Validation of Machine Learning-based Models for Estimating Low-Density Lipoprotein Cholesterol using OHDSI network

Sora Youn, Suk Jung, Taehoon Ko, Gyu Chul Oh, Kwangsoo Kim, Sae Won Choi, Yeseul Bae, Hae Young Lee, Hyung Jin Yoon, Kyung Hwan Kim
1 Biomedical research institute, Seoul National University Hospital, Seoul, South Korea;
2 Office of Hospital Information, Seoul National University Hospital, Seoul, South Korea;
3 Department of Internal Medicine, Division of Cardiology, Seoul National University Hospital, Seoul, South Korea;
4 Department of Radiology, Seoul National University College of Medicine, Seoul, Korea

Background

Low-Density Lipoprotein Cholesterol (LDL-C) is a very important factor in the field of cardiovascular disease as well as a therapeutic target [1]. LDL-C can be included in the standard lipid profile and measured directly. However, in some cases, LDL-C can be estimated based on a formula based on total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), LDL-C and triglyceride (TG). Especially in South Korea, when implementing all four standard lipid tests, the National Health Insurance of South Korea adopts a policy of recognizing only three tests among the standard lipid profile except for some diseases such as hypertension, hyperlipidemia. For this reason, calculating LDL_C have been made instead of directly measuring it.

There are 3 existing LDL_C calculation formula; Friedewald [2], Chen [4-6] and Martin formula [3]. Table 1 shows the 3 formula.

1. Friedewold formula : LDL-C=TC-HDL-C-TG/5
2. Chen formula : LDL-C=(TC-HDL-C)*0.9-TG*0.1
3. Martin formula : LDL-C=TC-HDL-C-TG/7

Existing LDL_C calculations, except for Martin formula, were mostly focused on specific groups with fewer patients and no large group was verified. In this study, LDL_C estimation models are learned from large-scale data using machine learning models. We also compared the predictive performance with some conventional formulas.

Methods

We compare the model performance of our proposed models to 3 existing LDL_C calculation formula. In this study, machine learning models are used. The 2nd-order polynomial regression model adds square terms and interaction terms constructed with existing independent variables. The existing LDL_C calculation formula describe above only uses the first-order terms of TC, HDL-C and TG. On the other hand, our proposed models use the square of each variable such as TC^2, HDL_C^2, TG^2 and interactions like TC*HDL_C, TC*TG and HDL_C*TG.

We use 4 machine learning algorithm to predict the LDL_C: Ridge regression, Random Forests, Gradient Boosting and Multi-Layer Perceptrons (MLP) regression. These models are called Ridge, Random Forests, XGBoost, and MLP, respectively.

For this study, we collect the cases where 4 standard lipid tests were performed on one blood sample among SNUH CDM from October 11, 1999 to February 28, 2019. There are 217,559 patients satisfying the above condition and 754,737 cases with all four standard lipid tests. TC, HDL_C and TG are used as input variables and LDL_C as a target variable. Before training the models, the whole data is divided into training set and test set at a ratio of 7:3. To optimize parameters of the model, cross-validation method is used.

Conclusions

In this study, machine learning models that predict LDL_C using the remaining three lipid tests such as TC, HDL_C and TG are trained. And proposed models have better performance on large-scale data than conventional methods.

In the future, we plan to apply our models to other 2 institutions’ CDM with agreement with SNUH; Seoul National University Bundang Hospital (SNUBH) and Asan Medical Center (AMC). We will conduct the external validation about how well our models predict the LDL_C in patients with 2 hospitals.

Results

7 models such as Friedewald formula, Chen formula, Martin formula, trained Ridge model, trained Random Forest model, trained XGBoost and trained MLP model are evaluated with test set previously separated. Martin formula shows the best performance among the existing methods. (RMSE=9.99, R^2=0.95) The proposed machine learning models show good performance for all evaluation criteria. In this study, XGBoost model shows the best performance on test set. (RMSE=8.73, R^2=0.96)

We compared the existing formula with and proposed model using concordance in guideline classification, and bland-altman plot (Table 3, Figure 1&2)

Table 3 shows the concordance in guideline classification by 3 existing LDL-C calculation formula and our 4 proposed models. In more than 190 mg/dL and less than 70 mg/dL, the concordance of the existing LDL-C calculation formula is higher than our machine learning models. But between 70 mg/dL and 190 mg/dL, our machine learning model is higher than the existing formula, especially Ridge and XGBoost.

Acknowledgements & References

This study used clinical data retrieved by Common Data Model (CDM) of Seoul National University Hospital (SNUH).