Machine Learning Prediction Models for Chronic Kidney Disease using National Health Insurance Claim Data in Taiwan

Surya Krishnamurthy, Kapeleshh KS, Erik Dovgan, Mitja Luštrek, Barbara Gradišek Piletič, Kathiravan Srinivasan, Yu-Chuan (Jack) Li, Anton Gradišek, Shabbir Syed-Abdul

Supplementary Materials

Table S1. Hyperparameter values of the tested models. Models with prefix “temp” use the temporal quarterly data.

Algorithm	Hyperparameter
Logistic Regression	solver = 'liblinear', class_weight = {1:4, 0:1}, C = 0.01, penalty = 'l1'
Decision Tree	**12 Month** max_depth=8, class_weight = {1:4,0:1}, min_samples_leaf = 30
	6 Month max_depth=10, class_weight = {1:4,0:1}, min_samples_leaf = 30
Random Forest	**12 Month** criterion='gini', class_weight = {1:4, 0:1} n_estimators = 700, max_depth = 18
	6 Month criterion='gini', class_weight = {1:4, 0:1} n_estimators = 800, max_depth = 18
LightBoost	**12 Month** boosting_type = 'gbdt',

	learning_rate = 0.01, n_estimators = 800, scale_pos_weight=4
	6 Month boosting_type = 'gbdt', learning_rate = 0.01, n_estimators = 1000, scale_pos_weight=4
CNN	**Conv layer:** n_filters = 64 filter_size = 3 stride = 1 activation = relu padding = same **Max pool layer:** filter_size = 3 stride = 1 **Dropout:** drop_probabiity = 0.4 **Dense layer:** n_units = 32 activation = relu initialization = random normal **Output layer:** activation: sigmoid **Optimizer:** algorithm: adadelta learning rate: 0.1 epochs: 15
BLSTM / BLSTM-qtr	**Bi-LSTM layer:** LSTM units: 64 activation: relu **Dense layer:** n_units = 32 activation = relu initialization = random normal **Output layer:** activation: sigmoid **Optimizer:** algorithm: adadelta learning rate: 0.1 epochs: 15
CNN-qtr	**Conv layer:**

	n_filters = 32 filter_size = 3 stride = 1 activation = relu padding = same **Max pool layer:** filter_size = 3 stride = 1 **Dense layer:** n_units = 32 activation = relu initialization = random normal **Output layer:** activation: sigmoid **Optimizer:** algorithm: adadelta learning rate: 0.1 epochs: 15
temp-lightgbm	**12 Month** boosting_type='gbdt', colsample_bytree=0.7275999182022543, learning_rate=0.06467440873590191, max_depth=10, min_child_samples=20, min_child_weight=0.001, min_split_gain=0.34234014807063695, n_estimators=750, num_leaves=50, reg_lambda=1.1, scale_pos_weight=4, subsample=0.8,
	6 Month boosting_type='gbdt', colsample_bytree=0.7942201755684853, learning_rate=0.08416317594127293, max_depth=15, min_child_samples=20, min_child_weight=0.001, min_split_gain=0.3230893825622149, n_estimators=400, num_leaves=100, reg_lambda=1.1, scale_pos_weight=4, subsample=0.7

temp-randomforest	**12 Month** criterion='gini', n_estimators = 800, max_depth = 15, class_weight = {1:4, 0:1}
	6 Month criterion='gini', n_estimators = 800, max_depth = 18, class_weight = {1:4, 0:1}
temp-decisiontree	**12 Month** class_weight={0: 1, 1: 4}, criterion='gini', max_depth=8, min_samples_leaf=35, splitter='best'
	6 Month class_weight={0: 1, 1: 4}, criterion='gini', max_depth=8, min_samples_leaf=35, splitter='best'
temp-LogisticRegression	**12 Month** C=1, class_weight={0: 1, 1: 4}, penalty='l2', solver='liblinear'
	6 Month C=0.1, class_weight={0: 1, 1: 4}, penalty='l2', solver='liblinear'

Table S2. The threshold values used for each model to computer the performance metrics.

Algo	Threshold	
	6 month	**12 month**
CNN	0.372	0.382
BLSTM	0.480	0.313
CNN-qrt	0.452	0.495
BLSTM-qrt	0.545	0.559
temp-lightgbm	0.436	0.429
temp-randomforest	0.436	0.470
temp-logistic	0.18	0.49
temp-decisiontree	0.484	0.486
lightgbm	0.483	0.514
randomforest	0.436	0.450
logistic	0.49	0.49
decisiontree	0.463	0.467

Dataset analysis

The following figures show the correlation plots of the top 25 features that have contributed to the occurrence of CKD. In these plots, the entries that start with an alphabet are the ATC codes for medications (eg: C03CA, C09CA, etc). The other entries that are fully numeric are the ICD9 codes for comorbidities (eg: 250, 274, etc).

- From the correlation plots, we can see that all three cases are similar. This indicates that the patient distribution in the dataset is similar among the CKD and the non-CKD groups.
- No two features are highly correlated with each other, with the highest being 0.5.
- The high correlation scores are between
    - 274 and M04AC, MO04AA, M04AB
      274 - ICD9 Code for Gout

M04 - ATC Code for Anti Gout medications

- None of the features are directly correlated to the occurrence of CKD (Figure S1), indicating the complexity of the relationship with the outcome.

Figure S1 represents the correlation plot of all the patients (CKD and non-CKD).

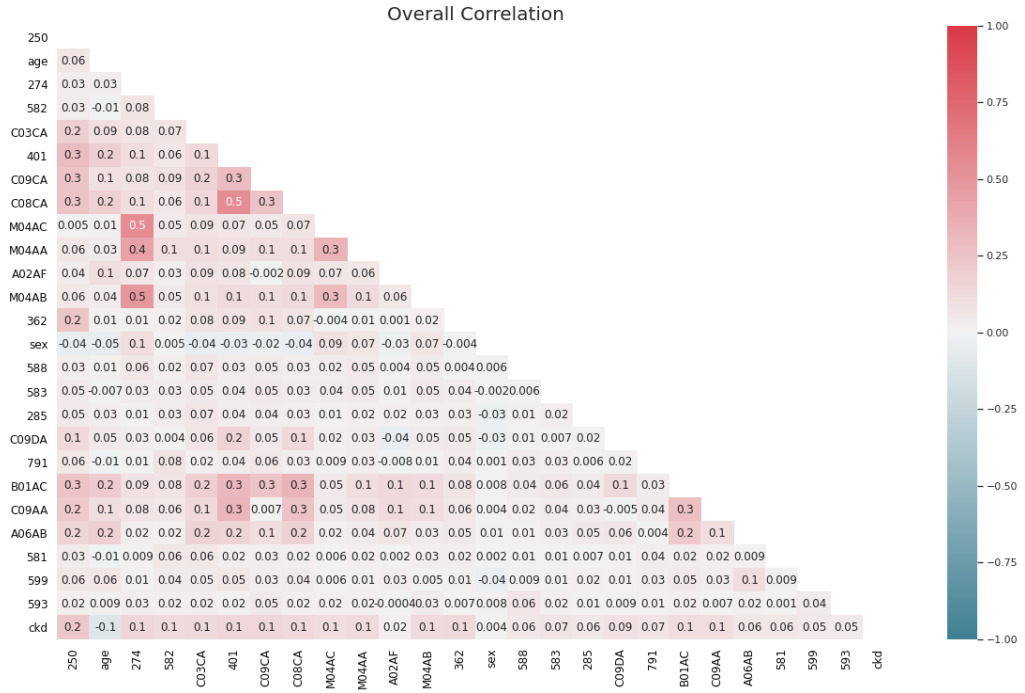

Figure S1

Figure S2 represents the correlation analysis for the CKD patients.

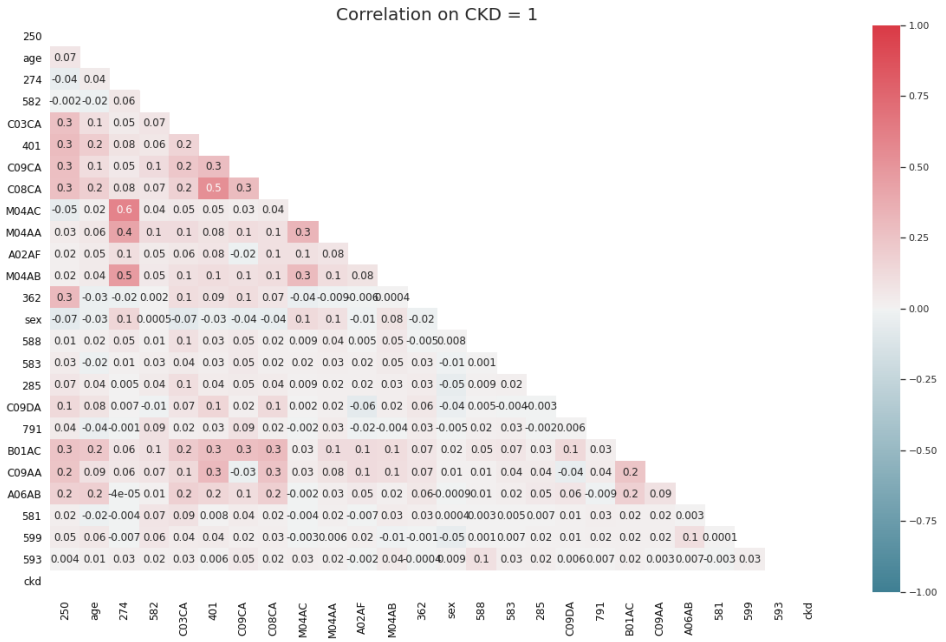

Figure S2

Figure S3 represents the correlation analysis for the non-CKD people.

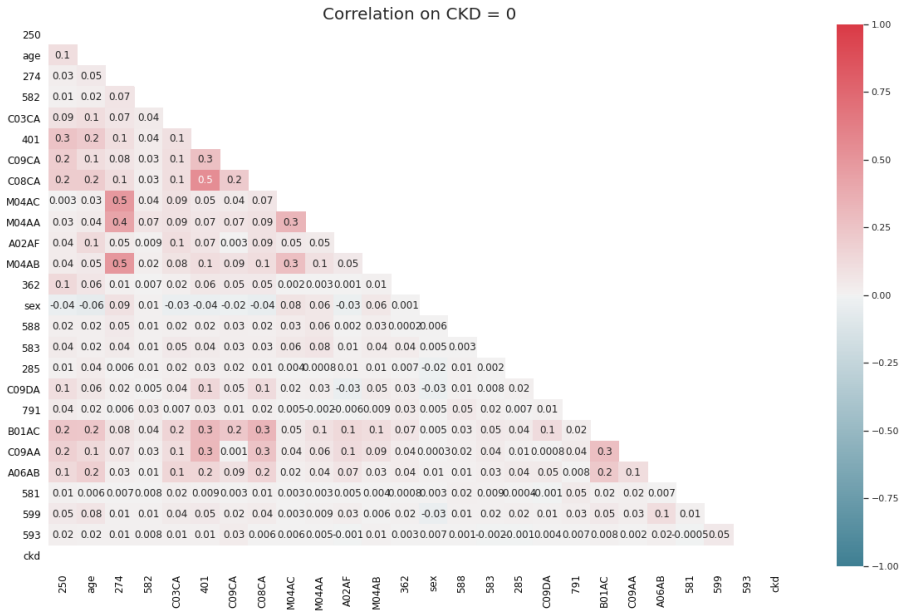

Figure S3

The following two figures show the histograms for the months visited by the patients for receiving a diagnosis or drug during the observation window in the entire dataset . Fig S4 looks at the CKD patients and Fig S5 at the non-CKD group. For example, from Fig S4 we can see that about 5800 patients visit every month (total 24 months in our observation period) for checkups and about 5000 patients visit every month to collect drugs. Both groups show similar trends (one should keep in mind that the non-CKD set is 4 times larger). We also see that most of the patients in the cohort visit every month.. .

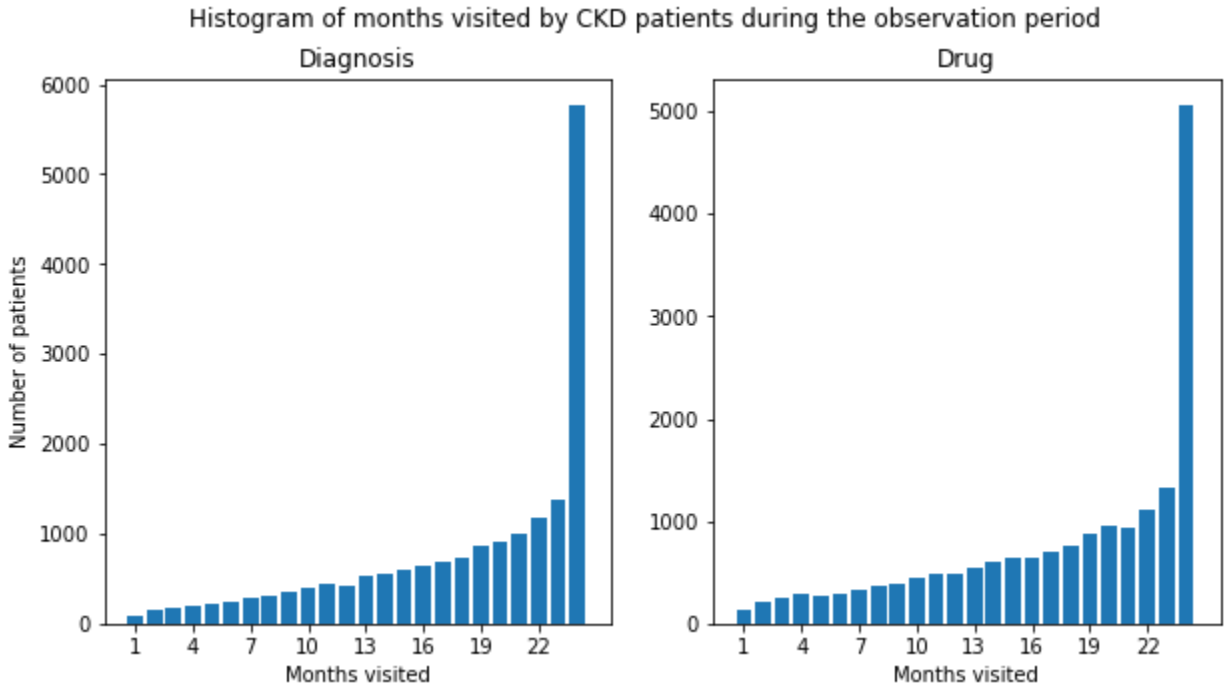

Figure S4: Number of months visited by the CKD patients for diagnosis/drugs during the observation period (24 months) as a histogram

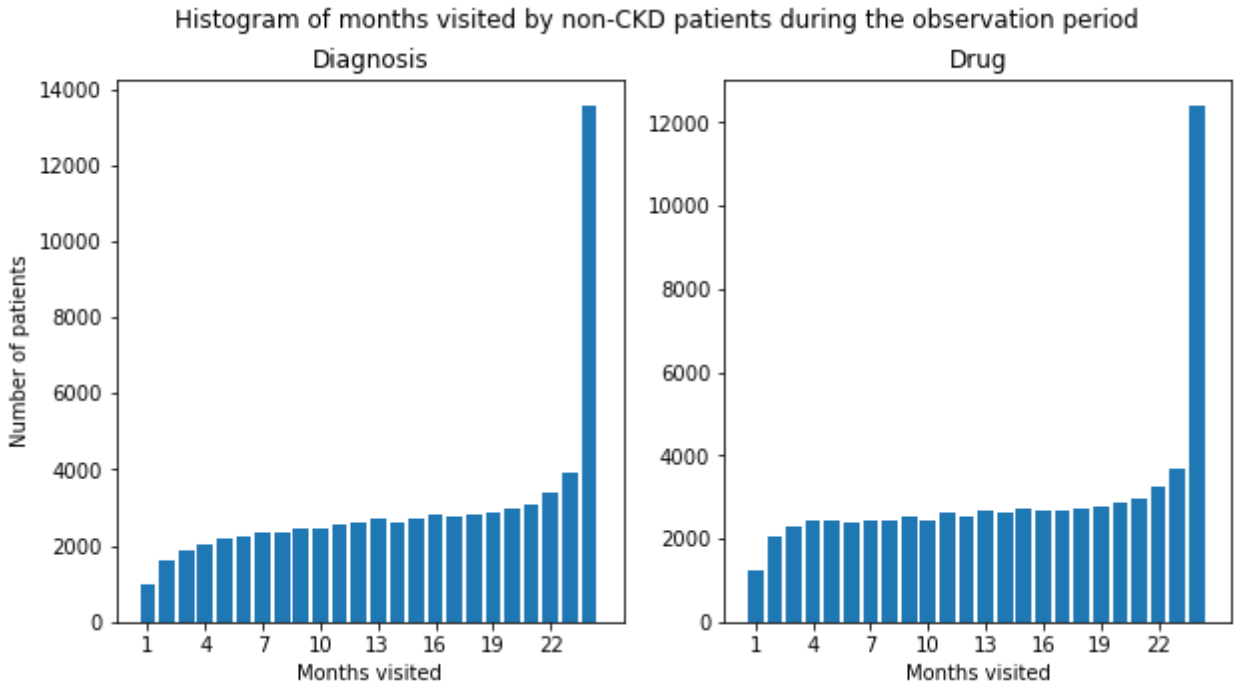

Figure S5: Number of months visited by the non-CKD patients for diagnosis/drugs during the observation period (24 months) as a histogram