

RESEARCH ARTICLE

# Can we predict when to start renal replacement therapy in patients with chronic kidney disease using 6 months of clinical data?

Min-Jeong Lee<sup>1,2</sup>, Joo-Han Park<sup>3</sup>, Yeo Rae Moon<sup>4</sup>, Soo-Yeon Jo<sup>5</sup>, Dukyong Yoon<sup>5</sup>, Rae Woong Park<sup>5</sup>, Jong Cheol Jeong<sup>1</sup>, Inwheel Park<sup>1</sup>, Gyu-Tae Shin<sup>1</sup>, Heungsoo Kim<sup>1\*</sup>

**1** Department of Nephrology, Ajou University School of Medicine, Suwon, Korea, **2** Department of Emergency Medicine, Ajou University School of Medicine, Suwon, Korea, **3** Department of Internal Medicine, Ajou University School of Medicine, Suwon, Korea, **4** Department of Biostatistics, Ajou University School of Medicine, Suwon, Korea, **5** Department of Medical Informatics, Ajou University School of Medicine, Suwon, Korea

These authors contributed equally to this work.

\* [nephrohs@ajou.ac.kr](mailto:nephrohs@ajou.ac.kr)



## Abstract

### Purpose

We aimed to develop a model of chronic kidney disease (CKD) progression for predicting the probability and time to progression from various CKD stages to renal replacement therapy (RRT), using 6 months of clinical data variables routinely measured at healthcare centers.

### Methods

Data were derived from the electronic medical records of Ajou University Hospital, Suwon, South Korea from October 1997 to September 2012. We included patients who were diagnosed with CKD (estimated glomerular filtration rate [eGFR] < 60 mL·min<sup>-1</sup>·1.73 m<sup>-2</sup> for ≥ 3 months) and followed up for at least 6 months. The study population was randomly divided into training and test sets.

### Results

We identified 4,509 patients who met reasonable diagnostic criteria. Patients were randomly divided into 2 groups, and after excluding patients with missing data, the training and test sets included 1,625 and 1,618 patients, respectively. The integral mean was the most powerful explanatory ( $R^2 = 0.404$ ) variable among the 8 modified values. Ten variables (age, sex, diabetes mellitus[DM], polycystic kidney disease[PKD], serum albumin, serum hemoglobin, serum phosphorus, serum potassium, eGFR (calculated by Chronic Kidney Disease Epidemiology Collaboration [CKD-EPI]), and urinary protein) were included in the final risk prediction model for CKD stage 3 ( $R^2 = 0.330$ ). Ten variables (age, sex, DM, GN, PKD, serum hemoglobin, serum blood urea nitrogen[BUN], serum calcium, eGFR(calculated by

## OPEN ACCESS

**Citation:** Lee M-J, Park J-H, Moon YR, Jo S-Y, Yoon D, Park RW, et al. (2018) Can we predict when to start renal replacement therapy in patients with chronic kidney disease using 6 months of clinical data? PLoS ONE 13(10): e0204586. <https://doi.org/10.1371/journal.pone.0204586>

**Editor:** Pasqual Barretti, Universidade Estadual Paulista Julio de Mesquita Filho, BRAZIL

**Received:** March 13, 2018

**Accepted:** September 11, 2018

**Published:** October 4, 2018

**Copyright:** © 2018 Lee et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** The authors received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

Modification of Diet in Renal Disease[MDRD]), and urinary protein) were included in the final risk prediction model for CKD stage 4 ( $R^2 = 0.386$ ). Four variables (serum hemoglobin, serum BUN, eGFR(calculated by MDRD) and urinary protein) were included in the final risk prediction model for CKD stage 5 ( $R^2 = 0.321$ ).

## Conclusion

We created a prediction model according to CKD stages by using integral means. Based on the results of the Brier score (BS) and Harrel's C statistics, we consider that our model has significant explanatory power to predict the probability and interval time to the initiation of RRT.

## Introduction

The incidences of chronic kidney disease (CKD) and end-stage renal disease (ESRD) have been increasing rapidly [1]. The overall prevalence of CKD was found to be 8.2% in South Korea according to a study published in 2016 [2], and most patients with CKD have concerns about starting dialysis or undergoing transplantation. However, accurate prediction of the progression of disease and the timing of renal replacement therapy (RRT) remain problematic because of the lack of an accepted predictive tool for CKD progression that is effective and precise. In clinical practice, it is common for physicians to perform prognostic evaluation of a patient's future disease progression based on a few recent measurements of glomerular filtration rate (GFR) or serum creatinine.

Therefore, physicians have difficulty in deciding which patients will ultimately progress to kidney failure and when they will require RRT. Identifying patients at risk of CKD progression may facilitate more optimal nephrology care. In the present study, we aimed to develop a model of CKD progression for predicting the probability and time to progression from CKD to RRT, using 6 months of clinical data variables routinely measured at healthcare centers. This developed model would provide more precise predictions than the commonly used Kidney Disease: Improving Global Outcomes (KDIGO) CKD stages, based eGFR and albuminuria.

## Material and methods

### Data source

The data were derived from the electronic medical record (EMR) database at Ajou University Hospital, Suwon, South Korea, from October 1997 to September 2012. This database contains information on patients and medical records, and includes data from all medical departments in the hospital. We extracted the data without personal identification to ensure patient confidentiality. The study was approved by the institutional review board of Ajou University Hospital.

### Study population

**Study set.** We included patients who were diagnosed with CKD and followed up for at least 6 months. The diagnostic criterion for CKD is estimated glomerular filtration rate (eGFR)  $< 60 \text{ mL} \cdot \text{min}^{-1} \cdot 1.73 \text{ m}^{-2}$  for  $\geq 3$  months [3]. The Modification of Diet in Renal Disease (MDRD) study equation or the Chronic Kidney Disease Epidemiology Collaboration

(CKD-EPI) equation was used to calculate eGFR; We used both equations and included the patients if even one of the two equations ( $eGFR < 60 \text{ mL}\cdot\text{min}^{-1}\cdot 1.73 \text{ m}^{-2}$ ) was satisfactory. We excluded patients who were  $< 19$  years old and those who had undergone RRT within 6 months of the study.

- MDRD equation

$$186 \times \text{serum creatinine}^{-1.154} \times \text{age}^{-0.203} \times 0.742 \text{ (if female)}$$

- CKD-EPI equation

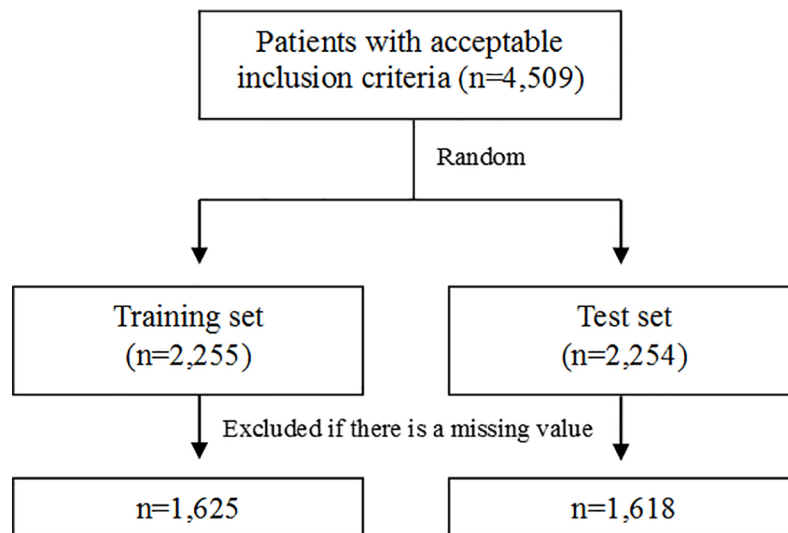
$$141 \times \min(\text{Scr}/\kappa, 1)^\alpha \times \max(\text{Scr}/\kappa, 1)^{-1.209} \times 0.993^{\text{Age}} \times 1.018 \text{ (if female)} \times 1.159 \text{ (if African)}$$

( $\kappa = 0.7$  if female,  $\kappa = 0.9$  if male,  $\alpha = -0.329$  if female,  $\alpha = -0.411$  if male, min = minimum Scr/ $\kappa$  or 1, max = maximum Scr/ $\kappa$  or 1)

**Training set and test set.** We randomly divided the final study population into a training set and a test set for the verification process (Fig 1).

### Observation period and study period

The observation period was defined as interval from the initial day of observation to the day of initiation of RRT or the day of censoring. The initial day is the first day on which the eGFR decreased to  $< 60 \text{ mL}\cdot\text{min}^{-1}\cdot 1.73 \text{ m}^{-2}$ . RRT included hemodialysis, peritoneal dialysis, and renal transplantation. The initial RRT point was defined as the first day of hemodialysis, day of catheter insertion for peritoneal dialysis, or the day of surgery for renal transplantation. If we could not identify the renal replacement event, we regarded the last follow-up date as the last observation day. The study period refers to the 180 days from the initial day of observation.



**Fig 1. Flow diagram for patient's selection.**

<https://doi.org/10.1371/journal.pone.0204586.g001>

## Variables

The variables were as follows: demographic variables, including age and sex; comorbid conditions, including diabetes mellitus (DM), hypertension (HTN), glomerular nephritis (GN), systemic lupus erythematosus (SLE), and polycystic kidney disease (PKD); laboratory variables, including levels of blood urea nitrogen (BUN), hemoglobin, serum creatinine, serum calcium, serum phosphate, serum albumin, serum bicarbonate, urinary creatinine, urinary protein, and urinary blood, eGFR by the MDRD, and eGFR by the CKD-EPI. We excluded urine albumin, urine hemoglobin, and urine creatinine levels as variables because they were not measured in more than 50% of the patients.

Data regarding laboratory examination and comorbidity variables were collected throughout the study period. For missing values, we included data for 30 days before the initial day of observation and 30 days after completion of the study period. Urinary protein by dipstick was reported semi-quantitatively as trace, 1+, 2+, 3+, or 4+ corresponding to albumin levels of 10, 30, 100, 300, or 1000mg/dl albumin respectively. Urinary protein level was coded as 5 dummy variables on the basis of negative values (trace, 1+, 2+, 3+, 4+). Criteria for the 5 comorbidities are described as follows.

1. DM: ICD-10 (E10–E14) code, serum HbA1c > 6.5%, or use of hypoglycemic medication
2. HTN: ICD-10 (I10–I15) code or use of antihypertensive medication
3. GN: ICD-10 (N01–N08) code
4. SLE: ICD-10 (M32) code
5. PCKD: ICD-10 (Q61) code

## Statistical analysis

**Development of representative value.** We developed 8 “modified values” that were potentially associated with CKD for 6 months and chose the “representative value” that demonstrated the greatest efficiency in a multivariate Cox proportional hazards regression model.

The modified values were: value at baseline, value at the end of the study period, minimum value, maximum value, ratio of the minimum to maximum values, slope of the minimum to maximum values, integral means, and slope of initial to integral means(details as follows).

1. The value at baseline: the value obtained closest to the initial day of observation ( $\pm 30$  days)
2. The value at the end of the study period: the value obtained closest to the end of the study period ( $\pm 30$  days)
3. The minimum value: the minimum value during the study period
4. The maximum value: the maximum value during the study period
5. The ratio of the minimum to maximum values: the maximum value/minimum value
6. The slope of the minimum to maximum values

$$\frac{\text{Maximum value} - \text{Minimum value}}{\text{Day (maximum value)} - \text{Day (minimum value)}}$$

7. The integral means

$$\sum_{i=1}^{n-1} \frac{(b_i + b_{i+1})(a_{i+1} - a_i)}{2(a_n - a_1)}$$

8. (n = number of values, i = order, a = day of value recording, b = value on that day)

The slope of initial to integral means

$$\frac{\text{The integral means} - \text{The value at baseline}}{\text{Day (maximum value)} - \text{Day (minimum value)} + 90 \text{ days}}$$

Values were excluded if 50% of the cases had missing data. Urinary protein (categorical variable) was only available at baseline.

**Model development.** Multivariate Cox proportional hazards regression was used for model development. We created a prediction model according to CKD stages [4]. The probability of the patient not undergoing RRT at time  $t$  (years) is as follows [5].

$$S(t) = S_0(t)^{\exp(\sum \beta_i \times X_i - \sum \beta_i \times \tilde{u}_i)}$$

$t$ : Followup time

$\beta_i$ : Regression coefficient

$X_i$ : Level of risk factor  $i$  of a patient

$\tilde{u}_i$ : Corresponding average value of population

$S_0(t)$ : Underlying probability of surviving

" $\sum \beta_i \times X_i - \sum \beta_i \times \tilde{u}_i$ " is defined as the risk index (RI): an increased value indicates a greater probability of RRT. We selected variables using clinical guidance and backward elimination (Wald) methods. The variables that did not contribute to the explanatory power of the RRT predictive model were removed until the remaining variables were significantly related to RRT ( $p < 0.05$ ).

**Evaluation of model performance.** To evaluate the expected prediction error of the training model, we calculated the Brier score (BS) [6] and Harrel's C statistics [7]. The BS is the square of deviation of the real value and the expected value. The higher the BS, the higher the expected error. If the BS is  $> 33\%$ , the expected data show random levels, and if the BS is close to 0%, the expected data show perfect prediction.

Harrel's C statistic is a common and well-validated measure to assess the discrimination. The higher the C-statistic, the better the model can discriminate between subjects who experience the outcome of interest and subjects who do not. C-statistics provide overall measures of predictive accuracy.

**Software.** We collected EMR data from Microsoft SQL Server 2012, and used PASW statistics (18.0.0) (SPSS Inc., Chicago, IL, USA) for selecting representative values. The multivariate Cox proportional hazards regression model, BS, and Harrel's C statistics were analysed using R package (3.4.3).

## Results

### Patient selection

We identified 4,509 patients who met reasonable diagnostic criteria. Patients were randomly divided into 2 groups, and after the exclusion of patients with missing values, the training and test sets included 1,625 and 1,618 patients, respectively (Table 1).

### Set description

Patients in the training and test sets were similar with regard to demographics, comorbidities, laboratory values, and outcomes, with the exception of eGFR (MDRD), eGFR (CKD-EPI), and

Table 1. Baseline characteristics of the patients.

| Characteristics   |                                  | No. (%) of patients         |                         | p value |
|-------------------|----------------------------------|-----------------------------|-------------------------|---------|
|                   |                                  | Training set<br>(n = 1,625) | Test set<br>(n = 1,618) |         |
| Demographics      |                                  |                             |                         |         |
|                   | Age (years)                      | 60 (19–90)                  | 59 (19–97)              | NS      |
|                   | Sex (male)                       | 728 (44.8)                  | 739 (45.7)              | NS      |
| Comorbidities     |                                  |                             |                         |         |
|                   | DM                               | 640 (39.4)                  | 626 (38.7)              | NS      |
|                   | HTN                              | 610 (37.5)                  | 562 (34.7)              | .096    |
|                   | GN                               | 299 (18.4)                  | 325 (20.1)              | NS      |
|                   | SLE                              | 8 (0.5)                     | 12 (0.7)                | NS      |
|                   | PKD                              | 28 (1.7)                    | 33 (2.0)                | NS      |
| Laboratory values |                                  |                             |                         |         |
|                   | Serum albumin* (g/dL)            | 4.00 (1.44–5.33)            | 4.00 (1.49–5.10)        | NS      |
|                   | Serum creatinine* (mg/dL)        | 1.59 (0.63–13.50)           | 1.55 (0.76–14.25)       | NS      |
|                   | Serum hemoglobin* (g/dL)         | 11.75 (5.30–20.28)          | 11.78 (5.44–17.60)      | NS      |
|                   | Serum bicarbonate* (mEq/L)       | 24.00 (11.50–33.57)         | 23.97 (11.56–32.00)     | NS      |
|                   | Serum BUN* (mg/dL)               | 23.40 (6.60–145.85)         | 23.19 (6.25–142.67)     | NS      |
|                   | Serum calcium* (mg/dL)           | 8.93 (5.95–12.03)           | 8.90 (4.90–12.50)       | NS      |
|                   | Serum phosphorus* (mg/dL)        | 3.63 (1.00–8.99)            | 3.64 (1.33–9.40)        | NS      |
|                   | Serum potassium* (mEq/L)         | 4.50 (2.41–6.73)            | 4.50 (2.81–7.60)        | NS      |
|                   | eGFR (MDRD)*                     | 39.89 (15.32)               | 41.07 (15.19)           | .027‡   |
|                   | eGFR (CKD-EPI)*                  | 39.41 (15.65)               | 40.60 (15.55)           | .031‡   |
|                   | Urine protein†                   |                             |                         | NS      |
|                   | Trace                            | 151 (9.3)                   | 162 (10.0)              |         |
|                   | 1 positive                       | 243 (15.0)                  | 236 (14.6)              |         |
|                   | 2 positive                       | 438 (27.0)                  | 402 (24.8)              |         |
|                   | 3 positive                       | 336 (20.7)                  | 317 (19.6)              |         |
|                   | 4 positive                       | 57 (3.5)                    | 58 (3.6)                |         |
| Outcome           |                                  |                             |                         |         |
|                   | Observation time                 | 1096 (182–5089)             | 1306 (1003)             | NS      |
|                   | Renal replacement therapy events | 530 (32.6)                  | 473 (29.2)              | .037‡   |

Note.  $\chi^2$ -test or Mann-Whitney U-test was used. Data are presented as number (%) or median (range)

\*Integral means

†initial value.

‡ $p < 0.05$ ,  $p$  value above 0.10 replaced with “NS”(not significant).

<https://doi.org/10.1371/journal.pone.0204586.t001>

RRT events. The eGFR (MDRD) and eGFR (CKD-EPI) were lower (39.9 mL·min<sup>-1</sup>·1.73 m<sup>-2</sup> vs. 41.1 mL·min<sup>-1</sup>·1.73 m<sup>-2</sup> and 39.4 mL·min<sup>-1</sup>·1.73 m<sup>-2</sup> vs. 40.6 mL·min<sup>-1</sup>·1.73 m<sup>-2</sup>,  $p < 0.05$ ) and RRT events were higher (530/1,625 vs. 473/1,618,  $p < 0.05$ ) in the training set than in the test set.

### Prediction model outcome

**Representative values.** We developed a multivariate Cox proportional hazards regression model with 8 modified values. We included 2,225 patients in the training set, and considered all collected variables. Eight modified values were all significantly effective, but the integral mean exhibited the most powerful explanatory value ( $R^2 = 0.404$ ), except for the end value

**Table 2. Outcomes of model development using different modified values of variables.**

| Modified values                   | R <sup>2</sup> | Patients (n) |
|-----------------------------------|----------------|--------------|
| Baseline value                    | 0.342          | 1,508        |
| End value                         | 0.546          | 685          |
| Maximum                           | 0.395          | 1,612        |
| Minimum                           | 0.373          | 1,612        |
| Ratio of minimum to maximum       | 0.260          | 1,612        |
| Integral mean                     | 0.404          | 1,612        |
| Slope of the minimum to maximum   | 0.244          | 1,168        |
| Slope of initial to integral mean | 0.298          | 1,508        |

<https://doi.org/10.1371/journal.pone.0204586.t002>

(R<sup>2</sup> = 0.546) (Table 2). We excluded the end value model because it included only 685 patients (< 50% of all patients). Thus, we used the integral mean as a representative value for each variable.

**Selection of prediction variables & outcome of the model.** Variable selection process underwent through the multivariate Cox proportional hazards regression model using the backward elimination method. We made 3 models for three separate analysis for CKD stage 3–5. The final model for CKD stage 3 that included 10 selected variables (age, sex DM, PKD, levels of serum albumin, serum hemoglobin, serum phosphorous, and serum potassium, eGFR [CKD-EPI], and urinary protein) had risk predictive power of approximately 33% in the Cox proportional hazards regression model (Table 3). The risk is greater in patients who are female or elderly and in those who have DM. The greater the levels of serum albumin and eGFR (CKD-EPI), the lower the risk; the greater the levels of serum phosphorus and urine protein, the higher the risk. The model for CKD stage 4 that included 10 selected variables (age, sex,

**Table 3. Regression coefficients and hazard ratios for variables in the risk prediction model at CKD stage 3 patients.**

|                   | Training set           |      |      |               |        | p value |
|-------------------|------------------------|------|------|---------------|--------|---------|
|                   | Regression coefficient | SE   | HR   | 95% CI for HR |        |         |
|                   |                        |      |      | Lower         | Higher |         |
| Age (years)       | -0.024***              | 0.06 | 0.98 | 0.97          | 0.99   | <0.001  |
| Sex (male)        | -0.424**               | 0.14 | 0.65 | 0.50          | 0.86   | 0.002   |
| DM                | 0.527***               | 0.14 | 1.69 | 1.28          | 2.24   | <0.001  |
| PKD               | 1.068                  | 0.62 | 2.91 | 0.89          | 9.56   | 0.079   |
| Serum albumin     | -0.976***              | 0.14 | 0.38 | 0.29          | 0.50   | <0.001  |
| Serum hemoglobin  | -0.078                 | 0.05 | 0.93 | 0.85          | 1.01   | 0.093   |
| Serum phosphorus  | 0.422***               | 0.10 | 1.53 | 1.24          | 1.87   | <0.001  |
| Serum potassium   | 0.217                  | 0.15 | 1.24 | 0.93          | 1.65   | NS      |
| eGFR(CKD-EPI)     | -0.050***              | 0.01 | 0.95 | 0.94          | 0.96   | <0.001  |
| Urine protein, 1+ | 1.085***               | 0.27 | 2.96 | 1.73          | 5.07   | <0.001  |
| Urine protein, 2+ | 1.253***               | 0.24 | 3.50 | 2.18          | 5.62   | <0.001  |
| Urine protein, 3+ | 1.289***               | 0.26 | 3.63 | 2.19          | 6.02   | <0.001  |
| Urine protein, 4+ | 1.520***               | 0.36 | 4.57 | 2.27          | 9.22   | <0.001  |

Note. R<sup>2</sup> = 0.330, p < 0.001.

HR = hazard ratio, CI = confidence interval.

\* p < 0.05

\*\* p < 0.01

\*\*\* p < 0.001, p value above 0.10 replaced with “NS”(not significant).

<https://doi.org/10.1371/journal.pone.0204586.t003>

Table 4. Regression coefficients and hazard ratios for variables in the risk prediction model at CKD stage 4 patients.

|                   | Training set           |      |       |               |        | p value |
|-------------------|------------------------|------|-------|---------------|--------|---------|
|                   | Regression coefficient | SE   | HR    | 95% CI for HR |        |         |
|                   |                        |      |       | Lower         | Higher |         |
| Age (years)       | -0.010                 | 0.01 | 0.99  | 0.98          | 1.00   | NS      |
| Sex (male)        | -0.433*                | 0.50 | 0.65  | 0.44          | 0.95   | 0.027   |
| DM                | 0.397                  | 0.21 | 1.49  | 0.98          | 2.25   | 0.059   |
| GN                | -0.318                 | 0.22 | 0.73  | 0.47          | 1.12   | NS      |
| PKD               | 1.786**                | 0.56 | 5.60  | 2.00          | 17.81  | 0.001   |
| Serum hemoglobin  | -0.109                 | 0.07 | 0.90  | 0.78          | 1.03   | NS      |
| Serum BUN         | -0.016                 | 0.01 | 0.98  | 0.97          | 1.00   | NS      |
| Serum calcium     | -0.318                 | 0.19 | 0.73  | 0.50          | 1.06   | NS      |
| eGFR(MDRD)        | -0.107***              | 0.02 | 0.90  | 0.87          | 0.93   | <0.001  |
| Urine protein, 1+ | 1.211**                | 0.42 | 3.36  | 1.46          | 7.70   | 0.004   |
| Urine protein, 2+ | 1.279**                | 0.39 | 3.59  | 1.67          | 7.72   | 0.001   |
| Urine protein, 3+ | 1.853***               | 0.41 | 6.38  | 2.88          | 14.16  | <0.001  |
| Urine protein, 4+ | 2.539***               | 0.56 | 12.67 | 4.23          | 37.89  | <0.001  |

Note. R<sup>2</sup> = 0.386, p<0.001.

HR = hazard ratio, CI = confidence interval.

\*p<0.05

\*\*p<0.01

\*\*\*p<0.001, p value above 0.10 replaced with "NS"(not significant).

<https://doi.org/10.1371/journal.pone.0204586.t004>

DM, GN, PKD, levels of serum haemoglobin, serum BUN, and serum calcium, eGFR [MDRD], and urinary protein) had risk predictive power of approximately 39% in the Cox proportional hazards regression model (Table 4). The risk is greater in patients who are female and in those who have PKD.

Table 5 shows the model for CKD stage 5 which had risk predictive power of approximately 32%. The model for CKD stage 5 included 4 selected variables, which was serum haemoglobin, serum BUN, eGFR[MDRD], and urinary protein.

Table 5. Regression coefficients and hazard ratios for variables in the risk prediction model at CKD stage 5 patients.

|                   | Training set           |      |      |               |        | p value |
|-------------------|------------------------|------|------|---------------|--------|---------|
|                   | Regression coefficient | SE   | HR   | 95% CI for HR |        |         |
|                   |                        |      |      | Lower         | Higher |         |
| Serum hemoglobin  | -0.100                 | 0.06 | 0.91 | 0.81          | 1.02   | 0.097   |
| Serum BUN         | 0.015                  | 0.01 | 1.02 | 1.00          | 1.03   | 0.013   |
| eGFR(MDRD)        | -0.070**               | 0.02 | 0.93 | 0.89          | 0.98   | 0.002   |
| Urine protein, 2+ | 0.669*                 | 0.29 | 1.95 | 1.11          | 3.43   | 0.019   |
| Urine protein, 3+ | 0.916**                | 0.41 | 2.50 | 1.40          | 4.47   | 0.002   |
| Urine protein, 4+ | 0.988                  | 0.56 | 2.69 | 0.90          | 8.06   | 0.078   |

Note. R<sup>2</sup> = 0.321, p<0.001.

HR = hazard ratio, CI = confidence interval.

\*p<0.05

\*\*p<0.01

\*\*\*p<0.001.

<https://doi.org/10.1371/journal.pone.0204586.t005>



**Risk prediction model.** The risk index (RI) of CKD stage 3 patients can be defined as follows.

$$\begin{aligned}
 \text{RI} &= -0.002 \times \text{age}(\text{years}) - 0.976 \times \text{albumin}(\text{g/dL}) - 0.078 \times \text{hemoglobin}(\text{g/dL}) \\
 &+ 0.422 \times \text{phosphorus}(\text{mg/dL}) + 0.217 \times \text{potassium}(\text{mEq/L}) - 0.050 \\
 &\times \text{eGFR}(\text{CKD EPI}) + 5.968 - 0.424 (\text{if female}) + 0.527(\text{if DM is present}) \\
 &+ 1.068(\text{if PKD is present}) + 1.085(\text{if urine protein} = 1+) \\
 &+ 1.253(\text{if urine protein} = 2+) + 1.289(\text{if urine protein} = 3+) \\
 &+ 1.520(\text{if urine protein} = 4+)
 \end{aligned}$$

The RI of CKD stage 4 patients can be defined as follows.

$$\begin{aligned}
 \text{RI} &= -0.002 \times \text{age}(\text{years}) - 0.976 \times \text{albumin}(\text{g/dL}) - 0.078 \times \text{hemoglobin}(\text{g/dL}) \\
 &+ 0.422 \times \text{phosphorus}(\text{mg/dL}) + 0.217 \times \text{potassium}(\text{mEq/L}) - 0.050 \\
 &\times \text{eGFR}(\text{CKD EPI}) + 5.968 - 0.424 (\text{if female}) + 0.527(\text{if DM is present}) \\
 &+ 1.068(\text{if PKD is present}) + 1.085(\text{if urine protein} = 1+) \\
 &+ 1.253(\text{if urine protein} = 2+) + 1.289(\text{if urine protein} = 3+) \\
 &+ 1.520(\text{if urine protein} = 4+)
 \end{aligned}$$

The RI of CKD stage 5 patients can be defined as follows.

$$\begin{aligned}
 \text{RI} &= -0.098 \times \text{hemoglobin}(\text{g/dL}) + 0.015 \times \text{BUN}(\text{mg/dL}) - 0.070 \times \text{eGFR}(\text{MDRD}) \\
 &+ 1.039 + 0.669(\text{if urine protein} = 2+) + 0.916(\text{if urine protein} = 3+) \\
 &+ 0.988(\text{if urine protein} = 4+)
 \end{aligned}$$

By using RI, the formula for the probability of a patient not undergoing RRT at some point ( $t$ , years) is as follows.

$$S(t) = S_0(t)^{\exp(RI)}$$

$S(t)$ : probability of not undergoing renal replacement therapy

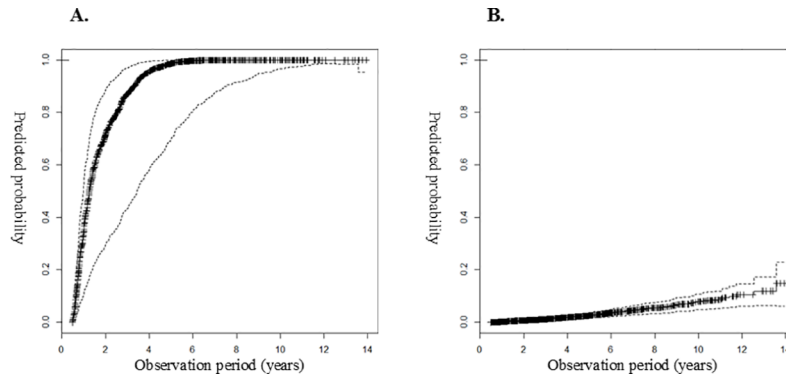
$S_0(t)$ : underlying probability

$(S_0(1) = 0.973, S_0(3) = 0.876, S_0(5) = 0.756, S_0(7) = 0.756, S_0(10) = 0.423)$

## Test set

**Brier score.** To evaluate the expected prediction error of the training set model, we calculated the weighted BS that gave the weighted value to censored data. The period during which the BS is  $< 0.33$  is approximately 5,000 days at the model of CKD stage 3 and 5. The period during which the BS  $< 0.33$  is approximately 4,000 days at the model of CKD stage 4. Thus, the prediction model gives a marginal predictive result up to approximately 4,000–5,000 days (S1–S3 Figs).

**Harrel's C statistics.** To evaluate the accuracy of the prediction model, we calculated Harrel's C statistics. The C-statistics of final model was 0.86 (0.83–0.88) at CKD stage 3, 0.80 (0.76–0.84) at CKD stage 4, 0.84 (0.78–0.90), respectively.



**Fig 2. Predicted probability of starting renal replacement therapy.** A) Patient with RRT after 5 years of follow-up (age = 56 years, sex = female, DM = no, PKD = yes, albumin = 4.06g/dl<sup>a</sup>, hemoglobin = 8.01g/dl<sup>a</sup>, calcium = 8.31mg/dl<sup>a</sup>, phosphorus = 3.16mg/dl<sup>a</sup>, potassium = 4.95mmol/L<sup>a</sup>, eGFR = 18.36, protein = 2+). B) Patient without RRT censored after 5 years of follow-up (age = 58 years, sex = male, DM = yes, PKD = no, albumin = 4.65g/dl<sup>a</sup>, hemoglobin = 12.11g/dl<sup>a</sup>, calcium = 9.70mg/dl<sup>a</sup>, phosphorus = 3.10mg/dl<sup>a</sup>, potassium = 4.02mmol/L<sup>a</sup>, eGFR = 54.81, protein = negative). (<sup>a</sup>integral mean value) Side lines are 95% confidence interval (CI).

<https://doi.org/10.1371/journal.pone.0204586.g002>

**Example cases of prediction model application.** We analysed 2 cases in which the observation period was approximately 5 years, using the risk prediction model in the test set. Fig 2A shows the graph for the probability of the event for a 56-year-old female patient who experienced progression to RRT after 5 years. The probability of the event was > 80% at 3 years and > 95% at 5 years. Fig 2B shows the probability of the event in a 58-year-old male patient who did not experience progression to RRT after 5 years. The probability of an event was < 20% at 10 years.

## Discussion

CKD is asymptomatic in the early stages, but symptoms appear in the later stages, accompanied by complications such as cardiovascular disease, anemia, infection, cognitive impairment, and impaired physical function [8–11]. The KDIGO clinical practice guideline suggested a prognostic classification system for CKD divided on the basis of 6 categories of GFR, 3 categories of albuminuria stage, and cause of disease. Based on these findings, KDIGO devised 3 broad risk categories based upon the likelihood of developing future kidney and cardiovascular complications [12]. However, eGFR assessment and ascertainment of albuminuria may not be sufficient for risk prediction in the clinic.

We considered many variables cited in previous articles that could affect renal function, including age, sex, laboratory findings, and comorbidities, to develop a risk prediction model. These included variables such as young age, male sex, African-American ethnicity, DM, HTN, obesity, urine protein, serum albumin, anemia, lipidemia, smoking, and cardiovascular disease [13]. In the Reduction of Endpoints in NIDDM with the Angiotensin II Antagonist Losartan (RENAAL) study, albuminuria, hypoalbuminemia, increased serum creatinine, and decreased hemoglobin were the risk factors associated with ESRD in patients with type 2 DM and nephropathy [14]. We collected data on the above variables, and identified data that were not measured in > 50% of the patients. Our study was performed retrospectively in order to identify missing variables that could significantly affect RRT.

We identified variables that were associated with RRT through the clinical guidance and backward elimination (Wald) methods. From a clinical point of view, models based on referral eGFR are more useful than an overall model. Predictions for patient with an eGFR of 60 would probably only be interesting to the patient, while predictions for a patient with an eGFR

15 are critical for dialysis preparation. So we underwent separate analysis according to CKD stages: 1) CKD stage 3: age, sex, DM, PKD, levels of serum albumin, serum hemoglobin, serum phosphate, and serum potassium, eGFR, and urinary protein. 2) CKD stage 4: age, sex, DM, GN, PKD, levels of serum hemoglobin, serum BUN, serum calcium, eGFR, and urinary protein. 3) CKD stage 5: level of serum hemoglobin, serum BUN, eGFR, and urinary protein. The results were similar to those of previous studies. First, one study reported that the risk of progression to ESRD was decreased among older patients with CKD stage 3 (hazard ratio [HR], 0.75; 95% confidence interval, 0.63–0.89 for each 10-year increase in age) [15]. Second, another study showed that male patients with CKD stage 4 and 5 had a shorter time to RRT than did female patients [16]. Third, it is thought that DM is rapidly becoming the most common cause of ESRD and is also associated with an increasing risk of ESRD [17]. In the African American Study of Kidney Disease and Hypertension (AASK) trial, the change in urinary protein level from baseline to 6 months predicted progression to RRT [18]. In the RENAAL study, baseline hemoglobin was an important independent variable for prediction of ESRD among diabetic patients [19]. Moreover, HTN has been found to be predictive of ESRD risk in several large population-based studies [17, 20]. However, the presence of HTN was not an independent predictor of kidney failure events in the present study. The RENAAL study showed similar findings, a result likely due to the fact that blood pressure was well controlled in the study patients [14].

To identify representative values that show renal function change over 6 months, we considered 8 modified values and developed a multivariate Cox proportional hazards regression model. The integral mean contains the time and the value in order to obtain sufficient power to explain the change in data over 6 months. The end value had the highest  $R^2$ , but the number of patients was inadequate to evaluate the model. We will compare the integral mean and end value in a larger dataset in a further study.

Finally, we developed the renal prediction model with several variables using integral means from continuous variables. To evaluate prediction error, we calculated the BS and the Harrel's C statistics. From the results of the BS and Harrel's C statistics, we consider that our model has sufficient explanatory power to predict renal progression.

The strength of our analysis is that we divided patients into 2 groups: the training set and the test set. Thus, we calculated the BS and Harrel's C statistics in order to confirm the accuracy of the model. Second, the prediction equation must include variables that are very routinely available in the nephrology clinic for convenience of use. Local healthcare facilities can collect laboratory data easily and integrate the risk prediction tool into decision-making for patients who require further evaluation or in preparation for RRT.

The limitations of our analysis are that the study was performed retrospectively, and therefore, the data obtained are insufficient including blood pressure measurements of the patients, which was important predictor in previous studies [17, 20]. We considered many variables from previous studies while developing the risk prediction tool, but insufficient data were available for evaluation from the EMR. Second, patients with missing data were excluded. Since missing data are usually selectively missing this causes a selection bias. Third, all of our study subjects were Asian, especially Korean, there is a limitation about applicability of the results in other occidental countries. Fourth, there is no standard procedure for determining the initiation of RRT; therefore, initiation of therapy may reflect personal opinions, and patients' economic, social, and environmental factors may also affect the timing. However, the selection of the test set and the training set from the same hospital in the present study meant that the prediction error was reduced because the characteristics of patients in the training set and test set were similar. Fifth, our study is a lack of renal diagnosis. Because we collected data not from accurate chart review but from the EMR, we have not been able to present the

primary cause of ESRD. Instead, we considered comorbidities with high prevalence as the primary cause of ESRD in Korea [21].

Many studies have identified a wide range of risk factors for the progression of CKD. Although many studies have identified similar risk factors, there has not been sufficient research performed on the risk prediction models for RRT. To develop accurate and easy-to-use models, further large prospective studies are required. Our predictive model for CKD may have sufficient power to predict RRT, as shown in 2 cases in the present study. However, there are also cases that did not fit the model. If data were collected from a greater number of patients with greater accuracy, a more precise model could be developed. The development of the representative value seems very complicated for a prediction tool in a clinical setting. To simplify this model could be achieved by cooperation among nephrologists and statisticians.

In summary, a model was developed and validated to predict the risk for ESRD. This model uses commonly available clinical variables and may provide more precise predictions than the commonly used KDIGO CKD stages, based on eGFR and albuminuria.

## Supporting information

**S1 Fig. Prediction error according to the observation period (days) at CKD stage 3.**  
(EMF)

**S2 Fig. Prediction error according to the observation period (days) at CKD stage 4.**  
(EMF)

**S3 Fig. Prediction error according to the observation period (days) at CKD stage 5.**  
(EMF)

**S1 Table. Modeling dataset.**  
(CSV)

**S2 Table. Validation dataset.**  
(CSV)

## Acknowledgments

The authors alone are responsible for the content and writing of the paper. Disclosure: M.J. Lee, None; J.H. Park, None; Y.R. Moon, None; S.Y. Jo, None; D.Y. Yoon, None; R.W. Park, None; J.C. Jeong, None; I. Park, None; G.Y. Shin, None; H. Kim, None.

## Author Contributions

**Conceptualization:** Min-Jeong Lee, Joo-Han Park, Jong Cheol Jeong, Heungsoo Kim.

**Data curation:** Min-Jeong Lee, Joo-Han Park, Soo-Yeon Jo, Dukyong Yoon, Rae Woong Park, Jong Cheol Jeong, Inwheel Park, Gyu-Tae Shin, Heungsoo Kim.

**Formal analysis:** Min-Jeong Lee, Joo-Han Park.

**Investigation:** Min-Jeong Lee, Joo-Han Park, Soo-Yeon Jo, Dukyong Yoon, Rae Woong Park, Gyu-Tae Shin, Heungsoo Kim.

**Methodology:** Min-Jeong Lee, Joo-Han Park, Soo-Yeon Jo, Dukyong Yoon, Rae Woong Park, Heungsoo Kim.

**Resources:** Jong Cheol Jeong, Inwheel Park, Gyu-Tae Shin, Heungsoo Kim.

**Software:** Soo-Yeon Jo, Dukyong Yoon, Rae Woong Park.

**Supervision:** Min-Jeong Lee, Joo-Han Park, Jong Cheol Jeong, Inwhee Park, Gyu-Tae Shin, Heungsoo Kim.

**Validation:** Min-Jeong Lee, Joo-Han Park, Yeo Rae Moon, Heungsoo Kim.

**Writing – original draft:** Min-Jeong Lee, Joo-Han Park, Jong Cheol Jeong, Heungsoo Kim.

**Writing – review & editing:** Min-Jeong Lee, Heungsoo Kim.

## References

1. Ishani A, Grandits GA, Grimm RH, Svendsen KH, Collins AJ, Prineas RJ, et al. Association of single measurements of dipstick proteinuria, estimated glomerular filtration rate, and hematocrit with 25-year incidence of end-stage renal disease in the multiple risk factor intervention trial. *Journal of the American Society of Nephrology: JASN*. 2006; 17(5):1444–52. Epub 2006/04/14. <https://doi.org/10.1681/ASN.2005091012> PMID: 16611715.
2. Park JI, Baek H, Jung HH. Prevalence of Chronic Kidney Disease in Korea: the Korean National Health and Nutritional Examination Survey 2011–2013. *J Korean Med Sci*. 2016; 31(6):915–23. <https://doi.org/10.3346/jkms.2016.31.6.915> PMID: 27247501; PubMed Central PMCID: PMC4853671.
3. Levey AS, de Jong PE, Coresh J, El Nahas M, Astor BC, Matsushita K, et al. The definition, classification, and prognosis of chronic kidney disease: a KDIGO Controversies Conference report. *Kidney international*. 2011; 80(1):17–28. Epub 2010/12/15. <https://doi.org/10.1038/ki.2010.483> PMID: 21150873.
4. Levey AS, Atkins R, Coresh J, Cohen EP, Collins AJ, Eckardt KU, et al. Chronic kidney disease as a global public health problem: approaches and initiatives—a position statement from Kidney Disease Improving Global Outcomes. *Kidney international*. 2007; 72(3):247–59. <https://doi.org/10.1038/sj.ki.5002343> PMID: 17568785.
5. Soveri I, Holme I, Holdaas H, Budde K, Jardine AG, Fellstrom B. A cardiovascular risk calculator for renal transplant recipients. *Transplantation*. 2012; 94(1):57–62. Epub 2012/06/12. <https://doi.org/10.1097/TP.0b013e3182516cdc> PMID: 22683851.
6. Gerds TA, Schumacher M. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*. 2006; 48(6):1029–40. <https://doi.org/10.1002/bimj.200610301> PMID: 17240660
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010; 21(1):128–38. <https://doi.org/10.1097/EDE.0b013e3181c30fb2> PMID: 20010215; PubMed Central PMCID: PMC3575184.
8. Fink JC, Brown J, Hsu VD, Seliger SL, Walker L, Zhan M. CKD as an underrecognized threat to patient safety. *American journal of kidney diseases: the official journal of the National Kidney Foundation*. 2009; 53(4):681–8. Epub 2009/02/28. <https://doi.org/10.1053/j.ajkd.2008.12.016> PMID: 19246142; PubMed Central PMCID: PMC3710448.
9. James MT, Quan H, Tonelli M, Manns BJ, Faris P, Laupland KB, et al. CKD and risk of hospitalization and death with pneumonia. *American journal of kidney diseases: the official journal of the National Kidney Foundation*. 2009; 54(1):24–32. Epub 2009/05/19. <https://doi.org/10.1053/j.ajkd.2009.04.005> PMID: 19447535.
10. Hailpern SM, Melamed ML, Cohen HW, Hostetter TH. Moderate chronic kidney disease and cognitive function in adults 20 to 59 years of age: Third National Health and Nutrition Examination Survey (NHANES III). *Journal of the American Society of Nephrology: JASN*. 2007; 18(7):2205–13. Epub 2007/06/08. <https://doi.org/10.1681/ASN.2006101165> PMID: 17554148.
11. Wilhelm-Leen ER, Hall YN, M KT, Chertow GM. Frailty and chronic kidney disease: the Third National Health and Nutrition Examination Survey. *The American journal of medicine*. 2009; 122(7):664–71 e2. Epub 2009/06/30. <https://doi.org/10.1016/j.amjmed.2009.01.026> PMID: 19559169.
12. Boeglin RJ, Caprioli J, Zulauf M. Long-term fluctuation of the visual field in glaucoma. *Am J Ophthalmol*. 1992; 113(4):396–400. Epub 1992/04/15. PMID: 1558113.
13. Taal MW, Brenner BM. Renal risk scores: progress and prospects. *Kidney international*. 2008; 73(11):1216–9. Epub 2008/03/07. <https://doi.org/10.1038/ki.2008.36> PMID: 18322541.
14. Keane WF, Zhang Z, Lyle PA, Cooper ME, de Zeeuw D, Grunfeld JP, et al. Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: the RENAAL study. *Clinical journal of the American Society of Nephrology: CJASN*. 2006; 1(4):761–7. Epub 2007/08/21. <https://doi.org/10.2215/CJN.01381005> PMID: 17699284.

15. Eriksen BO, Ingebretsen OC. The progression of chronic kidney disease: a 10-year population-based study of the effects of gender and age. *Kidney international*. 2006; 69(2):375–82. Epub 2006/01/13. <https://doi.org/10.1038/sj.ki.5000058> PMID: 16408129.
16. Evans M, Fryzek JP, Elinder CG, Cohen SS, McLaughlin JK, Nyren O, et al. The natural history of chronic renal failure: results from an unselected, population-based, inception cohort in Sweden. *American journal of kidney diseases: the official journal of the National Kidney Foundation*. 2005; 46(5):863–70. Epub 2005/10/29. <https://doi.org/10.1053/j.ajkd.2005.07.040> PMID: 16253726.
17. Haroun MK. Risk Factors for Chronic Kidney Disease: A Prospective Study of 23,534 Men and Women in Washington County, Maryland. *Journal of the American Society of Nephrology*. 2003; 14(11):2934–41. <https://doi.org/10.1097/01.asn.0000095249.99803.85> PMID: 14569104
18. Garway-Heath DF, Poinoosawmy D, Fitzke FW, Hitchings RA. Mapping the visual field to the optic disc in normal tension glaucoma eyes. *Ophthalmology*. 2000; 107(10):1809–15. PMID: 11013178.
19. Malik R, Swanson WH, Garway-Heath DF. 'Structure-function relationship' in glaucoma: past thinking and current concepts. *Clin Exp Ophthalmol*. 2012; 40(4):369–80. <https://doi.org/10.1111/j.1442-9071.2012.02770.x> PMID: 22339936; PubMed Central PMCID: PMC3693944.
20. Iseki K Fau—Iseki C, Iseki C Fau—Ikemiya Y, Ikemiya Y Fau—Fukuyama K, Fukuyama K. Risk of developing end-stage renal disease in a cohort of mass screening. *Kidney international*. 1996; 49:800–5. PMID: 8648923
21. Jin DC, Yun SR, Lee SW, Han SW, Kim W, Park J. Current characteristics of dialysis therapy in Korea: 2015 registry data focusing on elderly patients. *Kidney Res Clin Pract*. 2016; 35(4):204–11. <https://doi.org/10.1016/j.krcp.2016.09.006> PMID: 27957414; PubMed Central PMCID: PMC5142391.