APOB, CYC, HCT, Scr and Scys.)

|  |  |  |
| --- | --- | --- |
| CREAT | CYC | Model |
| <2 | <0.8 |  |
| <2 | >=0.8 |  |
| >= 2 | <0.8 |  |
| >= 2 | >=0.8 |  |

7、Choice of optimal knots of piecewise spline regression

We used the same methodology to determine the optimal knots in revised CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) equation. The knot of serum creatinine (Scr) were 1.6, whereas the knot of serum cystatin C (Scys) was 0.8 mg/L. For the 9-variable model, the knot of Scr was three values which were 1.8, 2 and 2.2. We chose the median 2 as the effect of choosing these three knots was minor. The knot of Scr was 0.8. The box plots of root of mean squared error (RMSE) of 5-fold cross-validation was shown below.

A screenshot of a cell phone

Description automatically generated

Fig.1 Optimal knot 1.6 for CREAT of 4-variable revised CKD-EPI model

A screenshot of a cell phone

Description automatically generated

Fig.2 Optimal knot 0.6 for CYC of 4-variable revised CKD-EPI model

A screenshot of a cell phone

Description automatically generated

Fig.3 Optimal knot for Scr of 9-variable spline regression model. 3 values (1.8, 2, 2.2) show equivalent mean RMSE, we chose median 2.

A screenshot of a cell phone

Description automatically generated

Fig.4 Optimal knot 0.6 for CYC of 9-variable spline regression model.