Review of chronic obstructive pulmonacy disease operational definition model using machine learning model

Sang Chul Lee, MD[1], JH Hong[2]*, MS; Hyunsun Lim, PhD[2].*

*[1]Department of Pulmonology, National Health Insurance Service Ilsan Hospital, [2]Department of Research and Analysis, National Health Insurance Service Ilsan Hospital, *Authors contributed equally*

Background. In recent years, research utilizing the claims data has increased. However, since the medical records in the database were collected for the purpose of claiming and reimbursing health insurance benefits, there may be inconsistencies between the actual disease status and the diagnosis based on the claimed information and ICD-10 codes. Therefore, it is necessary to review and consider this issue as it may lead to overestimation or underestimation of the disease prevalence. In this study, we aim to investigate whether machine learning models can be used for operational definitions of chronic obstructive pulmonary disease (COPD), which is difficult to define operationally, by comparing them with rule-based methods and expert-defined operational definitions using various hospital data.

Methods. We used data that combined the Korean National Health Insurance Service (NHIS) database with data from people who received pulmonary function tests at Ilsan Hospital between 2011 and 2020 (n=40,121). The rule-based method was defined as cases where J44 was diagnosed twice or more per year, and inhaled medications were prescribed twice or more per year in the same year. The silver standard was J43, J44 and was defined as a case with a pulmonary function test result of FEV1/FVC ratio (%) of 70 or less and a smokin history. Prediction models were estimating using Multiple logistic model, Ridge, Lasso, Elastic, support vector machine, random forest, XGBoost, random forest and XGBoost ensemble models. The model used 25 features: 1 silver standard definitions using hospital records, 3 personal information, 8 health examinations, and 13 medical records.

Results. A total of 40,121 participants were included in the study, of whom 2,353 were in the COPD group based on the silver standard, and 37,768 were in the control group. We defined COPD using rule-based methods, silver standard, and machine learning techniques and validated the models using 712 patients who were directly reviewed by experts (gold standard).

Operational definitions	AUROC	AUPRC	Accuracy	Recall	Precision	Specificity	NPV
Rule-based method	61.25	92.78	33.69	24.95	98.68	97.56	15.1
Silver standard	79.29	96.67	66.33	62.22	99.21	96.37	25.87
MLR	83.3	97.17	80.03	82.44	94.14	62.47	32.73
Ridge regression	82.17	96.89	84.55	89.3	92.87	49.87	38.93
Lasso regression	72.16	96.89	84.54	89.3	92.86	49.8	38.9
Elastic-Net regression	82.22	96.91	84.53	89.32	92.82	49.47	38.8
Support vecort machine	62.69	92.06	67.18	70.2	90.34	45.12	17.16
Random forest	83.5	97.21	78.67	79.46	95.55	72.96	32.7
XGBoost	83.28	97.15	79.8	81.96	94.34	64.05	32.67
Ensemble (Random Forest, XGBoost)	83.25	97.19	81.05	83.27	94.54	64.84	34.65

AUROC: The area under the receiver operating characteristic curve, AUPRC: The area under precision-recall curve, NPV: Negativie predictive value, MLR: Multiple logistic regression

The area under the ROC curve and precision-recall curve remained at around 62.6-0.83.5 and 97 espectively. In particular, the random forest model had a slightly lower recall (sensitivity) but higher specificity by about 2 compared to the logistic regression model. Based on the ROC curve, the ridge model with regularization was considered the most appropriate.

Conclusion. The Rule-based method had a higher specificity but lower recall and lower area under the ROC curve and precision-recall curve than the machine learning models. This trend may be due to the conservative definition of the disease, which is not easy to manipulate, and diseases with more conservative definitions may be more likely to choose a conservative definition. This conservative approach may make it easier for researchers to demonstrate the efficacy of the intervention. However, if the goal is disease prevention or early treatment, further evaluation of the appropriateness is necessary.