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

[1]Sora Youn, [1]Suk Jung, [2]Taehoon Ko, [3]Gyu Chul Oh, [1]Kwangsoo Kim, [2]Sae Won Choi, [2]Yeseul Bae, [2]Hae Young Lee, [4]Hyung Jin Yoon, [2]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 diretly. However, in some cases, LDL_C can be estimated based on a formula based on total cholesterol (TC), high-density lipoprotein cholesterol (HDL-cholesterol, HDL_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.

- Friedewald formula : LDL_C=TC−HDL_C−TG/5
- Chen formula : LDL_C=(TC−HDL_C)×0.9−TG×0.1
- Martin formula : LDL_C=TC−HDL_C−TG/k

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 methods 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 Feburary 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.

Contact: contact@ohdsi.org

Results

7 models such as Friedewald formula, Chen formula, Martin formula, trained Ridge model, trained Random Forests model, trained XGBoost and trained MLP model are evaluated with test set previsously 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 our 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.

LDL-C, mg/dL	No. Concordant/Total Group (%)						
	Friedewald	Chen	Martin	Ridge	Random Forests	XGBoost	MLP
≥190	**35,490/48,281 (92.88)**	33,526/41,315 (87.74)	32,642/39,294 (85.43)	32,008/36,262 (83.77)	32,095/37,699 (84.00)	32,401/37,085 (84.80)	32,952/38,651 (86.24)
160 to 189	62,219/78,952 (78.63)	66,591/85,692 (82.15)	65,566/80,816 (82.86)	**67,832/82,263 (85.72)**	65,560/80,638 (82.85)	67,239/81,138 (84.97)	66,942/81,271 (84.60)
130 to 159	45,794/58,744 (72.35)	46,079/62,880 (72.80)	48,837/63,085 (77.16)	**50,690/64,794 (80.09)**	48,989/63,908 (77.40)	50,541/64,343 (79.85)	49,856/63,424 (78.77)
100 to 129	22,141/28,933 (68.36)	19,578/27,715 (60.44)	23,332/31,200 (72.03)	23,792/31,359 (73.45)	23,558/32,043 (72.73)	**24,171/31,859 (74.62)**	23,699/31,190 (73.17)
70 to 99	6,552/8,916 (62.68)	4,731/6,940 (45.26)	6,757/9,410 (64.64)	6,751/9,254 (64.58)	6,761/9,629 (64.68)	**6,980/9,676 (66.78)**	6,801/9,375 (65.06)
<70	**2,105/2,596 (71.40)**	1,556/1,880 (52.78)	2,066/2,617 (70.08)	2,042/2,490 (69.27)	1,995/2,505 (67.67)	1,941/2,321 (65.84)	2,046/2,511 (69.40)

Table 3. Concordance in Guideline Classification by 3 existing LDL-C calculation formula vs our proposed models

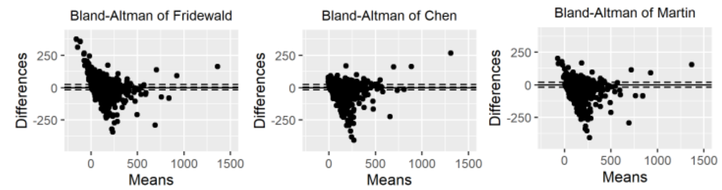

Figure 1. Bland-Altman plot of 3 existing LDL-C calculation formula

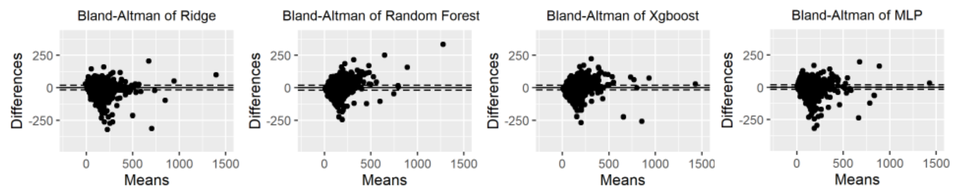

Figure 2. Bland-Altman plot of 3 existing LDL-C calculation formula our proposed models

Acknowledgements & References

This study used clinical data retrieved by Common Data Model(CDM) of Seoul National University Hospital(SNUH).
[1] P.S. Jellinger et al., American Association of Clinical Endocrinologists and American College of Endocrinology guidelines for management of dyslipidemia and prevention of cardiovascular disease, Endocrine Practice 23 (2017), 1-87.
[2] W.T. Friedewald, R.I. Levy, D.S. Fredrickson, Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge, Clinical chemistry 18(6) (1972), 499-502.
[3] S.S. Martin, Comparison of a novel method vs the Friedewald equation for estimating low-density lipoprotein cholesterol levels from the standard lipid profile, JAMA 310(18) (2013), 2061-2068.
[4] A. Rao et al., Calculation of low-density lipoprotein cholesterol with use of triglyceride/cholesterol ratios in lipoproteins compared with other calculation methods, Clinical chemistry 34(12) (1988), 2532-2534.
[5] S. Anandaraja et al., Low-density lipoprotein cholesterol estimation by a new formula in Indian population, International journal of cardiology 102(1) (2005), 117-120.
[6] T. Planella et al., Calculation of LDL-cholesterol by using apolipoprotein B for classification of nonchylomicronemic dyslipemia, Clinical chemistry 43(5) (1997). 808-815.
[7] Y. Chen et al., A modified formula for calculating low-density lipoprotein cholesterol values, Lipids in health and disease 9(1) (2010), 52.