RMJ Ramathibodi Medical Journal

Rama Med J | Original Article

Comparison of Machine Learning With Logistic Regression for Prediction of Chronic Kidney Disease in the Thai Adult Population

Ratchainant Thammasudjarit[1], Punnathorn Ingsathit[1, 2], Sigit Ari Saputro[1, 3], Atiporn Ingsathit[1], Ammarin Thakkinstian[1]

[1] Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[2] Triam Udom Suksa School, Bangkok, Thailand

[3] Division of Biostatistics and Health Informatics, Faculty of Public Health, Airlangga University, Surabaya, Indonesia

Background: Chronic kidney disease (CKD) takes huge amounts of resources for treatments. Early detection of patients by risk prediction model should be useful in identifying risk patients and providing early treatments.

Objective: To compare the performance of traditional logistic regression with machine learning (ML) in predicting the risk of CKD in Thai population.

Methods: This study used Thai Screening and Early Evaluation of Kidney Disease (SEEK) data. Seventeen features were firstly considered in constructing prediction models using logistic regression and 4 MLs (Random Forest, Naïve Bayes, Decision Tree, and Neural Network). Data were split into train and test data with a ratio of 70:30. Performances of the model were assessed by estimating recall, C statistics, accuracy, F1, and precision.

Results: Seven out of 17 features were included in the prediction models. A logistic regression model could well discriminate CKD from non-CKD patients with the C statistics of 0.79 and 0.78 in the train and test data. The Neural Network performed best among ML followed by a Random Forest, Naïve Bayes, and a Decision Tree with the corresponding C statistics of 0.82, 0.80, 0.78, and 0.77 in training data set. Performance of these corresponding models in testing data decreased about 5%, 3%, 1%, and 2% relative to the logistic model by 2%.

Conclusions: Risk prediction model of CKD constructed by the logistic regression, Neural Network, and Random Forest have comprehensible discrimination performance, but the logistic regression tends to have lower overfitting compared to Neural Network, and Random Forest.

Keywords: Chronic kidney disease, Machine learning, Clinical prediction model

Rama Med J: doi:10.33165/rmj.2021.44.4.250334

Received: June 24, 2021 **Revised:** October 15, 2021 **Accepted:** November 19, 2021

Corresponding Author:
Ratchainant Thammasudjarit
Department of Clinical Epidemiology and Biostatistics,
Faculty of Medicine Ramathibodi Hospital,
Mahidol University,
270 Rama VI Road, Ratchathewi,
Bangkok 10400, Thailand.
Telephone: +66 2201 1269
E-mail: ratchainant.tha@mahidol.edu

Introduction

Chronic kidney disease (CKD) is a major global health problem which was the 18th rank in reducing disability-adjusted life year (DALY),[1] and increasing the risk of cardiovascular disease.[2] CKD is defined as the presence of an abnormality in kidney structure or function persistently for 3 months or longer,[3] in which end stage renal disease (ESRD) is the worst severe stage that happens when the kidneys' function is less than 15%. However, most CKD patients are asymptomatic even with a severe stage ESRD, so they are often under recognized.[4]

In Thailand, the prevalence of CKD was as high as 17.5%, which was about 7 million Thai adult patients.[5] Moreover, the number of the ESRD patients who received renal replacement therapy (RRT) in Thailand increased from only 419 patients per million in 2007 to 2274 patients per million in 2019.

Because of the fact that most CKD patients may be asymptomatic, there is often a delay in recognition, diagnosis, and treatment allocations. Early identification of individuals with CKD should be encouraged and targeted in order to implement intervention strategies to delay CKD progression, such as lifestyle modifications (low-protein dietary changes, exercise, and education) and health monitoring programs for blood pressure, blood sugar control, lipid lowering, education, etc.[6]

Therefore, many risk prediction models[7-12] have been constructed including a simplified risk prediction score of CKD for Thai population.[13] These risk prediction models were developed mainly using traditional statistical models (logistic regression or Cox regression), but only a few of them had applied machine learning (ML) methods (Decision Tree, Support Vector Machine, Neural Network, etc).[14-17] The ML is claimed to be better than the traditional statistical models[18, 19] in dealing with nonlinear relationship among features and outcome, complex interactions among features and many potential predictive features, etc.

Therefore, we question if ML methods could be better in prediction of CKD risk in Thai population? This study was conducted to prove this hypothesis by applying ML methods and traditional logistic regression using a large-scale nationwide survey database in Thailand. Model performances of MLs and traditional logistic regression models were compared. Four ML methods including Decision Tree, Naïve Bayes, Random Forest, and Neural Network have different advantages. Decision Tree has intuitive interpretation compared to others but easy to get overfit. Naïve Bayes consumes noticeably short training time but relies on conditional independence assumption. Random Forest leverages ensemble learning to reduce overfit in Decision Tree. Neural Network is powerful but requires larger training data compared to other models. Logistic regression fits to clinical interpretation but weak against nonlinearly separable problem.

This study aimed to compare the performance of traditional logistic regression with ML in predicting the risk of CKD in Thai population using a large-scale nationwide survey database in Thailand.

Methods

Data Source

This study used data from a community-based cross-sectional Thai Screening and Early Evaluation of Kidney Disease (SEEK) study, which was a nationwide survey of CKD prevalence conducted between August 2007 and June 2008.[5]

Briefly, the study included all adult subjects aged 18 years or older, who had no menstruation period for at least a week prior to the examination date if women, and who were willing participants of the study and provided signed consent forms. For sampling method, 4 regions of Thailand (Northern, Northeastern, Central, and Southern) and Bangkok Metropolitan were treated as strata. Stratified-cluster random sampling was applied to selected subjects. At the first stage, 2 to 3 provinces in each region were randomly selected. Each selected province was next classified as either an urban or rural areas, and then one district from each area was randomly selected. There were, in total, 10 provinces

including Bangkok, Chon Buri, Lop Buri, Phayao, Phrae, Sakon Nakhon, Nong Bua Lam Phu, Maha Sarakham, Phuket, and Songkhla, and then 20 districts were chosen for study sampling. The dataset contained all 3459 volunteer participants and 408 variables in which for this study's data usage, we will use the 17 valid variables appropriate to our study. The personal information of each participant was removed for privacy protection.

Ethics

Our research has been approved by Ethical Committee referred to the document number COA MURA2021/188 on March 05, 2021.

Study Features

The study features were included physical examinations (weight, height, waist and hip circumference, blood pressure, and respiratory rate). Blood tests were consisted of serum blood sugar, total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TRIG), uric acid (UA), serum creatinine (SCr), complete blood count in part of hemoglobin (Hb) and white blood cell (WBC) count. Urine test (microprotein, macroprotein, and urinary analysis) was involved.

Subject classification with diabetes, hypertension, and high cholesterol was based on history, relevant medicines used, blood tests, and physical examinations. Subjects were classified as having diabetes if they had one of the following criteria: self-reported diabetes, taking oral hypoglycemic agents, or fasting plasma glucose levels of 126 mg/dL or higher. Subjects were classified as having hypertension if they were told by doctors, taking antihypertensive drug(s) or had systolic blood pressure of 140 mmHg or higher, or diastolic blood pressure of 90 mmHg of higher. Subjects were classified as having high cholesterol if they were told by doctors or taking cholesterol-lowering drug(s). History of kidney stones was measured by self-reporting kidney stones. Serum uric and waist/hip ratio were categorized into 3 groups according to the tertile distributions.

Outcome Measurements

The estimated glomerular filtration rate (eGFR) was then calculated using the Modification of Diet in Renal Disease (MDRD) equation for isotope dilution mass spectrometry (IDMS) traceable serum creatinine values as follows: eGFR (mL/min/1.73 m^2) = 175 × $(SCr)^{-1.154}$ × $(Age)^{-0.203}$ × (0.742 if female). CKD was defined as stage 1 and 2 if eGFR of 90 mL/min/1.73 m^2 or higher, and eGFR from 60 to 89 mL/min/1.73 m^2 with hematuria and/or albumin-creatinine ratio of 30 mg/g or greater, stage 3, 4, and 5 if the eGFR was 30 to 59, 15 to 29, or less than 15 mL/min/1.73 m^2, respectively. The CKD stages 1 to 5 were combined and then compared with non-CKD in the whole analyses.

Statistical Analysis

Data were described using mean and standard deviation (SD) or median and interquartile range (IQR) where appropriate for continuous variables, and percentage for categorical variables. These corresponding data were compared between CKD and non-CKD groups using one-way analysis of variance, *t* test and χ^2 test, and Fisher exact test where appropriated.

To construct prediction models, data was randomly split into training data and testing data using a 70:30 ratio. Number of training samples and test samples were 2421 and 1038, respectively. Univariate analysis was performed to determine association between each feature and CKD using the simple logistic regression. Features with *P* value less than .10 were used to construct traditional logistic regression and 4 MLs including Decision Tree, Naïve Bayes, Random Forest, and Neural Network. Model specification and configuration were set (Table 1 in Supplement). Features were selected based on ranking χ^2 test, Gini index, and information gain. Evaluation metrics for each model were estimated including accuracy, recall, precision, F1, and C statistic. The scales of all evaluation metrics ranged from 0 (incorrect at all) to 1 (perfect). All analyses were performed using Orange Data Mining Software version 3.28 (Biolab open-source)[9] and STATA version 16.0 (StataCorp. Version 16. College Edition, TX: StataCorp LLC; 2019).

Results

A total of 3459 subjects were included into analyses. Most subjects were aged 40 years or older, 54.64% were females, and 65.02% had body mass index (BMI) less than 25 kg/m^2. Among them, 30.40% of subjects had waist/hip ratio less than 0.81, more than 50% of them had taken exercise, and 85.31% had physical activities. The majority of subjects had never smoked (63.69%), and some of them were currently alcohol drinkers (60.51%). Prevalence rates of diabetes, hypertension, and high cholesterol were frequent, 12.58%, 27.49%, and 52.56 %, respectively. However, a history of kidney stones was quite low with only 5.19%. Interestingly, history of taking nonsteroidal anti-inflammatory drug (NSAID), and traditional medicines were as high as 45.59% and 33.20%, respectively. The CKD prevalence was 18.13%, with 95% confidence interval (CI) of 16.86% to 19.45% (Table 1).

Seventeen features were assessed if they were associated with CKD in the univariate analysis (Table 1). Among them, 15 features had *P* value of less than .10 and so they were simultaneously included in a multivariate logistic model. Model selection by likelihood ratio test was done indicating only 7 significant features (age group, alcohol use, diabetes, hypertension, kidney stones, serum uric acid, and history of traditional medicine use) maintained in the final model. Subjects aged 60 to 69 years and 70 years or older were about 3.92 and 7.22 times higher odds of having CKD than subjects aged younger than 40 years. Use of traditional medicine and alcohol drinking were about 1.26 and 1.42 times higher odds of CKD than never used traditional medicine and non-alcohol drinking, respectively. Furthermore, subjects with history of diabetes, hypertension, kidney stone, and high uric acid had the odds of CKD about 1.81 to 2.67 times than subjects who did not have any of these conditions (Table 2).

Table 1. Patient Characteristics According to CKD And Non-CKD Groups

Characteristic	No. (%)			*P* Value*
	Total (N = 3459)	Non-CKD (n = 2832)	CKD (n = 627)	
Age, y				
< 40	1325 (38.31)	1223 (43.19)	102 (16.27)	< .001
40 - 59	1464 (42.32)	1226 (43.29)	238 (37.96)	
60 - 69	403 (11.65)	255 (9.00)	148 (23.60)	
≥ 70	267 (7.72)	128 (4.52)	139 (22.17)	
Sex				
Male	1569 (45.36)	1299 (45.87)	270 (43.06)	.20
Female	1890 (54.64)	1533 (54.13)	357 (56.94)	
Education				
Primary	1985 (57.67)	1546 (54.88)	439 (70.24)	.001
Secondary	975 (28.33)	868 (30.81)	107 (17.12)	
Diploma	148 (4.30)	132 (4.69)	16 (2.56)	
Bachelor	194 (5.64)	175 (6.21)	19 (3.04)	
Master	12 (0.35)	12 (0.43)	0 (0.00)	
None	128 (3.71)	84 (2.98)	44 (7.04)	
BMI, kg/m^2				
< 25	2249 (65.02)	1879 (66.35)	370 (59.01)	.002
25 - 29.9	924 (26.71)	733 (25.88)	191 (30.46)	
≥ 30	286 (8.27)	220 (7.77)	66 (10.53)	

Table 1. Patient Characteristics According to CKD And Non-CKD Groups (Continued)

Characteristic	No. (%)			*P* Value*
	Total (N = 3459)	Non-CKD (n = 2832)	CKD (n = 627)	
Waist/hip ratio				
< 0.81	1051 (30.40)	926 (32.71)	125 (19.94)	< .001
0.81 - 0.86	1152 (33.31)	969 (34.23)	183 (29.18)	
> 0.86	1255 (36.29)	936 (33.06)	319 (50.88)	
Smoking				
No	2194 (63.69)	1802 (63.88)	392 (62.82)	.62
Yes	1251 (36.31)	1019 (36.12)	232 (37.18)	
Alcohol				
No	1360 (39.49)	1060 (37.62)	300 (47.92)	< .001
Yes	2084 (60.51)	1758 (62.38)	326 (52.08)	
Exercise				
No	1390 (40.46)	1148 (40.77)	242 (39.03)	< .001
Mild	356 (10.36)	250 (8.88)	106 (17.10)	
Moderate	467 (13.59)	350 (12.43)	117 (18.87)	
Severe	1223 (35.59)	1068 (37.92)	155 (25.00)	
Work				
No	1296 (37.99)	999 (35.81)	297 (47.83)	< .001
Yes	2115 (62.01)	1791 (64.19)	324 (52.17)	
Physical activity				
No	502 (14.69)	396 (14.14)	106 (17.15)	< 0.001
Mild	151 (4.42)	90 (3.22)	61 (9.87)	
Moderate	194 (5.68)	135 (4.82)	59 (9.55)	
Severe	2571 (75.21)	2,179 (77.82)	392 (63.43)	
Diabetes				
No	3024 (87.42)	2580 (91.10)	444 (70.81)	< .001
Yes	435 (12.58)	252 (8.90)	183 (29.19)	
Hypertension				
No	2508 (72.51)	2207 (77.93)	301 (48.01)	< .001
Yes	951 (27.49)	625 (22.07)	326 (51.99)	
High cholesterol				
No	1641 (47.44)	1395 (49.26)	436 (39.23)	< .001
Yes	1818 (52.56)	1437 (50.74)	381 (60.77)	
Kidney Stone				
No	3085 (94.81)	2569 (96.43)	516 (87.46)	< .001
Yes	169 (5.19)	95 (3.57)	74 (12.54)	
Serum uric acid, mg/dL				
< 4.40	1064 (30.76)	935 (33.02)	129 (20.57)	< .001
4.40 - 5.61	1126 (32.55)	959 (33.86)	167 (26.63)	
> 5.61	1269 (36.69)	938 (33.12)	331 (52.80)	

Table 1. Patient Characteristics According to CKD And Non-CKD Groups (Continued)

Characteristic	No. (%)			*P* Value*
	Total (N = 3459)	Non-CKD (n = 2832)	CKD (n = 627)	
NSAID				
No	1882 (54.41)	1564 (55.23)	368 (50.72)	.04
Yes	1577 (45.59)	1268 (44.77)	309 (49.28)	
Traditional medicine				
No	2300 (66.80)	1938 (68.77)	362 (57.92)	< .001
Yes	1143 (33.20)	880 (31.23)	263 (42.08)	

Abbreviations: BMI, body mass index; CKD, chronic kidney disease; NSAID, nonsteroidal antiinflammatory drug.

* Significance threshold, $P < .005$.

Table 2. Factors Associated With CKD by Multiple Logistic Regression Analysis

Feature	OR (95% CI)	SE	z	*P* Value*
Age, y				
< 40	1.00 [Reference]	NA	NA	NA
40 - 59	1.57 (1.19 - 2.06)	0.22	3.20	.001
60 - 69	3.92 (2.83 - 5.43)	0.65	8.23	< .001
≥ 70	7.22 (5.10 - 10.41)	1.35	10.59	< .001
Alcohol	1.42 (1.15 - 1.66)	0.08	-3.24	.001
Diabetes	2.66 (2.05 - 3.45)	0.35	7.37	< .001
Hypertension	1.81 (1.45 - 2.26)	0.21	5.22	< .001
Kidney stone	2.67 (1.63 - 4.38)	0.67	3.90	< .001
Serum uric acid, mg/dL				
< 4.40	1.00 [Reference]	NA	NA	NA
4.40 - 5.61	1.30 (0.98 - 1.72)	0.19	1.79	.07
> 5.61	2.23 (1.71 - 2.93)	0.31	5.83	< .001
Traditional medicine	1.26 (1.02 - 1.56)	0.14	2.13	.03

Abbreviations: CI, confidence interval; CKD, chronic kidney disease; NA, not applicable; OR, odds ratio; SE, standard error, z, Z statistics from Wald test.

*Significance threshold, $P < .005$.

Performances of the predictive models were estimated for training and testing data. For logistic regression model, the C statistics, and F1 score in the training data were respectively 0.79 and 0.81, whereas these corresponding performances reduced to 0.77 and 0.80 in the test data. Among ML models in training data, the Neural Network yielded highest discriminative performance followed by a Random Forest, Naïve Baye, and Decision Tree with the corresponding C statistics of 0.82, 0.80, 0.78, and 0.77. However, performances in testing data of these corresponding ML models decreased about 5%, 3%, 1%, and 2% whereas the logistic model decreased about 2%. As a result, a Naïve Bayes, Decision Tree, and logistic model were less likely to be overfitting, followed by Random Forest and Neural Network (Table 3).

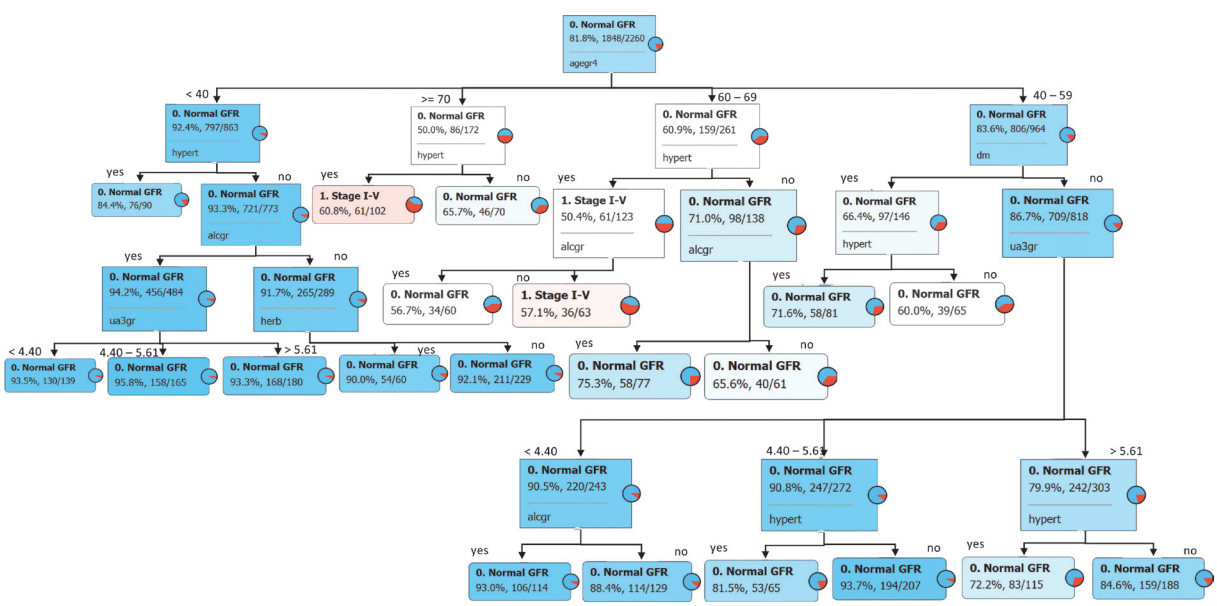
Applying a logistic regression in clinical practice should be straight forward by calculation of probability of CKD occurrence following the logit equation. For instance,

subjects aged 70 years or older with hypertension would have probability of CKD occurrence by logistic and Naïve Bayes models of 69.4% and 54.0%, respectively. A Decision Tree should be also easily in applying. A tree was firstly split to the right-side if subjects aged 60 to 69 years and 70 years or older. For those patients aged 70 years or older were split further if they had hypertension, then they were finally classified as CKD at the end tree leaf with probability of 60.8%. For those aged 60 to 69 years with hypertension, the tree was split further if they had alcohol consumption and then finally classified as CKD with probability of 57.1%. Conversely, for those subjects aged younger than 60 years with diabetes, hypertension, uric acid (> 5.61 mg/dL), alcohol consumption, and use of traditional medicine, a tree missclassified them as non-CKD (Figure 1).

Table 3. Prediction Performances

Model	Data type	Recall	Precision	C Statistics	Accuracy	F1
Logistic Regression	train	0.84	0.82	0.79	0.84	0.81
	test	0.83	0.80	0.77	0.83	0.80
Decision Tree	train	0.83	0.80	0.77	0.83	0.77
	test	0.83	0.81	0.75	0.83	0.80
Naïve Bayes	train	0.83	0.81	0.78	0.83	0.82
	test	0.81	0.79	0.77	0.81	0.80
Random Forest	train	0.85	0.84	0.80	0.85	0.82
	test	0.83	0.79	0.77	0.83	0.80
Neural Network	train	0.86	0.81	0.82	0.83	0.84
	test	0.82	0.79	0.77	0.82	0.80

Figure 1. Prediction of CKD Occurrence Using a Decision Tree Analysis



Abbreviations: agegr4, age group;alcgr, alcohol group; CKD, chronic kidney disease; dm, diabetes mellitus.; GFR, glomerular filtration rate; hypert, hypertension; ua3gr, uric acid group.

Discussion

We had applied ML algorithms for CKD prediction relative to a traditional logistic regression using data of Thai SEEK study. Seven features were included in the prediction models. By comparison performance between training and testing, our study demonstrated that logistic regression, Neural Network, and Random Forest have comprehensible discrimination performance, but the logistic regression tends to have lower overfitting compared to Neural Network, and Random Forest.

We have also demonstrated how to apply our prediction models in clinical practice based on information of 7 features including age, alcohol consumption, diabetes, hypertension, kidney stone, uric acid, and traditional medicine usage. To ease of use in clinical practice, an application in a mobile device should be developed further, particularly for a logistic regression model in which model performance was good and interpretation was clinically meaningful.

Conversely, a Decision Tree, although its performance was quite good in general, it may be faced with false negative results in subjects aged younger than 60 years leading to clinically uninterpretable outcome. For instance, subjects age younger than 60 years with diabetes, hypertension, and uric acid more than 5.61 mg/dL would be missclassified as non-CKD. Therefore, the Decision Tree further requires to re-tune parameters in larger updated training and testing data sets.

A risk-based approach to screening which was suggested by clinical practice guidelines is including those aged older than 60 years, or with having any history of hypertension or diabetes,[3, 20] which is similar to our findings. In addition, our study found that serum uric acid is significantly associated with CKD which is consistent with previous studies.[21, 22] Experimental studies have documented pathophysiological mechanisms beyond monosodium urate (MSU) crystal deposition in kidney, which may possibly involve overproduction of chemotactic cytokines, cell proliferation, and inflammation, eventually leading to the development of both tubule-interstitial and glomerular damage.[23] A recent research showed proinflammatory pathways, possibly mediated by toll-like receptor 4 and implicated in tubule-interstitial damage, are induced in an additive manner by uric acid and angiotensin II in proximal tubular epithelial cells.[24]

Many risk prediction models had previously been developed which mainly used traditional statistical models of logistic or Cox regressions with varied discriminative C statistics of about 0.7 to 0.9.[25] For those models with high performance, they considered more features such as glycated hemoglobin, duration of diabetes, diabetic controlled drugs, urine-albumin level, etc. Some of those features are available in a routine clinical practice, but some may be not and also costly in performing.

Considering more number of features in the models may cause interactions among them, in which traditional statistic models like logistic regression may be limited. Unlike traditional statistics, the MLs could deal with more complex interactions among features, but they could not avoid model overfitting particularly for Neural Network and Random Forest models.

Our finding showed that ML methods could not improve in prediction of CKD more than traditional logistic regression. This was similar to commentary on the use of MLs in prediction of myocardial infraction,[26] suggesting use of MLs in clinical research when there are unstructured features such as signal and image data. Further research should consider dynamic and signal of clinical features to improve prediction of CKD using ML models.

The strength of our study was that we applied 4 ML algorithms in the nationwide cross-sectional study for CKD detection in adult Thai population which was the first study in Thailand including Decision Tree, Naïve Bayes, Random Forest, and Neural Network. In addition, performance of these MLs were compared with traditional statistic methods in CKD predictions. We also demonstrated the clinical application of

Decision Tree which could be apply in clinical based on common features easily. However, some limitations could be not avoided. First, this prediction model derived from a cross-sectional study which could not claim a causal association of CKD risk. Second, we have yet performed external validation, thus, the generalizability and reproducibility of predictive model is still questionable. Further study should be conducted to include signal and image data in ML models.

Conclusions

Seven features are used to predict CKD occurrence including age, uric acid, hypertension, diabetes, alcohol consumption, traditional medicine use, and kidney stone. Logistic regression, Neural Network, and Random Forest have comprehensible discrimination performance, but the logistic regression tends to have lower overfitting compared to Neural Network, and Random Forest.

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet.* 2020;396(10258): 1204-1222. doi:10.1016/S0140-6736(20)30925-9
2. GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the lobal Burden of Disease Study 2017. *Lancet.* 2020;395(10225): 709-733. doi:10.1016/S0140-6736(20)30045-3
3. Inker LA, Astor BC, Fox CH, et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. *Am J Kidney Dis.* 2014;63(5):713-735. doi:10.1053/j.ajkd.2014.01.416
4. Plantinga LC, Boulware LE, Coresh J, et al. Patient awareness of chronic kidney disease: trends and predictors. *Arch Intern Med.* 2008;168(20):2268-2275. doi:10.1001/archinte.168.20.2268
5. Ingsathit A, Thakkinstian A, Chaiprasert A, et al. Prevalence and risk factors of chronic kidney disease in the Thai adult population: Thai SEEK study. *Nephrol Dial Transplant.* 2010;25(5):1567-1575. doi:10.1093/ndt/gfp669
6. Chen TK, Knicely DH, Grams ME. Chronic kidney disease diagnosis and management: a review. *JAMA.* 2019;322(13):1294-1304. doi:10.1001/jama.2019.14745
7. Elley CR, Robinson T, Moyes SA, et al. Derivation and validation of a renal risk score for people with type 2 diabetes. *Diabetes Care.* 2013;36(10):3113-3120. doi:10.2337/dc13-0190
8. Lin CC, Li CI, Liu CS, et al. Development and validation of a risk prediction model for end-stage renal disease in patients with type 2 diabetes. *Sci Rep.* 2017;7(1):10177. doi:10.1038/s41598-017-09243-9
9. Demšar J, Curk T, Erjavec A, et al. Orange: data mining toolbox in python. *J Mach Learn Res.* 2013;14(1):2349-2353.
10. Miao DD, Pan EC, Zhang Q, Sun ZM, Qin Y, Wu M. Development and validation of a model for predicting diabetic nephropathy in Chinese people. *Biomed Environ Sci.* 2017;30(2): 106-112. doi:10.3967/bes2017.014
11. Wan EYF, Fong DYT, Fung CSC, et al. Prediction of new onset of end stage renal disease in Chinese patients with type 2 diabetes mellitus - a population-based retrospective cohort study. *BMC Nephrol.* 2017; 18(1):257. doi:10.1186/s12882-017-0671-x
12. Wu M, Lu J, Zhang L, et al. A non-laboratory-based risk score for predicting diabetic kidney disease in Chinese patients with type 2 diabetes. *Oncotarget.* 2017;8(60):102550-102558. doi:10.18632/oncotarget.21684
13. Thakkinstian A, Ingsathit A, Chaiprasert A, et al. A simplified clinical prediction score of chronic kidney disease: a cross-sectional-survey study. *BMC Nephrol.* 2011;12:45. doi:10.1186/1471-2369-12-45
14. Dagliati A, Marini S, Sacchi L, et al. Machine learning methods to predict diabetes complications. *J Diabetes Sci Technol.* 2018;

12(2):295-302. doi:10.1177/2193296817706375

15. Rodriguez-Romero V, Bergstrom RF, Decker BS, Lahu G, Vakilynejad M, Bies RR. Prediction of nephropathy in type 2 diabetes: an analysis of the accord trial applying machine learning techniques. *Clin Transl Sci.* 2019;12(5):519-528. doi:10.1111/cts.12647
16. Song X, Waitman LR, Hu Y, Yu ASL, Robbins DC, Liu M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *J Am Med Inform Assoc.* 2019;26(3): 242-253. doi:10.1093/jamia/ocy165
17. Goldfarb-Rumyantzev AS, Pappas L. Prediction of renal insufficiency in Pima Indians with nephropathy of type 2 diabetes mellitus. *Am J Kidney Dis.* 2002;40(2):252-264. doi:10.1053/ajkd.2002.34503
18. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA.* 2018; 319(13):1317-1318. doi:10.1001/jama.2017.18391
19. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019; 110:12-22. doi:10.1016/j.jclinepi.2019.02.004
20. Farrington K, Covic A, Aucella F, et al. Clinical practice guideline on management of older patients with chronic kidney disease stage 3b or higher (eGFR <45 mL/min/1.73 m^2). *Nephrol Dial Transplant.* 2016; 31(suppl 2):ii1-ii66. doi:10.1093/ndt/gfw356
21. De Cosmo S, Viazzi F, Pacilli A, et al. Serum uric acid and risk of CKD in type 2 diabetes. *Clin J Am Soc Nephrol.* 2015;10(11): 1921-1929. doi:10.2215/CJN.03140315
22. Takae K, Nagata M, Hata J, et al. Serum uric acid as a risk factor for chronic kidney disease in a Japanese Community- The Hisayama Study. *Circ J.* 2016;80(8):1857-1862. doi:10.1253/circj.CJ-16-0030
23. Liu H, Xiong J, He T, et al. High uric acid-induced epithelial-mesenchymal transition of renal tubular epithelial cells via the TLR4/NF-kB signaling pathway. *Am J Nephrol.* 2017; 46(4):333-342. doi:10.1159/000481668
24. Milanesi S, Verzola D, Cappadona F, et al. Uric acid and angiotensin II additively promote inflammation and oxidative stress in human proximal tubule cells by activation of toll-like receptor 4. *J Cell Physiol*. 2019;234(7): 10868-10876. doi:10.1002/jcp.27929
25. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS Med.* 2012;9(11): e1001344. doi:10.1371/journal.pmed.1001344
26. Engelhard MM, Navar AM, Pencina MJ. Incremental benefits of machine learning-when do we need a better mousetrap? *JAMA Cardiol.* 2021;6(6):621-623. doi:10.1001/jamacardio.2021.0139

Supplement

Table 1. Machine Model Calibration

Model	Hyper Parameter	Value	Optimum Value
Logistic Regression	Penalty Strength	[0.001, 1000][1]	10
	Regularization	L1, L2	L1
Decision Tree	Number of instances per leaves	[10, 100][2]	60
	Minimum instances for splitting	[1, 10][3]	5
	Limit maximal tree depth	[1, 10][3]	4
	Stop when majority reached	[50 - 99][3]	95
Naïve Bayes	None	None	None
Random Forest	Number of trees	[1, 10][3]	9
	Number of attributes considered	[1, 10][3]	4
	Depth of each tree	[1, 10][3]	6
	Minimum instances for splitting	[1, 10][3]	7
Neural Network	Neurons in layers	[100 - 400][2]	230
	Activation	ReLu, identity, logistic, tanh	ReLu
	Solver	Adam, SGD, L-BSGF-B	Adam
	Regulation	[0.0001, 10000][1]	0.001
	Maximal number of iterations	[10 - 100][2]	70

[1] Log scale.

[2] 10x scale.

[3] Linear scale.

RMJ Ramathibodi Medical Journal

Rama Med J | Original Article

การเรียนรู้ด้วยเครื่องสำหรับการพยากรณ์โรคไตเรื้อรังในประชากรไทยวัยผู้ใหญ่

รัตน์ชัยนันท์ ธรรมสุจริต[1], ปัณณธร อิงค์สาธิต[1,2], ชิจิต อาริ ซาปูโตร[1,3], อติพร อิงค์สาธิต[1], อัมรินทร์ ทักขิญเสถียร[1]

[1] ภาควิชาระบาดวิทยาคลินิกและชีวสถิติ คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล กรุงเทพฯ ประเทศไทย

[2] โรงเรียนเตรียมอุดมศึกษา กรุงเทพฯ ประเทศไทย

[3] ภาควิชาระบาดวิทยา ชีวสถิติ ประชากร และการส่งเสริมสุขภาพ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยแอร์ลังกา เมืองซูราบายา ประเทศอินโดนีเซีย

บทนำ: โรคไตเรื้อรังเป็นโรคที่ใช้ทรัพยากรอย่างมากในการดูแลรักษา การค้นหาโรคได้ตั้งแต่ระยะเริ่มต้นด้วยการใช้แบบจำลองทำนายความเสี่ยง จะเป็นประโยชน์ในการค้นหาผู้ป่วยที่เสี่ยงในการเกิดโรคและสามารถให้การรักษาได้ตั้งแต่ระยะเริ่มต้น

วัตถุประสงค์: เพื่อเปรียบเทียบแบบจำลองที่สร้างจากสมการ Logistic regression กับการเรียนรู้ของเครื่อง (Machine learning) ในการทำนายความเสี่ยงของการเกิดโรคไตเรื้อรังในประชากรไทยวัยผู้ใหญ่

วิธีการศึกษา: ข้อมูลสำหรับการศึกษานี้นำมาจากโครงการ Thai Screening and Early Evaluation of Kidney Disease (SEEK) ประกอบด้วย 17 ตัวแปรในการสร้างแบบจำลองทำนายความเสี่ยงโดยใช้วิธี Logistic regression และการเรียนรู้ของเครื่อง 4 วิธี (Random Forest, Naïve Bayes, Decision Tree, Neural Network) โดยข้อมูลถูกแบ่งออกเป็นการฝึกและการทดสอบข้อมูลในสัดส่วน 70:30 การประเมินสมรรถนะใช้ค่า Recall, C statistics, Accuracy, F1, และ Precision

ผลการศึกษา: ตัวแปร 7 จาก 17 ตัวแปรได้ถูกคัดเลือกในการสร้างแบบจำลองทำนายความเสี่ยงพบว่า แบบจำลองจากสมการ Logistic regression สามารถจำแนกผู้ป่วยที่เป็นโรคไตเรื้อรังออกจากผู้ป่วยที่ไม่เป็นโรคได้ดี โดยมีค่า C statistics เท่ากับ 0.78 และ 0.78 จากการฝึกและการทดสอบข้อมูลตามลำดับ ในขณะที่แบบจำลองจาก Neural Network ให้ผลลัพธ์ดีที่สุดเมื่อเทียบกับตัวแบบอื่นๆ ที่สร้างจากการเรียนรู้ของเครื่อง เช่น Random Forest, Naïve Bayes, และ Decision Tree โดยมีสมรรถนะในการฝึกวัดจากค่า C statistics เท่ากับ 0.82, 0.80, 0.78, และ 0.77 ตามลำดับ ส่วนสมรรถนะในการทดสอบลดลงร้อยละ 5 ร้อยละ 3 ร้อยละ 1 และร้อยละ 2 ตามลำดับ ในขณะที่แบบจาก Logistic ลดลงร้อยละ 2

สรุป: ในบรรดาแบบจำลองที่สร้างขึ้นมานั้น Logistic regression, Neural Network, และ Random Forest มีสมรรถนะในการจำแนกผู้ป่วยได้ใกล้เคียงกัน แต่ Logistic regression มีแนวโน้มที่จะ Overfit น้อยกว่า Neural Network และ Random Forest

คำสำคัญ: โรคไตเรื้อรัง การเรียนรู้ของเครื่อง แบบจำลองทำนายทางคลินิก

Rama Med J: doi:10.33165/rmj.2021.44.4.250334

Received: June 24, 2021 **Revised:** October 15, 2021 **Accepted:** November 19, 2021

Corresponding Author:
รัตน์ชัยนันท์ ธรรมสุจริต
ภาควิชาระบาดวิทยาคลินิก
และชีวสถิติ
คณะแพทยศาสตร์
โรงพยาบาลรามาธิบดี
มหาวิทยาลัยมหิดล
270 ถนนพระรามที่ 6
แขวงทุ่งพญาไท เขตราชเทวี
กรุงเทพฯ 10400 ประเทศไทย
โทรศัพท์ +66 2201 1269
อีเมล ratchainant.tha@mahidol.edu