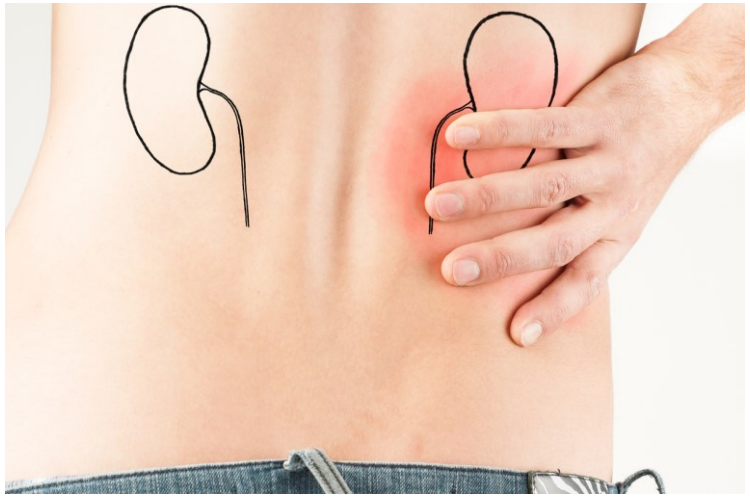
Predicting chronic kidney disease

Yeo Sing Chen

25th August 2021

Chronic kidney disease (CKD)

The sustained reduction of kidney function



Prevalence (2019) = 13.4%

End-stage therapy cost = $5-7 million

Early detection of CKD can improve clinical and economic outcomes

Targeted therapy

Limited resource

Drug safety

Financial relief

Symptoms of CKD

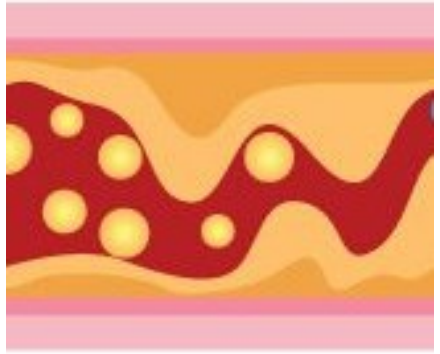
High blood pressure

High creatinine

Low hemoglobin

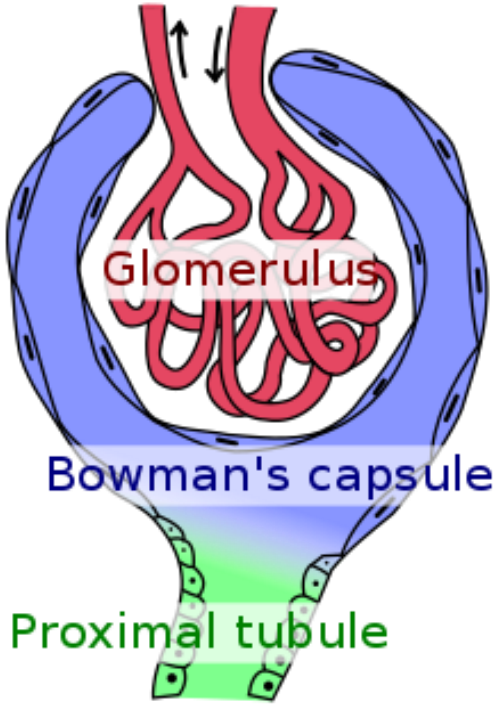
High blood glucose

High LDL-c



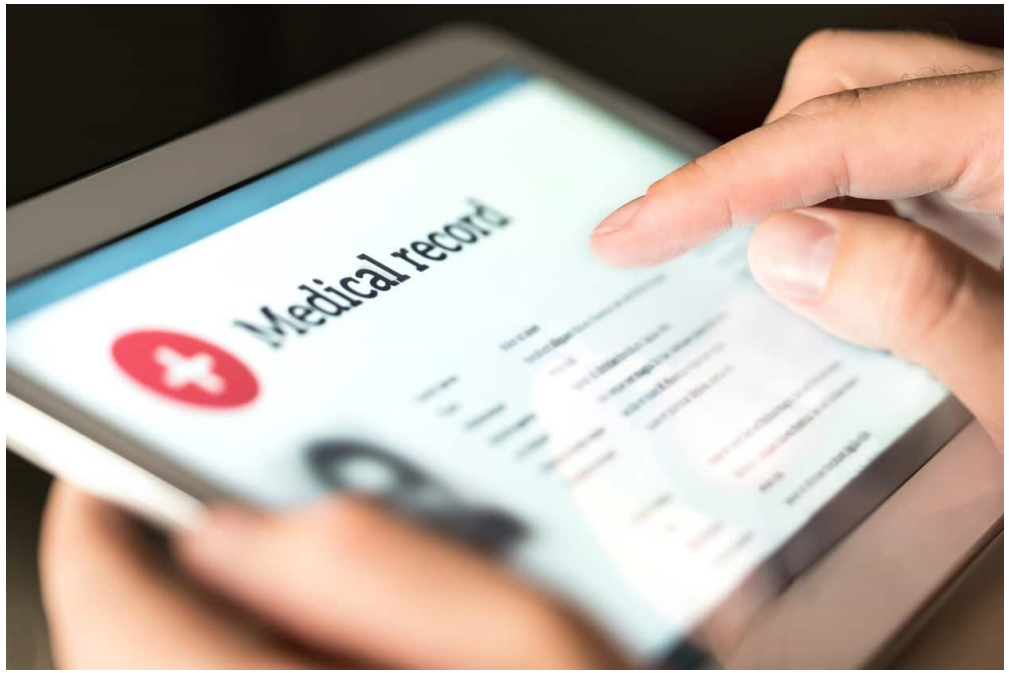
Estimated Glomerular Filtration Rate (eGFR)

- Glomerular filtration rate (GFR) estimates how much blood passes through the glomeruli each minute
- eGFR is calculated based on gender, race, age, and serum creatinine
- Current detection of kidney disease depends on routine reporting of whether eGFR <60 ml/min/1.73 m^2



Objectives

To build models that predict onset of CKD in immediate and far future for patients with at least 2-year of medical record

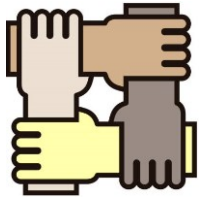


Data provided

Gender

Age

Race

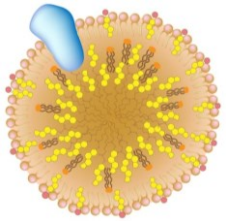


Creatinine

Glucose

Blood pressure

LDL-c

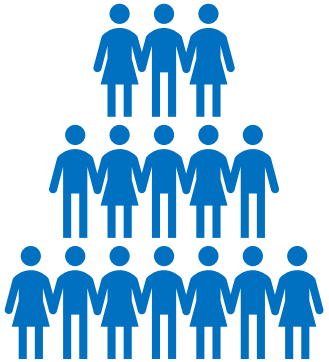


Hemoglobin

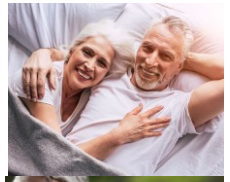
Medications

Progress of CKD

Demographics



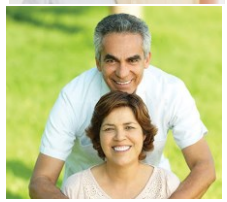
n = 300



75.3%

8.0%

5.7%



2.3%

8.7%

58.7%

70.4 ± 9.2 years
(mean ± SD, at baseline)

Longitudinal medical records

- Time points are record in day, with baseline as 0
- All measurements available at day 0
- Irregular time points for collecting data
- All data ended at day 699, except for hemoglobin
- Time points for drugs are in range (start day and end day provided)
- Bin time points by every 180-day
- Analyze data based on first 720 days (~2 years)

Progress of CKD

- Binary (True/False)
- 100 subjects developed CKD (one-third of sample)
- Assume to be diagnosed by the last available data point

Processing data

Missing values

- Race
  - Convert “Unknown” to the mode (“White”)
- Continuous measuremets:
  - KNNImputer from sklearn.impute
  - Imputed separately for each 180-day bin

Pre-processing

- Numerical features
  - Standardization
- Categorical features
  - OneHotEncoder

One hot encoding

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

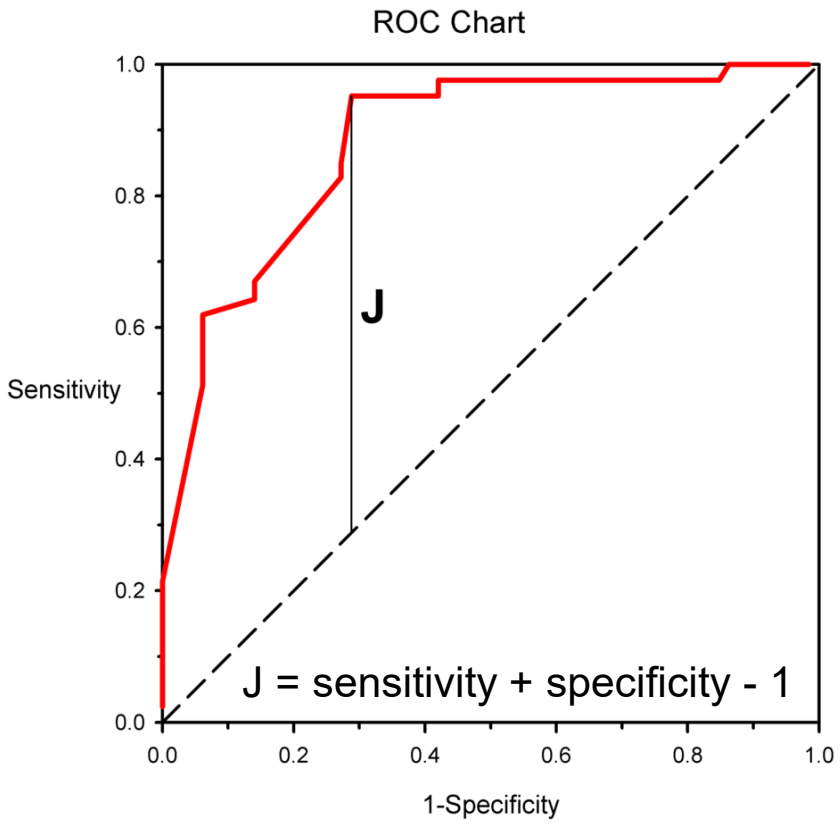
Comparison between patients with CKD and non-CKD at the 4th 180-day bin

	non-CKD	CKD	t/X^2	*P*
Gender				
Female	129 (65.8)	42 (43.8)	12.0	<0.001
Male	67 (34.2)	54 (56.2)		
Age in years	71.0 ± 9.0	69.2 ± 9.8	1.5	0.130
Race				
White	163 (83.2)	82 (85.4)	0.9	0.819
Black	17 (8.7)	6 (6.2)		
Asian	12 (6.1)	5 (5.2)		
Hispanic	4 (2.0)	3 (3.1)		
Physiological readings				
Systolic blood pressure	131.5 ± 12.6	136.1 ± 15.2	-2.6	0.011
Disatolic blood pressure	78.7 ± 8.1	80.7 ± 10.3	-1.7	0.097
Creatinine	1.4 ± 0.3	1.3 ± 0.4	0.7	0.503
Glucose	6.5 ± 1.1	7.1 ± 1.9	-3.0	0.004
Low-density lipoprotein	80.6 ± 23.0	89.8 ± 27.7	-2.8	0.006
Hemoglobin	13.9 ± 1.4	13.4 ± 1.5	2.3	0.022

- Gender and ethnicity are expressed in n (%)
- Age and physiological readings are expressed in Mean ± SD

Machine learning

- Data split
  - 80% training, 20% testing
  - Stratify by progress of CKD
- Applied GridSearchCV to obtain optimal algorithm (F1 as scorer)
- Area under the ROC (AUROC) as the main evaluation metric for final models
- Implemented Youden's J statistic to define cut-off point for optimal probability thresholds



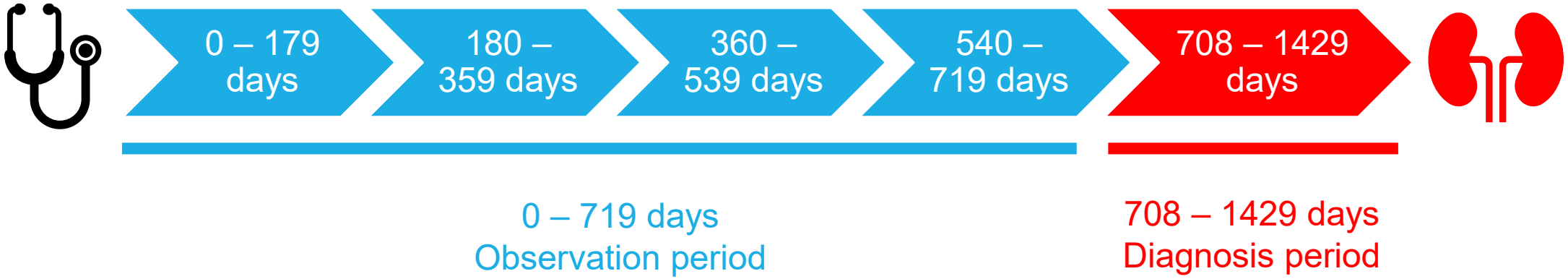
Algorithms

- Logistic
- LightGBM (LGBM)
- Decision tree
- Random forest

- Ensemble
  - Soft voting
    - Estimators = optimal algorithms (Logistic, LightGBM, Decision tree, Random forest)
  - StackingClassifier
    - Final estimator = Logistic / LightGBM / Random forest

180-day bin data

- Bin every 180 days (~6 months)
- One model for each 180-day bin
- For each subject, within the 180-day bin…
  - Compute median for the continuous measurements (BP, creatinine, glucose, HGB, LDL)
  - Sum the drug prescribed (daily dosage X total day)



Aggregated data

- Discard temporal component
- For each subject…
  - Compute mean of all 180-day bins
  - Sum the drug prescribed (daily dosage X total day)

id	SBP	DBP	Creatinine	Glucose	HGB	LDL	Drug 1	Drug 2	…	Drug 21
0	120	90	1	5	15	110	10000	15000	...	14000

Temporal data

- For each feature, enter four 180-day bin of data
- For each subject, within the 180-day bin…
  - Compute median for the continuous measurements (BP, creatinine, glucose, HGB, LDL)
  - Sum the drug prescribed (daily dosage X total day)

id	SBP_1	SBP_2	SBP_3	SBP_4	DBP_1	…	Drug 1_1	Drug 1_2	…	Drug 21_4
0	120	118	119	115	90	...	10000	11000	...	13800

Prediction results of continuous values

Day / data structure	Single algorithm	AUROC	Ensemble algorithm	AUROC
0-179	Logistic	0.582	Stacking (logistic)	0.550
180-359	Random forest	0.699	Voting	0.682
360-539	Decision tree	0.737	Voting	0.722
540-719	LGBM	0.547	Stacking (random forest)	0.561
Aggregated	Random forest	0.691	Voting	0.599
Temporal	Random forest	0.781	Stacking (logistic)	0.742

Note: Stacking (final estimator)

Improving predictions: categorization

- Categorized continuous physiological measurements based on clinical definition
  - Low (1), Normal (2), and High (3)
- Categorized drug dosage using 75th percentile of total dosage as threshold
  - Drug not taken (0), Low (1), High (2)
  - Compute separately for each bin in the 180-day bin data
- Applied LabelEncoder to encode gender and race
- Binned age by tertile split (1 < 2 < 3)

Categorization of physiological measurements

Measurement	Low (1)	Normal (2)	High (3)
Creatinine (mg/dL)			
Men	<0.74	0.74 – 1.35	>1.35
Women	<0.59	0.59 – 1.04	>1.04
Blood pressure (mmHg)			
Systolic	<90	90 – 120	>120
Diastolic	<60	60 – 80	>80
Glucose (mmol/L)	<3.9	3.9 – 7.8	>7.8
HGB (g/dL)			
Men	<14	14 – 17.5	>17.5
Women	<12.3	12.3 – 15.3	>15.3
LDL (mg/dL)	<100	100 – 129	>129

Prediction results of categorized values

Day / data structure	Single algorithm	AUROC	Ensemble algorithm	AUROC
0-179	Decision tree	0.558	Stacking (random forest)	0.608*
180-359	Logistic	0.624	Voting	0.628
360-539	Logistic	0.589	Voting	0.592
540-719	Random forest	0.691*	Stacking (logistic)	0.683*
Aggregated	Logistic	0.606	Stacking (LGBM)	0.616*
Temporal	Random forest	0.681	Stacking (logistic)	0.691

* Indicates better than non-categorize
Note: Stacking (final estimator)

Group drugs by treatments

Drug	Treatment
Canagliflozin	Glucose
Dapagliflozin	Glucose
Metformin	Glucose
Atenolol	Blood pressure
Bisoprolol	Blood pressure
Carvedilol	Blood pressure
Labetalol	Blood pressure
Losartan	Blood pressure
Metoprolol	Blood pressure
Nebivolol	Blood pressure
Propranolol	Blood pressure
Irbesartan	Blood pressure
Olmesartan	Blood pressure
Telmisartan	Blood pressure
Valsartan	Blood pressure
Atorvastatin	Cholesterol
Lovastatin	Cholesterol
Pitavastatin	Cholesterol
Pravastatin	Cholesterol
Rosuvastatin	Cholesterol
Simvastatin	Cholesterol

- Group all drugs into 3 treatments and sum the total daily dosage
- Categorize drug dosage based on 75th percentile threshold (0 = not taken, 1 = normal, 2 = high)
- Sum the code to indicate the severity for treatment (higher value indicates higher dosage)

Further categorizing drug by treatment improved predictions

Day / data structure	Single algorithm	AUROC	Ensemble algorithm	AUROC
0-179	LGBM	0.626*	Stacking (random forest)	0.608*
180-359	Random forest	0.738*	Voting	0.695*
360-539	Logistic	0.592	Stacking (logistic)	0.596
540-719	Logistic	0.708*	Stacking (logistic)	0.689*
Aggregated	Decision tree	0.648	Stacking (random forest)	0.688*
Temporal	Logistic	0.819*	Stacking (LGBM)	0.788*

* Indicates better than non-categorize
Note: Stacking (final estimator)

Building models based on eGFR only

- $eGFR = 141 \times \min(S_{cr}/\kappa, 1)^{\alpha} \times \max(S_{cr}/\kappa, 1)^{-1.209} \times 0.993^{Age} \times 1.018$ [if female] $\times 1.159$ [if black]

  S_{cr} is serum creatinine in mg/dL
  κ is 0.7 for females and 0.9 for males
  α is -0.329 for females and -0.411 for males
  min indicates the minimum of S_{cr}/κ or 1
  max indicates the maximum of S_{cr}/κ or 1

- 24 models built
  - 12 Continuous eGFR
  - 12 Binary eGFR
    - <60 is kidney disease (1), ≥60 is normal (2)

Low prediction power with only eGFR as feature

- The AUROC for all 24 models were around 0.5 – 0.6
- All AUROC were lower than models built with more features
- Might be caused by the similar serum creatinine level between CKD and non-CKD patients

Summary

- Built models that can predict onset of CKD after gathering medical records at different time lengths
- Different data processing and algorithms were required to obtain best outcomes for each purpose
- Retaining temporal component improved prediction power

Best models

Day / data structure	Data processing	Algorithm	AUROC
0-179	Categorization*	LGBM	0.626
180-359	Categorization*	Random forest	0.738
360-539	Standardization & One-hot-encoding	Decision tree	0.737
540-719	Categorization	Random forest	0.691
Aggregated	Standardization & One-hot-encoding	Random forest	0.691
Temporal	Categorization*	Logistic	0.819

*Drug categorized by treatment

Future directions

- Collect more data to build deep learning models (e.g., CNN, BLSTM)
- Combine eGFR with other physiological measurements and drug dosage
- Compute “days to diagnosis” by counting days from last time point of record to build models that can predict onset of CKD in different time length
- When aggregate data, try other measurements such as mean or standard deviation
- For temporal data, compute difference between day-bin
- Investigate the influence of drugs on different physiological measurements
- Other tuning
  - Hyperparameters
  - Other algorithms (naïve Bayes, KNN, SVM)
  - Adjust age at different time points

References

- Levin & Stevens, Early detection of CKD: the benefits, limitations and effects on prognosis. 2011; Nat. Rev. Nephrol. 7, 446–457. doi:10.1038/nrneph.2011.86
- Lv JC, Zhang LX. Prevalence and Disease Burden of Chronic Kidney Disease. Adv Exp Med Biol. 2019;1165:3-15. doi: 10.1007/978-981-13-8871-2_1
- Massy ZA, de Zeeuw D. LDL cholesterol in CKD--to treat or not to treat? Kidney Int. 2013 Sep;84(3):451-6. doi: 10.1038/ki.2013.181. Epub 2013 May 22.
- Krishnamurthy et al., Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan. Healthcare 2021, 9, 546. doi: 10.3390/healthcare9050546
- Levey et al., A New Equation to Estimate Glomerular Filtration Rate. Ann Intern Med 2009, *150*(9), 604–612. doi: 10.7326/0003-4819-150-9-200905050-00006

Thank you for your attention