Research Article

Ferhat Demirci*, Murat Emec, Ozlem Gursoy Doruk, Murat Ormen, Pınar Akan and Mehmet Hilal Ozcanhan

Prediction of LDL in hypertriglyceridemic subjects using an innovative ensemble machine learning technique

https://doi.org/10.1515/tjb-2023-0154
Received July 12, 2023; accepted October 17, 2023;
published online November 21, 2023

Abstract

Objectives: Determining low-density lipoprotein (LDL) is a costly and time-consuming operation, but triglyceride value above 400 (TG>400) always requires LDL measurement. Obtaining a fast LDL forecast by accurate prediction can be valuable to experts. However, if a high error margin exists, LDL prediction can be critical and unusable. Our objective is LDL value and level prediction with an error less than low total acceptable error rate (% TEa).
Methods: Our present work used 6392 lab records to predict the patient LDL value using state-of-the-art Artificial Intelligence methods. The designed model, p-LDL-M, predicts LDL value and class with an overall average test score of 98.70 %, using custom, hyper-parameter-tuned Ensemble Machine Learning algorithm.
Results: The results show that using our innovative p-LDL-M is advisable for subjects with critical TG>400. Analysis proved that our model is positively affected by the Hopkins and Friedewald equations normally used for (TG≤400). The conclusion follows that the test score performance of p-LDL-M using only (TG>400) is 7.72 % inferior to the same p-LDL-M, using Hopkins and Friedewald supported data. In addition, the test score performance of the NIH-Equ-2 for (TG>400) is much inferior to p-LDL-M prediction results.
Conclusions: In conclusion, obtaining an accurate and fast LDL value and level forecast for people with (TG>400) using our innovative p-LDL-M is highly recommendable.

Keywords: Artificial Intelligence; LDL; machine learning; medical care; prediction methods

***Corresponding author: Ferhat Demirci**, MD, Clinical Biochemistry Laboratory, Dr. Suat Seren Chest Disease and Thoracic Surgery Training and Research Hospital, Yenisehir, 35110 Izmir, Türkiye; and Department of Neurosciences, The Institute of Health Sciences, Dokuz Eylul University, Izmir, Türkiye, Phone: +90 232 433 33 33/2357, E-mail: drferhat5505@hotmail.com. https://orcid.org/0000-0002-5999-3399
Murat Emec, Department of Computer Engineering, Faculty of Computer and Informatics, Istanbul University, Istanbul, Türkiye, E-mail: murat.emec@istanbul.edu.tr. https://orcid.org/0000-0002-9407-1728
Ozlem Gursoy Doruk and Pınar Akan, Department of Neurosciences, The Institute of Health Sciences, Dokuz Eylul University, Izmir, Türkiye; and Department of Biochemistry, Faculty of Medicine, Dokuz Eylul University, Izmir, Türkiye, E-mail: ozlemgursoy2008@gmail.com (O. Gursoy Doruk), pinar.akan@deu.edu.tr (P. Akan). https://orcid.org/0000-0002-1312-3777 (O. Gursoy Doruk). https://orcid.org/0000-0001-9211-1944 (P. Akan)
Murat Ormen, Department of Biochemistry, Faculty of Medicine, Dokuz Eylul University, Izmir, Türkiye, E-mail: murat.ormen@deu.edu.tr. https://orcid.org/0000-0003-4206-411X
Mehmet Hilal Ozcanhan, Department of Computer Engineering, Faculty of Engineering, Dokuz Eylul University, Izmir, Türkiye, E-mail: hozcanhan@cs.deu.edu.tr. https://orcid.org/0000-0002-5619-6722

Introduction

Clinical status in patients at cardiovascular risk depends on their low-density lipoprotein (LDL) levels [1]. Therefore, LDL levels are one of the most important target parameters in cholesterol-lowering treatment regimens worldwide. In biochemistry laboratory practice, LDL is usually a member of a group of tests called the lipid profile. Other tests in this panel are total cholesterol (TC), triglycerides (TG), and high-density lipoprotein (HDL). Beta quantification is the most commonly used ultracentrifugation-based method for LDL measurement [2]. Although it is the gold standard, new measurement methods have been developed because this method is time-consuming and expensive. Many manufacturers have developed calibrated kits validated by the gold-standard method compatible auto-analyzers. The auto-analyzer-supported test methods are based on homogeneous immunoassays. However, medical laboratories are units that adopt the principle of best cost-effective operation for fast and accurate test results. Therefore, scientists have started to develop new laboratory tests and methods, especially due to increasing health expenditures. The Friedewald equation (LDL=TC-(HDL + TG/5)) has been widely used in clinical practice for several decades.

Open Access. © 2023 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

The Friedewald equation calculates the LDL value using the patients' TC, HDL, and TG test results [3]. However, other studies have shown that LDL calculation using the Friedewald formula is inappropriate in cases with TG>400 mg/dL (High-TG) [4]. Consequently, new formulae for more accurate LDL calculations have been developed, and their results have been reflected in routine laboratory result reports. In particular, for patients with low LDL levels (LDL<70 mg/dL), the Martin-Hopkins equation (MH-E) equation (LDL=TC – HDL – TG/novel factor) has been considered the most successful [5, 6]. Later in 2020, Sampson M et al. planned a study in normolipidemic and/or hypertriglyceridemic patients and developed the equation (LDL=TC/0.948 – HDL/0.971 – (TG/8.56 + [TG × NonHDL]/2,140 – TG^2/16,100) – 9.44). Although NIH-Equ-2 has the best accuracy in the TG 400–800 mg/dL range, it is far from the results of direct measurement (Direct LDL), as it has an error up to 30 mg/dL at 800 mg/dL of TG [7]. Analysis of the above three formulae shows different superiorities over each other concerning TG and LDL-level relationship. Nevertheless, the NIH-Equ-2 is usually accepted as more successful than the Friedewald and MH-E in the TG 400–800 mg/dL range.

Currently, the LDL levels of patients with TG≤400 mg/dL (Low-TG) are calculated using the Friedewald formula in Turkish routine biochemistry laboratory tests. Meanwhile, in high-TG patients, the result is obtained by direct LDL measurement. Recently, machine learning (ML) and deep learning algorithms of artificial intelligence have started to take the place of formulae in medical science. In the literature, studies target patients with low TG that use ML. Naturally, the performances of ML-supported LDL calculation studies are compared with mainly the Friedewald and MH-E [8–13]. However, to our knowledge, there is currently no ML study targeting the LDL levels in high-TG patients. Very lately, ensemble ML techniques have been practiced in many health-related works and diabetes prediction [14, 15].

This study aims to develop a model using the Ensemble ML technique that predicts the patient LDL value and level classification using the age, gender data, and other non-LDL test results for high TG cases. The accuracy of our designed model and NIH-Equ-2 results will be evaluated by comparison to Direct LDL. The goal is not to exceed the total acceptable error rate (% TEa) in predicting the patient's LDL value and level.

Materials and methods

Study design

Our present research has been conducted at Dr. Suat Seren Chest Diseases and Thoracic Surgery Education and Research Hospital (Hospital). The medical participants of the research requested LDL value and level prediction software from the Computer Engineering members of Dokuz Eylul and Istanbul Universities. A total of 109,591 (high-TG: 6,404 + low-TG: 103,187) patient records collected at the Biochemistry Laboratory of the Hospital were provided. Records missing total cholesterol/triglycerides/HDL/LDL results, beyond the linearity limit of the specific assays, containing zero or negative valued results, records of patients <18 years of age, or records lacking numerical values were excluded from the study. The remaining 6,392 records (High-TG) of a total of 106,390 records (All-TG) were uploaded to a Phyton® software (Wilmington, Delaware, USA) and divided into a training subset (80 % of All-TG) and a test subset (20 % of All-TG), as in typical ML testing set-ups. Low-TG patient results were used only for training the algorithm. Our study results are only for high-TG patients. The TC, HDL, LDL, and TG records were obtained by measurements made using the Beckman Coulter (Brea, CA) AU-640 and Roche Cobas (Mannheim, Germany) c702 auto-analyzers.

A detailed literature research revealed the previous algorithms used in LDL prediction [8, 10, 11]. Our data analysis method was based on an investigation using state-of-the-art ensemble ML algorithms that were a subset of artificial intelligence. The candidate algorithms were based on tree-type algorithms that could detect linear or non-linear relationships between independent and dependent variables. Tests were conducted on the pre-conditioned data, and the best prediction performers were chosen to be included in the final prediction model. Performances were measured by comparing the predicted LDL value/class to the actual measured value/class.

Study population/subjects

Our study population consisted of the laboratory test results of 109,591 blood samples obtained at the Hospital's biochemistry laboratory between January 2010 and December 2022. The standard lipid profile data collected from the laboratory database included the patient's TC, TG, HDL, and the same-day measured LDL value. The baseline characteristics of our study population (High-TG) are shown in Table 1. While 3,431 subjects were male, the remaining 2,961 were female. The average age of the males and females were 49.72 and 54.07, respectively. The measured average Direct LDL was 149.76 ± 45.28 mg/dL.

The Standards for Reporting Diagnostic Accuracy diagram is shown in Figure 1. The diagram reports the flow of the subjects throughout the study. The total number was divided into two sets for statistical and ML analysis. The first set (training set) was created as 80 % of the total number of subjects, and the second set (test set) was created as 20 %. The first set was used for training the designed ML model, and the second set was used for testing the performance of the designed model. The training set was divided into two groups. The first group (n=105,111) contained LDL data with TG≤400 mg/dL and LDL values calculated with MH-E. The second group (n=5,113) contained direct LDL data with TG>400 mg/dL. In the p-LDL-M {1} model, the training set contained only the second group data, while in the p-LDL-M {2} model, both groups were included. The distribution of the test set of 1,279 subjects according to their LDL levels is shown in Figure 1. The classification is according to the 2019 ESC/EAS Guidelines for managing dyslipidemias [16]. The most undesired error in the LDL-level classification determination is putting a patient into a lower LDL level than the patient's under-classification. Therefore, under-classification is a very critical preference point in new model designs.

Lipid profile testing

All lipid profile parameters were measured on automatic chemistry analyzers Beckman Coulter AU 640 and Roche Cobas c702 of the Hospital's

Table 1: Characteristics of the study population (n=6392).

Characteristics		Value ± SD
Age, years		51.73 ± 11.61
Male		49.72 ± 11.20
Female		54.07 ± 11.61
Sex		
Male		3,431 (% 53.7)
Female		2,961 (% 46.3)
Total cholesterol	mg/dL	243.16 ± 52.79
	mmol/L	6.29 ± 1.37
Triglycerides	mg/dL	510.98 ± 96.71
	mmol/L	5.77 ± 1.09
HDL	mg/dL	37.57 ± 8.97
	mmol/L	0.97 ± 0.23
Non-HDL cholesterol	mg/dL	205.62 ± 48.12
	mmol/L	5.32 ± 1.24
Direct LDL	mg/dL	149.76 ± 45.28
	mmol/L	3.87 ± 1.17

biochemistry laboratory. Beckman auto analyzers were used for lipid profile measurements in our Hospital until 2017, and Roche auto analyzers afterward. The results of each patient's first test results were included in the study, and duplicate measurements were not included. TC was estimated using the enzymatic cholesterol esterase/oxidase method, while TG was estimated during the enzymatic glycerol phosphate oxidase method. HDL was measured by direct homogenous assays without precipitation, LDL was measured by a direct homogenous assay that uses a selective protective agent to separate LDL from chylomicrons, HDL, and VLDL, and then estimated by cholesterol esterase/oxidase method. The allowable % TEa for the LDL test for these measurement principles is 11.9 % [17]. All tests using Bio-Rad internal quality control material were performed using Beckman Coulter AU and Roche Cobas reagents and calibrators.

Statistical analysis

Assuming Direct LDL was the most accurate, it was compared to the corresponding predicted and calculated LDL values. Statistical analysis was performed using IBM® SPSS® Statistics 26 for Windows®.

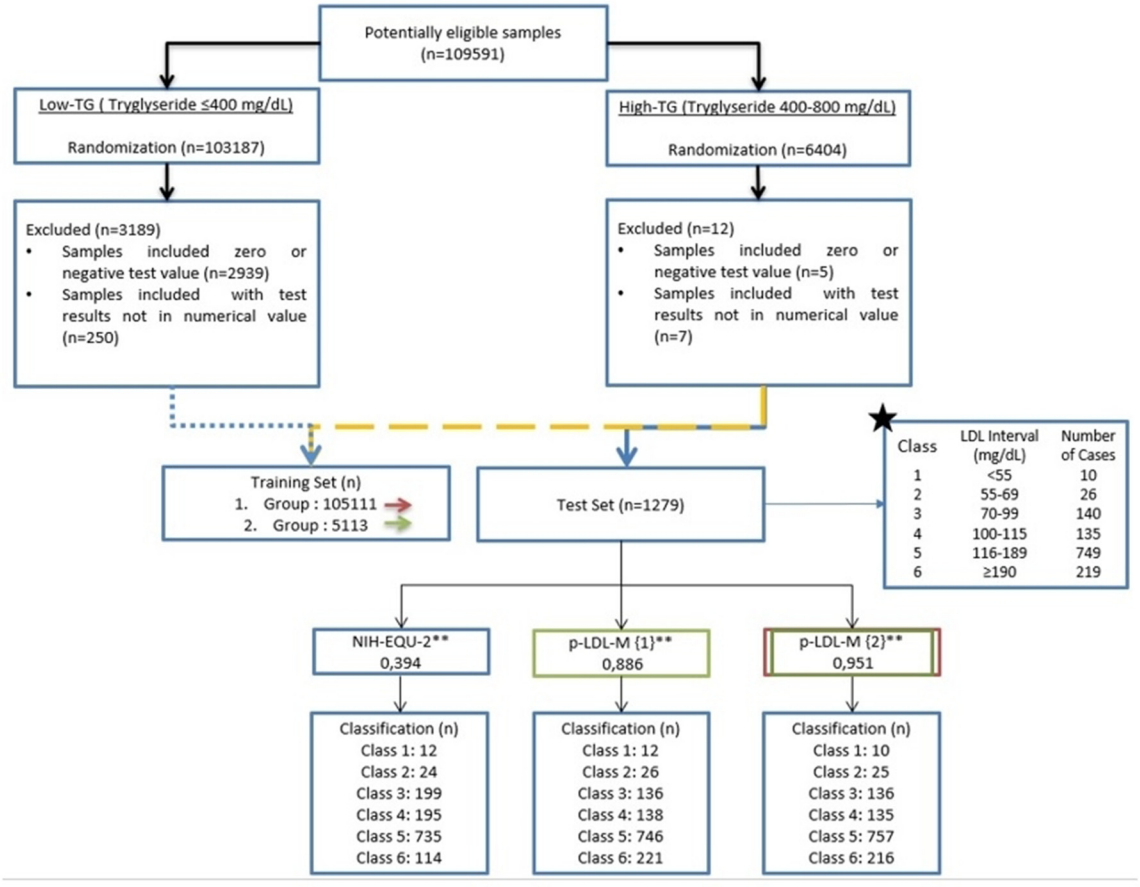


Figure 1: Standards for reporting diagnostic accuracy diagram showing the flow of the subjects through the study.

Paired t-test was used for the comparison of means. Pearson correlation test was performed to assess the correlation of Direct LDL with the results of the designed ML model and NIH-Equ-2. The p-value <0.05 was taken as statistically significant. Bland-Altman plots were used to assess systematic bias for different direct LDL concentrations. In the Bland-Altman plots, differences among the methods were plotted against direct LDL measurements.

We also assessed the performance of the designed model and NIH-Equ-2 in classifying the LDL levels as per the 2019 ESC/EAS Guidelines. If the subject was correctly classified into the same treatment class by the model and NIH-Equ-2, it was taken as correctly classified. The percentage of subjects classified correctly into appropriate treatment classes was calculated and compared. Cohen's kappa score was used to check the accuracy of the designed model and NIH-Equ-2 classifications. The Kappa result can be interpreted as follows: values≤0 indicating no agreement, 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement [18].

A universal method of model performance measurement was the statistical calculation of the accuracy, precision, and recall values. The definitions of the parameters used in the accuracy, precision, and recall calculations are as follows:

- True Positive (TP): the number of cases in the subject's LDL class is correctly identified.
- False Positive (FP): the number of cases in which the subject's LDL class is incorrectly identified.
- True Negative (TN): the number of cases that subjects out of an LDL class is correctly identified. Not applicable.
- False Negative (FN): the number of cases that subjects out of LDL class (1–6) is incorrectly identified.

Machine learning analysis

In our study, Python version 3.9 was used as the software programming language. The Pandas software library (Version: 1.4.4) in Python was used for data processing and analysis. The NumPy library (Version: 1.21.5) was used for the Python programming language, which supports large, multi-dimensional arrays and provides high-level mathematical functions for working with these arrays. The Sklearn library (Version: 1.0.2) was used to train machine learning models. The SHAP library (Version: 0.42.1) was used to feature importance. Lastly, the XGBoost library (Version: 1.5.0) was used to train and test the XGBoost algorithm.

In this part, the first step was the determination of Pearson and Spearman correlation coefficient matrices. Sometimes, due to outliers, these two matrices disagree on the strength of the correlation between an independent and dependent variable [19]. Therefore, both were included in the first step of statistical investigation. In the present study, there was no correlation strength disagreement between the two matrices, but only minor differences of less than 1.000. The results and implications of the matrices are in Figure 2. The figure summarizes the relationship between the different parameters involved in our research. A correlation higher than 0.7 is an indication of a strong relationship. A correlation higher than 0.85 is an indication of a very strong relationship [19].

The SHapley Additive exPlanations (SHAP) graph can determine each feature's significance in our proposed model. The SHAP values provide the contribution of each feature to every individual prediction in a proposed model. In other words, SHAP values quantify the impact of each feature on the model's predictions by considering all the possible feature combinations. The SHAP graph and the Pearson correlation results of our model were provided in the discussion section.

The step in ensemble ML analysis was the regression investigation of the High-TG subset. Seven ML algorithms (Random Forest, Decision Tree, XGBoost, Linear Regression, KNN, AdaBoost, SVR) were used to test the linearity of pre-processed data. After verifying the linearity of the High-TG dataset, the LDL values of the patients with Low-TG were calculated using the MH-E and Friedewald equations. After preliminary analysis showed that MH-E performed better in predicting LDL, a dataset was obtained by calculating low-TG LDL. A new dataset, All-TG, was formed by combining the Low-TG subset with the High-TG dataset. The linearity of All-TG was also verified.

In the third step, test scores of the seven individual ML algorithms were determined for the High-TG subset and the All-TG data set. The "bucket of models" is an ensemble technique that tests models to choose the best-performing model for the problem. Avoiding intermediate results, Random Forest and XGBoost were the best performers in both data sets, taking first place and second place. Their test scores were around 1 % apart.

A Machine Learning Model was designed in the next step, as shown in Figure 3. The model has been named "**p-LDL-M**" because our model predicts the LDL value using Artificial Intelligence rather than measuring it. For the High-TG dataset, the p-LDL-M was tested using both Random Forest and XGBoost algorithms (Table 2: p-LDL-M {1}). After observing the XGBoost algorithm's slightly better performance (around 1 %), the designed model was finalized by hyper-parameter tuning of the used XGBoost algorithms. Next, p-LDL-M prediction was conducted using the All-TG dataset (Table 2: p-LDL-M {2}). In the final step, the predicted LDL values were placed in their LDL-level classes, separately for p-LDL-M {1} and p-LDL-M {2}, as shown in Figure 3.

Results

Basic statistics results

The Pearson correlation basic statistical results of our study are presented in Figure 2. According to Figure 2, there was a very strong relationship between LDL and TC, LDL and Non_HDL, also between TC and Non_HDL. It is also clear that NIH-Equ-2 can calculate LDL value fairly correctly. Surprisingly, the rest of the parameters have moderate or weak relationships.

The basic statistical accuracy, precision, and recall results of our proposed p-LDL-M {2} model classification are presented. The total number of subjects is 1,279. But first, the p-LDL-M {2} model's statistical TP, FP, TN, and FN findings are given below, using Supplementary Table 1. Since our classification was not about identifying a subject's "not belonging to a class," True Negative was considered non-applicable to our results.

- p-LDL-M {2} true positive: 12, 24, 131, 127, 730, 218 → total: 1,242.
- p-LDL-M {2} false positive: 0, 2, 2, 4, 6, 0 → total: 14.
- p-LDL-M {2} true negative: not applicable.
- p-LDL-M {2} false negative: 0, 0, 3, 7, 10, 3 → total: 23.

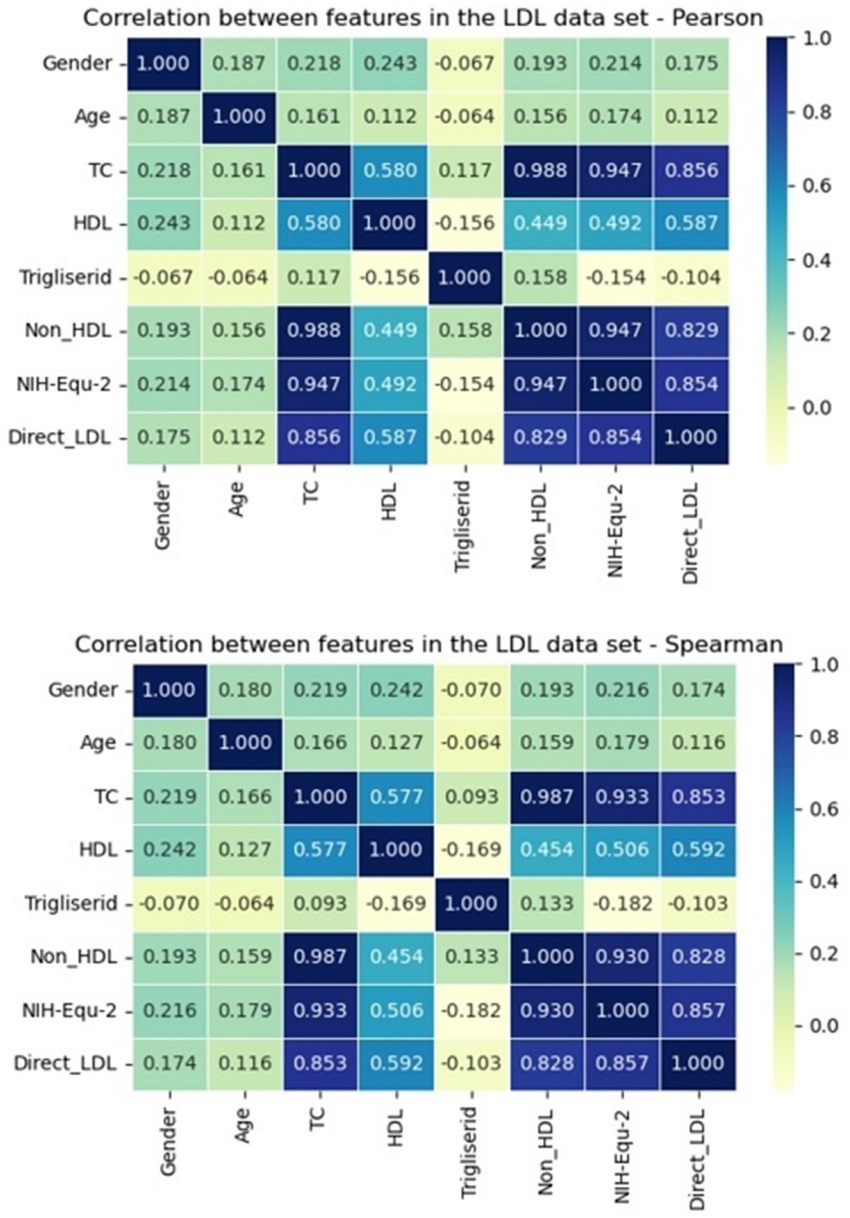


Figure 2: Correlation matrices of LDL-related parameters and NIH-Equ-2.

Using the universal accuracy and precision equations, we get:

$$\text{Accuracy of p} - \text{LDL} - \text{M}\{2\} = (\text{TP} + \text{TN}) \times / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) = (1,242/1,279) = 97.11\%$$

$$\text{Precision of p} - \text{LDL} - \text{M}\{2\} = (\text{TP})/(\text{TP} + \text{FP}) = (1,242/1,256) = 98.89\%$$

$$\text{Recall of p} - \text{LDL} - \text{M}\{2\} = (\text{TP})/(\text{TP} + \text{FN}) = (1,242/1,265) = 98.18\%$$

The accuracy, precision, and recall results of our recommended p-LDL-M {2} model classification are all above 97 %. The results of the other two models have not been included to save the reader's focus on our recommended model due to obvious inferior performances.

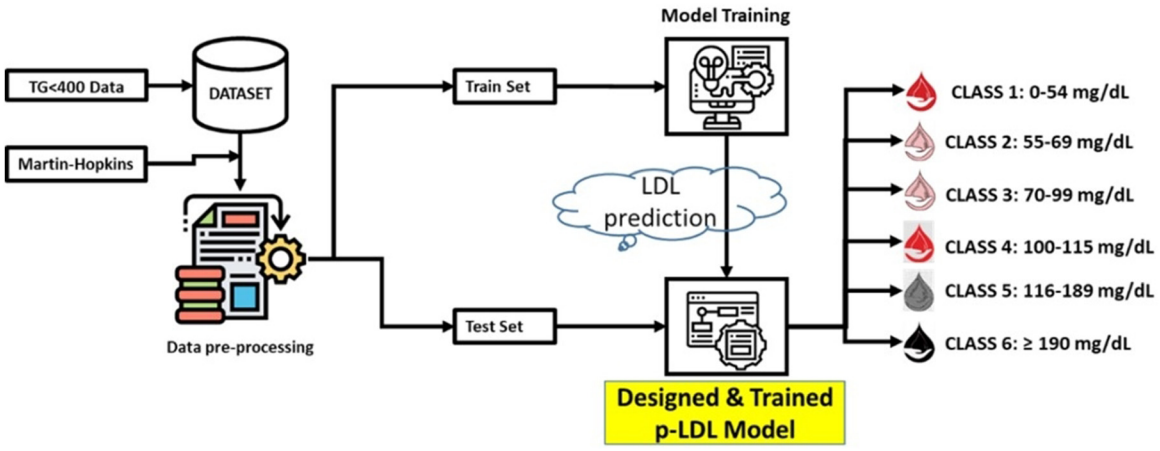


Figure 3: Our proposed LDL prediction model: p-LDL-M.

Table 2: Description of model abbreviations according to data sets.

Model	Training set[a]	Test set
p-LDL-M {1}	80 % records of High-TG subjects only	20 %
p-LDL-M {2}	80 % of (records of Low-TG subjects with LDL calculated using MH-E[b] + direct LDL of High-TG subjects)	20 %
NIH-Equ-2	Records of High-TG subjects	20 %

[a]Training set includes: age, sex, total cholesterol, triglyceride (TG), HDL and Direct LDL. [b]MH-E, Martin–Hopkins equation.

Machine learning results

The ML analysis results are shown in Table 3. The first six rows show the test score of predicting the LDL values in that specific LDL-level Class only. The last row uses the entire dataset to show the LDL value prediction for all classes. For the High-TG subset, the test score for predicting LDL values in Class 1 was the lowest at 74.31 % and highest in Class 4 (95.77 %). In all classes, prediction performance was 91.14 %. The rest of the scores were all above 90 %.

Table 3: p-LDL-M {1} and p-LDL-M {2} test results.

Class	LDL interval, mg/dL	Number of cases	MAE[a]	MSE[b]	MAPE[c]	Balanced accuracy, %	F1-score, %
			p-LDL-M {1}				
1	<55	12	8.22	1.23	25.69	91.67	90.91
2	55–69	26	3.86	0.20	6.23	94.11	88.46
3	70–99	136	6.18	0.58	7.19	94.48	89.13
4	100–115	138	4.57	0.29	4.23	91.20	84.98
5	116–189	746	8.97	1.33	6.07	92.00	93.75
6	≥190	221	10.60	1.91	4.86	90.84	87.98
	All classes	1,279	7.06	0.92	9.04	92.01	91.18
			p-LDL-M {2}				
1	<55	12	0.44	0.03	1.33	100.00	100.00
2	55–69	26	0.23	0.01	0.36	98.03	90.57
3	70–99	136	0.28	0.02	0.33	97.77	94.93
4	100–115	138	0.27	0.01	0.25	95.49	91.70
5	116–189	746	1.06	0.17	0.72	98.27	98.45
6	≥190	221	11.59	2.50	5.14	99.04	97.98
	All classes	1,279	2.31	0.45	0.96	98.02	97.12

[a]MAE, mean absolute error. [b]MSE, mean squared error. [c]MAPE, mean absolute percentage error.

The first general observation of Table 3 was that p-LDL-M {2} performs better with the All-TG dataset, which the MH-E supported. All test scores were above 94 %, the lowest being 94.86 % in Class 6 and the highest again in Class 4 with 99.75 %. When using all 106,390 patient records, the test score was 99.04 % when all classes were considered. A performance increase of almost 8 % over p-LDL-M {1}. Accordingly, all p-LDL-M {2} prediction errors were lower at all three error types.

The LDL level class prediction results are shown in Figure 1. Compared to the actual distribution, the predicted Class distributions of p-LDL-M {2} look similar for the p-LDL-M {1} model. However, the classification predictions were further detailed in Table 5. In this table, the over and under-classification errors were shown.

Comparison of model predicted and NIH-Equ-2 results with direct LDL

A study of the mean of the predicted and calculated LDL values by a paired t-test showed that all of the means significantly agreed with the mean of Direct LDL. However, the mean LDL values calculated by NIH-Equ-2 were distinguishably lower than the mean of Direct LDL and the other models. There was a 12.72 mg/dL difference between the mean of the values calculated by the NIH-Equ-2 (135.22 mg/dL) and the mean of the Direct LDL values (147.94 mg/dL). This difference was 8.60 % of the Direct LDL value and smaller than the LDL test TEa (11.9 %).

The Pearson correlation analysis between the predicted NIH-Equ-2 and Direct-LDL results showed very strong correlations. While the correlation of predicted results was >0.950, the correlation score of the NIH-Equ-2 was 0.865 (Table 4).

Analysis showed that the NIH-Equ-2 results were scattered, and the R^2 value was low (R^2=0.7483). The p-LDL-M {2} produced the best results, with an R^2 value close to 1 (R^2=0.9905) and a low scatter. The p-LDL-M {1} model had a slightly high scatter, with 0.9231 R^2 value. The scattering of all competitors can be seen in Figure 4a. In the Bland Altman plot, it can be seen that most of the measurements are above the mean line and are mostly within the 95 % confidence interval. It can be mentioned that the values that do not fall within the 95 % confidence interval from the results produced by our model tend to be high (Figure 4b). The receiver operating characteristic (ROC) curves of the predicted six classes in the p-LDL-M {2} model are shown in Figure 4c. The area under the ROC curve (AUC) indicates a model's performance across all possible classification thresholds. A value of more than 0.9 is considered an outstanding performance. Our obtained ROC curves indicate a 98 % AUC value success rate in five classes. The AUC value of Class 3 is the lowest (95 %). Therefore, the average AUC value proves that our proposed model has good prediction accuracy and precision performance across all classes.

Table 4: Comparison of LDL predicted by p-LDL-M, and NIH-Equ-2 with Direct LDL (n=1,279).

Algorithms	Mean + SD, mg/dL	Correlation, r	p-Value	Cohen's kappa score
Direct LDL	147.94 ± 45.03			
p-LDL-M {1}	147.75 ± 45.41	0.961	0.590	0.886[b]
p-LDL-M {2}	147.96 ± 44.63	0.995	0.873	0.951[b]
NIH-Equ-2	135.22 ± 41.01	0.865	<0.005	0.394[a]

[a]Fair agreement. [b]Almost perfect agreement.

The class of LDL levels of patients is as important as the quantitative value of LDL in lipid-lowering therapies. Clinicians apply variable treatment options ranging from dietary changes and exercise to multi-drug therapies, depending on the patient LDL level category. Therefore, the LDL values under study were categorized into classes according to the 2019 ESC/EAS guideline. As shown in Figure 1, class 1 (0–54 mg/dL) had the least number of cases (10 cases), while class 5 (116–189 mg/dL) had the highest number (749 cases).

The kappa scores were obtained (Table 4) by comparing the actual LDL levels with the predicted and calculated LDL levels. The lowest kappa score was 0.394 for the NIH-Equ-2's LDL level classification, while the highest was for the p-LDL-M {2} (0.951). The p-LDL-M {2}, p-LDL-M {1}, and the NIH-Equ-2 classification according to the Direct-LDL results were given in Table 5. It shows that the p-LDL-M {2} misclassifications were 38 (2.97 %), the p-LDL-M {1} misclassifications were 88 (6.78 %), and the NIH-Equ-2 misclassifications were 476 (37.2 %).

Discussion

Our study aims at LDL value and level prediction of High-TG people based on measured and calculated values. According to our literature review, our study is the first for high-TG subjects. Below, the most important findings of our study are discussed.

The mean value of the used datasets is important in giving a general idea about the studied population. When the lipid profiles of the population included in our study were analyzed, the mean values of TC, TG, and LDL values were found to be high, while the mean HDL values were found to be low compared to other related ML studies [8–11]. The dietary habits in our country may explain the differences. However, except for the TG results, the lipid profile values of the National Institutes of Health center were similar to our results in

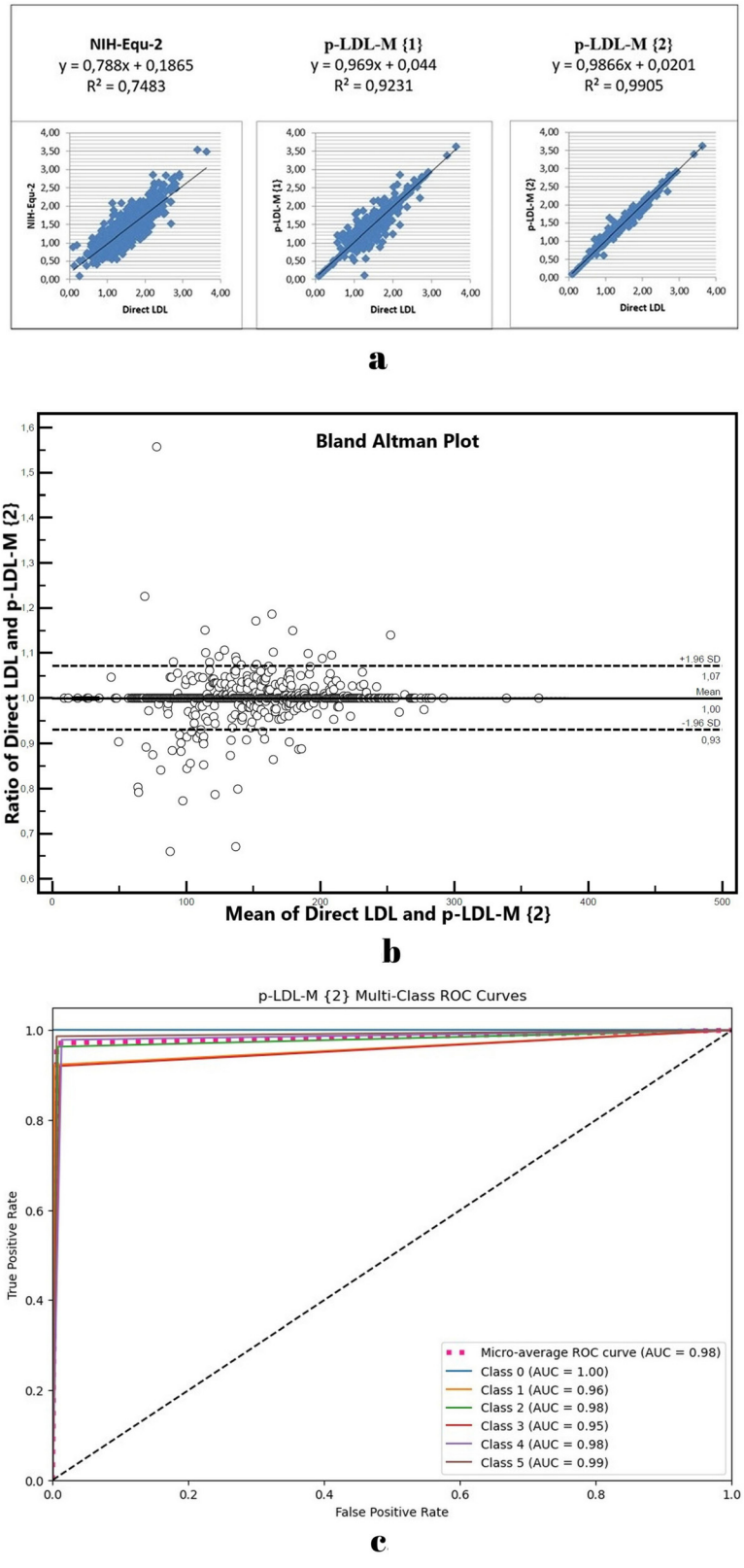


Figure 4: Statistical performance characteristics of our model. (a) Scatter plots of correlations between predicted, calculated, and direct LDL. (b) Bland–Altman plot between direct LDL and p-LDL-M {2}. (c) ROC curves of p-LDL-M {2} model class predictions.

the multicenter study by Sampson M. et al. The lipid results from the other four centers were lower than our study [7].

In previous Low-TG LDL prediction studies, Random Forest has been the most preferred ML algorithm, but XGBoost, Deep Neural Network (DNN), Support Vector Machine (SVM), Linear Regression, and k-Nearest Neighbors (KNN) methods have also been used [8, 10]. In our study, Ensemble ML techniques proved XGBoost as the preferred algorithm, as it provided the best prediction results.

Our p-LDL-M predictions in patient LDL value and LDL-level class have produced interesting results. Comparison of the LDL value prediction accuracy of our p-LDL-M with the direct LDL measurement results showed that our model's prediction errors did not exceed the total acceptable error rate (%TEa) and obtained the highest test score (99.04 %), the lowest MAE, MSE and MAPE results. Interestingly, the best performance was obtained by using All-TG and not by using measured High-TG only. Applying the Martin-Hopkins equation to Low-TG and using the calculated LDL values increased the prediction performance by a 7.90 % margin over predictions when only High-TG data was used. The p-LDL-M prediction results outperformed the famous NIH-Equ-2 results by a margin of 8.60 %. In addition, LDL level classification results had better performance and the least scatter when All-TG was used.

The study clearly shows the positive effect of using the Martin-Hopkins equation to calculate LDL values. Furthermore, our study confirmed the following:

- The previously known strong correlations between TC, Non-HDL, and Direct LDL,
- The good performance of NIH-Equ-2 in calculating LDL (Figure 2).
- The success of the Random Forest algorithm in estimating LDL values.
- The linear relationship between TG and LDL.

A thorough statistical analysis of the p-LDL-M results has proven the strong performance of our designed model. The p-LDL-M predicted, and direct LDL measurement results were significantly correlated (0.995), while the correlation of the NIH-Equ-2 results was lower (0.865). The algorithm results of Anudeep PP. and Singh G. et al. were also significantly correlated with the direct-LDL measurements (0.98 and 0.982) [9, 10]. The above strong correlation results indicate that ML algorithms can predict LDL values better than previously developed formulae [20]. Another interesting result was the difference between the mean value of the NIH-Equ-2 results and the direct LDL mean (147.94 ± 45.03, 135.22 ± 41.01 mg/dL). In contrast, the mean values obtained in our model and the other compared methods had no significant mean differences. Our p-LDL-M {2} model's superior statistical performance is further supported by the accuracy, precision, and recall values in the results sections.

The SHAP graph of our model is shown in supplementary material. The highest impacting feature is Non_HDL. This is validated by the highest Pearson correlation value, 0.8229 in Figure 2. Triglyceride's negative −0.104 Pearson correlation value is observed in the minus section of the SHAP graph. The protruding line in the positive section depicts the natural correlation between cholesterol and LDL levels.

The scatter plot of p-LDL-M results is shown in Figure 4a. While our results' R^2 value is 0.9905, the R^2 value of the NIH-Equ-2 formula is 0.7483. The graphs show that our model results are almost linear, in contrast to the scattered results of NIH-Equ-2. NIH-Equ-2 has more scatter with increasing TG levels. Our scatter performance is again in agreement with Anudeep PP. et al.'s low scatter study. Our study also agreed that different formulae can produce negative results, even though their correlation values (r) vary between 0.89 and 0.94 [10].

Numerous epidemiological, mendelian randomization, and randomized controlled trials have shown that atherosclerotic cardiovascular diseases (ASCVD) and LDL have a log-linear relationship. Therefore, according to previous guidelines and committees, lipid-lowering therapies are of great importance for ASCVD-related survival. The lipid-lowering therapeutic interventions have been demonstrated by basic science, clinical observations, genetic factors, randomized controlled trials, and epidemiological studies [21–23]. It is also known that LDL levels determine the type and dose of cholesterol-lowering therapies to be administered to patients. One study demonstrated that a 1 mmol/L reduction in LDL levels can reduce major cardiovascular events by 20 % (24). In some cases, the p-LDL-M has overestimated LDL levels by one class (Table 5). However, we believe there is little danger of overestimating LDL levels by one class because ezetimibe and monoclonal antibody therapies added to statins lower LDL levels more, thus improving cardiovascular outcomes. It has been demonstrated that extra ezetimibe in statin therapy reduced the risk of possible cardiovascular events but did not cause any side effects or toxic effects, even in individuals with acute coronary syndromes and desirable LDL levels. As a result, the survival rate has improved with decreased LDL levels. In addition, there is no study on the harm of ezetimibe addition in statin therapy at medium LDL levels [25, 26].

The costs of lipid-lowering therapies to healthcare providers are significant worldwide [27]. Although beta quantification is a reference method, it is unsuitable for routine use due to its high cost and effort [28]. Therefore, beta quantification has been replaced by enzymatic and homogeneous immune methods. After the correlation of LDL with other

cholesterol derivatives was demonstrated, the direct measurement was replaced by formulae such as Friedewald, Martin Hopkins, and NIH-Equ-2. The aforementioned formulae have different advantages over each other. While the Martin Hopkins and Friedewald formulae are successful at TG levels <400 mg/dL, the NIH-Equ-2 can calculate results up to 800 mg/dL. More remarkably, since the Martin Hopkins formula uses different factors for different TG levels, it can be used even in non-fasting people [29]. Recently, research eliminating the disadvantages of used formulae has focused on ML LDL prediction models. According to our literature review, our study is the first to predict LDL levels using ML algorithms for cases with TG between 400 and 800 mg/dL. Our work becomes important when clinicians need accurate LDL level classifications for appropriate treatment [16]. To demonstrate the classification accuracy of our model, we calculated the kappa score of our classification results. The kappa analysis revealed that the kappa score of our p-LDL-M model was significantly higher than the NIH-Equ-2 (0.951 and 0.394, respectively). According to the kappa score, the p-LDL-M model was in "almost perfect agreement" with the direct LDL measurements, while NIH-Equ-2 was evaluated as "fair agreement".

Literature shows that previous models and formulae have larger prediction error margins at low LDL levels (<70 mg/dL). Our p-LDL-M model has an 8.3 % (3/36) classification error, while NIH-Equ-2 has 44.4 % (16/36), as shown in Table 5. The same error rate was 7.5 % in the Weill-Cornell model (9) [9]. According to the 2019 ESC/EAS Guidelines, our model's error rate was 6.8 % (12/176) in the first three classes (LDL<100 mg/dL). The same error rates were 43.75 % (77/176) in the NIH-Equ-2 formulae, 11.4 % (143/1,254) in the Weill-Cornell model, and 3.47 % (53/1,528) in the Anudeep PP et al. model [9, 10]. The study by Baraket-Hamade V. et al. divided the LDL cholesterol levels into three categories, the lowest being LDL<80 mg/dL. The error rate of misclassification in the mentioned lowest category was 12.5 % (793/6327) [11].

It is known that classifying an LDL value to a lower class (under-classification) than it actually may prevent the patient from receiving the appropriate treatment. On the other hand, putting LDL levels in the correct or a higher class (over-classification) will result in the appropriate or more intensive treatment. This situation may only lead to some more patients receiving treatment and some more treatment side effects. However, considering the morbidity and mortality statistics of cardiovascular disease patients due to high LDL levels, intensive hyperlipidemia treatment can positively affect the quality of life of over-classified patients [24, 30, 31]. Table 5 shows that The p-LDL-M{2}had 15 under-classified cases (1.17 %, n=15), while The p-LDL-M{1} had 42 (3.28 %, n=42) and NIH-Equ-2 had (27.7 % n=355). More strikingly, in the p-LDL-M {2}algorithm, none of the patients were misclassified under two classes. Furthermore, while there were only three misclassifications for classes 1 and 2 in prediction models, there were 23 misclassifications in NIH-Equ-2.

In classes that include patients with more risk factors (classes 1–3), our model placed some subjects incorrectly in a higher class. As a consequence, such misclassified patients would receive more intensive treatment, which is clinically less critical than inadequate treatment [4]. Our study has the lowest sample size in the first 3 clusters among all studies. We believe that our study's relatively large classification error is due to the low number of samples available, and it will decrease with a larger number of samples.

Many publications are defending the success of NIH-Equ-2 LDL value estimation [32, 33]. It should be noted that the results of patients with TG<400 mg/dL were also used in the development of NIH-Equ-2. In addition, NIH-Equ-2 was based on LDL levels estimated by the reference beta-quantification method. Although our model's accuracy was superior to NIH-Equ-2 accuracy, the LDL levels for patients with TG<400 mg/dL were obtained by direct homogeneous immune methods in our study. In addition, our study used results based on two different auto-analyzers, unlike many publications in the literature. Therefore, different accuracy rates in results may arise because of other methods, models, and auto-analyzers [34].

Our study has several limitations. First, since our data were obtained retrospectively and beta quantification is not used in laboratory practice, the more frequently preferred direct homogeneous immune method was used instead of the reference beta-quantification method. Secondly, the effects of diseases that may affect the lipid profiles could not be evaluated separately due to the study's retrospective design. Thirdly, ethnic group sub-categorization could not be made. Since the target TG range was >400 mg/dL, triglyceride values could not be subcategorized or analyzed in detail. Our study's sample size is lower than other studies in the literature because our Hospital is a chest disease hospital specializing in lung diseases. However, we believe the sample size is still sufficient to obtain statistically significant results. Due to the low number of participants, the model must be validated with larger datasets, even if the values produced are attributed to success.

Conclusions

As a result, the estimation model created by our study gives more accurate results for target TG values than the NIH-Equ-2 formula, considering the proportional relationship of very low-density lipoprotein (VLDL) with non-HDL cholesterol and TG. After validation of this model in more cases, it can take a place in the routine lipid profile. Its use as a clinical

support system to predict LDL levels in patients with high TG may be a safe and time-efficient/cost-effective solution for both laboratory professionals and clinicians in the future.

Research ethics: The study was approved by the University of Health Sciences, Dr. Suat Seren Chest Diseases and Thoracic Surgery Education and Training Hospital Ethics Committee (Resolution Number 2023/13-23 dated 12.04.2023)
Informed consent: Informed consent was obtained from all individuals included in this study.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Research funding: None declared.

References

1. Ference BA, Ginsberg HN, Graham I, Ray KK, Packard CJ, Bruckert E, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. Eur Heart J 2017;38:2459–72.
2. Ordóñez-Llanos J, Wägner AM, Bonet-Marqués R, Sánchez-Quesada JL, Blanco-Vaca F, González-Sastre F. Which cholesterol are we measuring with the Roche direct, homogeneous LDL-C plus assay? Clin Chem 2001; 47:124–6.
3. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 1972;18: 499–502.
4. Sathiyakumar V, Blumenthal RS, Elshazly MB. American College of Cardiology; 2020. [cited 2022 Dec 26]. New Information on Accuracy of LDL-C Estimation. Available from: https://www.acc.org/latest-in-cardiology/articles/2020/03/19/16/00/new-information-on-accuracy-of-ldl-c-estimation.
5. Martin SS, Blaha MJ, Elshazly MB, Toth PP, Kwiterovich PO, Blumenthal RS, et al. Comparison of a novel method vs the Friedewald equation for estimating low-density lipoprotein cholesterol levels from the standard lipid profile. JAMA 2013;310:2061–8.
6. Martin SS, Blaha MJ, Elshazly MB, Brinton EA, Toth PP, McEvoy JW, et al. Friedewald-estimated versus directly measured low-density lipoprotein cholesterol and treatment implications. J Am Coll Cardiol 2013;62:732–9.
7. Sampson M, Ling C, Sun Q, Harb R, Ashmaig M, Warnick R, et al. A new equation for calculation of low-density lipoprotein cholesterol in patients with normolipidemia and/or hypertriglyceridemia. JAMA Cardiol 2020;5:540.
8. Kwon YJ, Lee H, Baik SJ, Chang HJ, Lee JW. Comparison of a machine learning method and various equations for estimating low-density lipoprotein cholesterol in Korean populations. Front Cardiovasc Med 2022;9:824574.
9. Singh G, Hussain Y, Xu Z, Sholle E, Michalak K, Dolan K, et al. Comparing a novel machine learning method to the Friedewald formula and Martin-Hopkins equation for low-density lipoprotein estimation. PLoS One 2020;15: e0239934.
10. Anudeep PP, Kumari S, Rajasimman AS, Nayak S, Priyadarsini P. Machine learning predictive models of LDL-C in the population of eastern India and its comparison with directly measured and calculated LDL-C. Ann Clin Biochem 2022;59:76–86.
11. Barakett-Hamade V, Ghayad JP, Mchantaf G, Sleilaty G. Is Machine Learning-derived Low-Density Lipoprotein Cholesterol estimation more reliable than standard closed-form equations? Insights from a laboratory database by comparison with a direct homogeneous assay. Clin Chim Acta 2021;519:220–6.
12. Fan G, Zhang S, Wu Q, Song Y, Jia A, Li D, et al. A machine learning-based approach for low-density lipoprotein cholesterol calculation using age and lipid parameters. Clin Chim Acta 2022;535:53–60.
13. Palmer MK, Barter PJ, Lundman P, Nicholls SJ, Toth PP, Karlson BW. Comparing a novel equation for calculating low-density lipoprotein cholesterol with the Friedewald equation: a VOYAGER analysis. Clin Biochem 2019;64:24–9.
14. Ihnaini B, Khan MA, Khan TA, Abbas S, Daoud MSH, Ahmad M, et al. A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning. Comput Intell Neurosci 2021;2021:1–11.
15. Xie JC, Pun CM. Deep and ordinal ensemble learning for human age estimation from facial images. IEEE Trans Inf Forensics Secur 2020;15: 2361–74.
16. Mach F, Baigent C, Catapano AL, Koskinas KC, Casula M, Badimon L, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. Eur Heart J 2020;41: 111–88.
17. Westgard J. Desirable Specifications for Total Error, Imprecision, and Bias, derived from intra- and inter-individual biologic variation 2014. [Internet]. [cited 2023 Mar 4]. Available from: https://www.westgard.com/biodatabase1.htm.
18. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276–82.
19. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018;18:91–3.
20. Atabi F, Mohammadi R. Clinical validation of eleven formulas for calculating LDL-C in Iran. Iran J Pathol 2020;15:261–7.
21. Lloyd-Jones DM, Morris PB, Ballantyne CM, Birtcher KK, Daly DD, DePalma SM, et al. 2016 ACC expert consensus decision pathway on the role of non-statin therapies for LDL-cholesterol lowering in the management of atherosclerotic cardiovascular disease risk. J Am Coll Cardiol 2016;68:92–125.
22. Reiner Z, Catapano AL, De Backer G, Graham I, Taskinen MR, Wiklund O, et al. ESC/EAS guidelines for the management of dyslipidaemias: the task force for the management of dyslipidaemias of the European society of cardiology (ESC) and the European atherosclerosis society (EAS). Eur Heart J 2011;32:1769–818.
23. Catapano AL, Graham I, De Backer G, Wiklund O, Chapman MJ, Drexel H, et al. 2016 ESC/EAS guidelines for the management of dyslipidaemias. Eur Heart J. 2016;37:2999–3058.
24. Baigent C, Blackwell L, Emberson J, Holland LE, Reith C, Bhala N, et al. Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. Lancet 2010;376:1670–81.
25. Cannon CP, Blazing MA, Giugliano RP, McCagg A, White JA, Theroux P, et al. Ezetimibe added to statin therapy after acute coronary syndromes. N Engl J Med 2015;372:2387–97.

26. Harvey PD, Sabbagh MN, Harrison JE, Ginsberg HN, Chapman MJ, Manvelian G, et al. No evidence of neurocognitive adverse events associated with alirocumab treatment in 3340 patients from 14 randomized Phase 2 and 3 controlled trials: a meta-analysis of individual patient data. Eur Heart J 2018;39:374–81.
27. Greving J, Visseren F, de Wit G, Algra A. Statin treatment for primary prevention of vascular disease: whom to treat? Cost-effectiveness analysis. BMJ 2011;342:d1672–2.
28. Nauck M, Warnick GR, Rifai N. Methods for measurement of LDL-cholesterol: a critical assessment of direct measurement by homogeneous assays versus calculation. Clin Chem 2002;48:236–54.
29. Sathiyakumar V, Park J, Golozar A, Lazo M, Quispe R, Guallar E, et al. Fasting versus nonfasting and low-density lipoprotein cholesterol accuracy. Circulation 2018;137:10–9.
30. Ference BA, Yoo W, Alesh I, Mahajan N, Mirowska KK, Mewada A, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease. J Am Coll Cardiol 2012;60:2631–9.
31. Di Angelantonio E, Gao P, Pennells L, Kaptoge S, Caslake M, Thompson A, et al. Lipid-related markers and cardiovascular disease prediction. JAMA 2012;307:2499–506.
32. Vargas-Vázquez A, Bello-Chavolla OY, Antonio-Villa NE, Mehta R, Cruz-Bautista I, Aguilar-Salinas CA. Comparative assessment of LDL-C and VLDL-C estimation in familial combined hyperlipidemia using Sampson's, Martin's and Friedewald's equations. Lipids Health Dis 2021;20:46.
33. Piani F, Cicero AFG, Borghi C, D'Addato S. Is the 2020 Sampson equation the best formula for LDL-C estimation? Eur J Intern Med 2021; 83:99–101.
34. Ertürk Zararsız G, Bolat S, Cephe A, Kochan N, Yerlitaş Sİ, Doğan HO, et al. Validation of Friedewald, Martin-Hopkins and Sampson low-density lipoprotein cholesterol equations. PLoS One 2022;17:e0263860.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/tjb-2023-0154).