Prediction of End Stage Renal Disease in Patients with Type 2 Diabetes Mellitus Patients Using Common Data Model and Machine Learning Algorithm

Hyuna Yoon[1,2], Kyungseon Choi[1,2], Sang Youl Rhee[3], Hae Sun Suh[1,2,4*]

1. Department of Regulatory Science, Graduate School, Kyung Hee University
2. Institute of Regulatory Innovation through Science (IRIS), Kyung Hee University
3. College of Medicine, Kyung Hee University
4. College of Pharmacy, Kyung Hee University

* corresponding author

Background

End Stage Renal Disease (ESRD) is a devastating disease for patients and induces significant socio-economic burden. One leading cause of ESRD is type 2 diabetes mellitus (T2DM), which is very common disease worldwide. Predicting the risk of ESRD in T2DM patients could be helpful in preventing ESRD. The aim of our study was to develop a 10-year ESRD risk prediction model among Korean T2DM patients using a single hospital registry and externally validate the developed model using the other hospitals' registry data. We developed prediction models using several machine learning algorithms to include as many predictive variables as possible.

Methods

Retrospective cohort study was conducted using electronic health record (EHR) data from five secondary or tertiary hospitals: Ajou university medical center, Kangdong sacred heart hospital, Kyung Hee university hospital, Myongji hospital, Wonkwang university hospital. The EHR databases were standardized to the Observational Medical Outcomes Partnership common data model version 5.3.1. For developing prediction model and internal validation, data from Ajou university medical center, Kandong sacred heart hospital, Myoungji hospital and Wonkwang university were used, and for external validation, data from Kyung Hee university was used. Study population consisted of the patients who were newly diagnosed with T2DM at a single hospital and had no history of ESRD. The index date was defined as the day the patient was first diagnosed with T2DM at a single hospital. The primary endpoint of our study was the occurrence of ESRD. To be included in the analysis, patients in the non-ESRD group were required to have a follow-up period of at least 3650 days. All patients in the ESRD group were included in the analysis. Several prediction models were developed using L1-regularized logistic regression, gradient boosting machine, random forest algorithms. Among the developed models, the model with the best performance assessed by the area under

the receiver operator characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) was selected. Gender, age groups, condition group era, drug group era, measurement range group, and observation recorded withing 365 days before first diagnosis of T2DM were included as candidate predictors. After model development, model was externally validated using data from external hospital.

Results

The best model was developed using random forest algorithm and the data from Myongji hospital, which included 1,022 target patients (786 with no ESRD, and 236 with ESRD). External validation of model was done by using Kyung Hee university hospitals' data, which included 2,517 target patients (2,062 with no ESRD, and 455 with ESRD). Final model included 270 covariates, including above normal range of creatinine, urea nitrogen, having condition of hypertension, renal impairment, etc. The internal validation results showed receiver operating characteristic (ROC) curve and precision-recall (PR) curve depicted in figure 1, having AUROC of 0.956 (95% CI: 0.935-0.977) and AUPRC of 0.921. External validation resulted in ROC and PR curve shown in figure 2. The AUROC and AUPRC were calculated to be 0.793 (95% CI; 0.775-0.812) and 0.534 respectively. Based on the results of internal validation, the model can be evaluated as having outstanding performance, while the external validation results indicate that the model's performance is either acceptable or good.

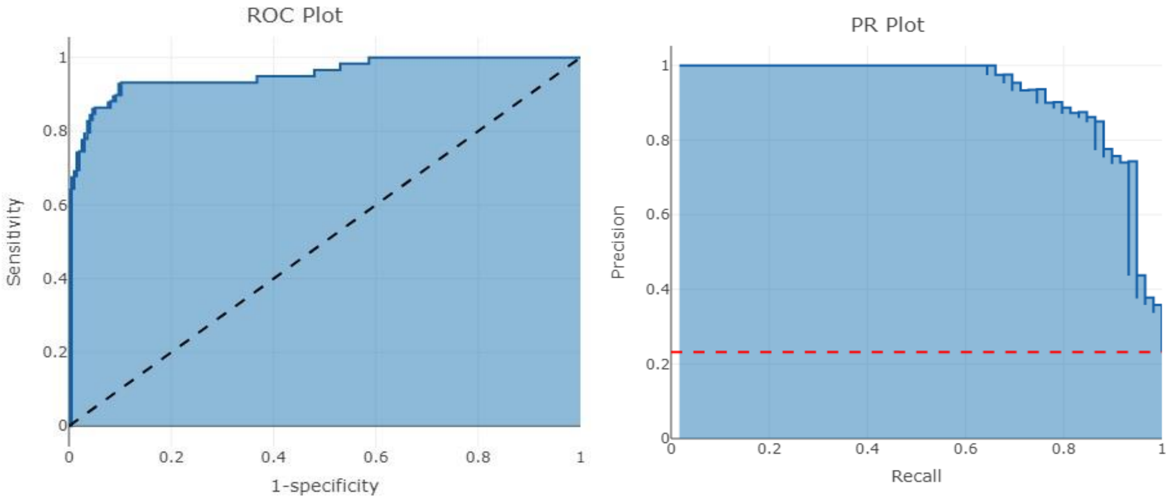

Figure 1 Receiver Operation Characteristic (ROC) curve and Precision Recall (PR) curve: Internal validation results

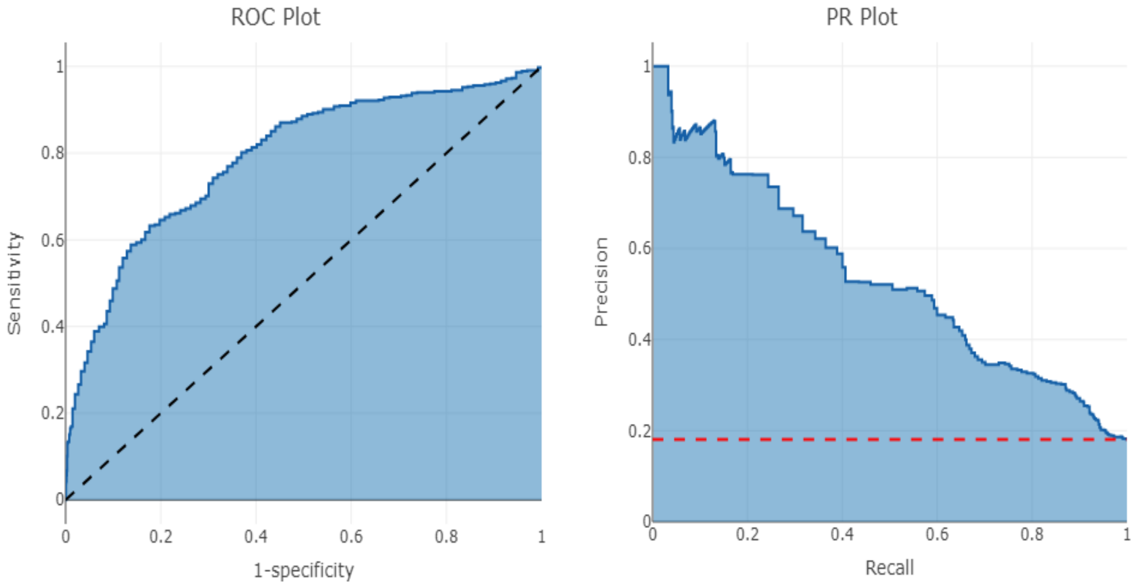

Figure 2 Receiver Operation Characteristic (ROC) curve and Precision Recall (PR) curve: External validation results

Conclusion

we developed a prediction model using the random forest model algorithm to predict ESRD in patients with T2DM. We used EHR data standardized into a CDM. The internal validation results of our model demonstrated outstanding performance, and external validation result showed acceptable performance. If further external validation is conducted confirming the consistent predictive performance of the model, it can be applied in real clinical practice to assist in identifying the group of patients with T2DM who require ESRD prevention and contribute to reducing the incidence rate of ESRD.