Predicting chronic kidney disease

YEO SING CHEN

2021

Chronic kidney disease (CKD)

- The global estimated prevalence of CKD is 13.4%
- Main diagnosis is by estimated Glomerular Filtration Rate (eGFR) less than 60 ml/min per 1.73 m^2
- Patients with end-stage kidney disease (ESKD) needing renal replacement therapy is estimated between 4.902 and 7.083 million

Symptoms and causes of CKD

- High blood pressure is one of the major causes and effects of chronic kidney disease
- Doctors determine the stage of kidney disease using the GFR, which includes serum creatinine in the formula. The kidneys normally remove creatinine from the blood that comes from muscle activity. As kidney function slows down, creatinine levels go up
- Kidneys make a hormone called erythropoietin (EPO), which is important for the production of red blood cells. Kidney disease will cause low EPO, which lead to low red blood cell and anemia
- Insulin resistance is an early metabolic alteration in CKD patients, being apparent when the GFR is still within the normal range and becoming almost universal in those who reach the end stage of kidney failure
- In the majority of patients with CKD, the low-density lipoprotein (LDL) cholesterol are usually normal. Reduction in LDL cholesterol in patients with CKD is associated with proportional reduction of cardiovascular risk

Demographics

- Sample size = 300 patients
- Race: 75.3% White, 8.0 % Black, 5.7% Asian, 2.3% Hispanic, 8.7% Unknown
- Gender: 58.7% Female
- Age (mean ± SD): 70.4 ± 9.2 years

Data provided

- Demographic (gender, age, race)
- Serum creatinine in mg/dL
- Glucose level in mmol/L
- Diastolic and systolic blood pressure in mmHg
- Hemoglobin (HGB) level in g/dL
- Low-density lipoprotein (LDL-c) level in mg/dL
- 21 medications prescribed in mg
  - Glucose control (Canagliflozin, Dapagliflozin, Metformin)
  - Blood pressure control (Atenolol, Bisoprolol, Carvedilol, Labetalol, Losartan, Metoprolol, Nebivolol, Propranolol, Irbesartan, Olmesartan, Telmisartan, Valsartan)
  - Cholesterol control (Atorvastatin, Lovastatin, Pitavastatin, Pravastatin, Rosuvastatin, Simvastatin)

Time points

- Measured 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 point by 180-day bin
- Analyze data based on first 720 days (~2 years)

Outcome

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

Objectives

To build models that predict onset of CKD for patient…

- With medical record of different time length (180-day bin data)
- With at least 2 years of medical record

Processing data

Missing values

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

Preparing for machine learning

- Numerical features
  - Standardization
- Categorical features
  - OneHotEncoder

Compare between stage_progress group

- Compare at the 4th 180-day bin (last available data)

	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
Ethnicity				
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

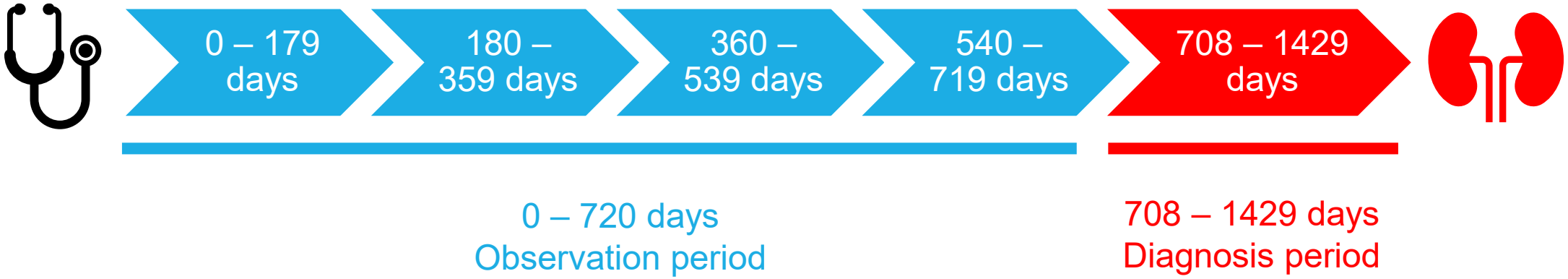
- Gender and ethnicity are expressed in n (%)
- Age and physiological readings are expressed in mean ± sd

Machine learning

- Data split
  - 80% training, 20% testing
  - Stratify by outcome (stage_progress)
- Area under the ROC (AUROC) curve as the main evaluation metric
- Implemented Youden's J statistic to define cut-off point for optimal probability thresholds
- Applied GridSearchCV to obtain optimal algorithm
- Algorithms
  - Logistic, LightGBM, Decision tree, Random forest
- Ensemble
  - Soft voting and StackingClassifier
  - Estimators = optimal algorithms (Logistic, LightGBM, Decision tree, Random forest)
  - Final estimator for StackingClassifier = 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



Aggregated data

- Discard temporal component
- For each subject…
  - Compute mean of the median values in each 180-day bin
  - 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 “raw” values

Day / data structure	Single algorithm	AUROC	Ensemble algorithm	AUROC
0-179	Logistic	0.582	Stacking (logistic)	0.544
180-359	Random forest	0.694	Voting	0.676
360-539	Decision tree	0.737	Stacking (logistic)	0.719
540-719	Random forest	0.518	Stacking (random forest)	0.591
Aggregated	Random forest	0.700	Stacking (random forest)	0.658
Temporal	Random forest	0.784	Stacking (random forest)	0.782

Improving predictions

- Categorized continuous measurements into low (1), normal (2), and high (3)
- Binarized drug dosage based on 75^{th} percentile as threshold: low (1), high (2)
  - 0 means drug was not taken
  - 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)

Feature categorization

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

Note: (2) is between low and high (inclusive)

Prediction results of categorized values

Day / data structure	Single algorithm	AUROC	Ensemble algorithm	AUROC
0-179	Decision tree	0.558	Stacking (LGBM)	0.630*
180-359	Logistic	0.624	Voting	0.640
360-539	Logistic	0.589	Voting	0.591
540-719	Random forest	0.688*	Stacking (logistic)	0.670*
Aggregated	Logistic	0.606	Stacking (random forest)	0.648
Temporal	Random forest	0.718	Stacking (logistic)	0.721

* Indicates better than non-categorize

Group drug by treatment

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

- Binarize drug dosage based on 75th percentile threshold (1 = normal, 2 = high)
- Sum the binary 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)	0.689*
180-359	Random forest	0.723*	Voting	0.700*
360-539	Logistic	0.592	Stacking (logistic)	0.600
540-719	Logistic	0.708*	Stacking (logistic)	0.700*
Aggregated	Decision tree	0.648	Stacking (random forest)	0.698*
Temporal	Logistic	0.820*	Stacking (logistic)	0.807*

All AUROC larger than categorizing all features

* Indicates better than non-categorize

Glomerular Filtration Rate (GFR)

- Sum of the filtration rates of functioning nephrons (filtering units of kidney)
- CKD-EPI equation
  - $GFR = 141 \times \min(S_{cr}/\kappa, 1)^{\alpha} \times \max(S_{cr}/\kappa, 1)^{-1.209} \times 0.993^{Age} \times 1.018\ [\text{if female}] \times 1.159\ [\text{if black}]$
- Binarize: <60 is kidney disease (1), ≥60 is normal (2)
- GFR includes race, gender, age, and one physiological measurement
- Is it possible to obtain better predictions based on GFR only?
  - Models showed low predicting power if only using GFR to predict CKD
  - The AUROC was around 0.5 for all models

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

Summary

- Built models that can predict onset of CKD after gathering medical records at different time lengths
- Different data processing and algorithms required for different purposes
- Retaining temporal component improved prediction power

Best models

Day / data structure	Data processing	Algorithm	AUROC
0-179	Categorization	Stacking (random forest)	0.689
180-359	Categorization	Random forest	0.723
360-539	Standardization & One-hot-encoding	Decision tree	0.737
540-719	Categorization	Logistic	0.708
Aggregated	Standardization & One-hot-encoding	Random forest	0.700
Temporal	Categorization	Logistic	0.820

Future directions

- Collect more data to build deep learning model
- Combine GFR 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

- 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