Abant Tıp Dergisi

Abant Medical Journal

Abant Med J 2023;12(1):63-75, doi:10.47493/abantmedj.1217478

Estimation of LDL-C Using Machine Learning Models and its Comparison with Directly Measured and Calculated LDL-C in Turkish Pediatric Population

Türk Pediatrik Popülasyonunda Makine Öğrenimi Modelleri Kullanılarak LDL-K Tahmini ve Doğrudan Ölçülen ve Hesaplanan LDL-K ile Karşılaştırılması

Necla KOÇHAN [1*]

[1*] İzmir Biomedicine and Genome Center (IBG), İzmir, Türkiye

Geliş Tarihi (*Received*): 11.12.2022 **Kabul Tarihi (*Accepted*):** 03.04.2023 **Yayın Tarihi (*Published*):** 28.04.2023

Abstract

Objective: The assessment of lipid profiles in children is critical for the early detection of dyslipidemia. Low-density lipoprotein cholesterol (LDL-C) is one of the most often used measures in diagnosing and treating patients with dyslipidemia. Therefore, accurate determination of LDL-C levels is critical for managing lipid abnormalities. In this study, we aimed to compare various LDL-C estimating formulas with powerful machine-learning (ML) algorithms in a Turkish pediatric population.

Materials and Methods: This study included 2,563 children under 18 who were treated at Sivas Cumhuriyet University Hospital in Sivas, Türkiye. LDL-C was measured directly using Roche direct assay and estimated using Friedewald's, Martin/Hopkins', Chen's, Anandaraja's, and Hattori's formulas, as well as ML predictive models (i.e., Ridge, Lasso, elastic net, support vector regression, random forest, gradient boosting and extreme gradient boosting). The concordances between the estimates and direct measurements were assessed overall and separately for the LDL-C and TG sublevels. Linear regression analyses were also carried out, and residual error plots were created between each LDL-C estimation and direct measurement method.

Results: The concordance was approximately 0.92-0.93 percent for ML models, and around 0.85 percent for LDL-C estimating formulas. The SVR formula generated the most concordant results (concordance=0.938), while the Hattori and Martin-Hopkins formulas produced the least concordant results (concordance=0.851).

Conclusion: Since ML models produced more concordant LDL-C estimates compared to LDL-C estimating formulas, ML models can be used in place of traditional LDL-C estimating formulas and direct assays.

Keywords: Cardiovascular Diseases, Cholesterol, Lipoproteins, Low-Density Lipoprotein, Machine Learning

&

Öz

Amaç: Çocuklarda lipid profillerinin değerlendirilmesi, dislipideminin erken saptanması için kritik öneme sahiptir. Düşük yoğunluklu lipoprotein kolesterol (LDL-K), dislipidemik hastaların teşhis ve tedavisinde en sık kullanılan ölçümlerden biridir. Bu nedenle, LDL-K düzeylerinin doğru belirlenmesi, lipid anormalliklerinin yönetimi için kritik öneme sahiptir. Bu çalışmada, Türk pediatrik popülasyonunda çeşitli LDL-K tahmin formüllerini güçlü makine öğrenmesi algoritmalarıyla karşılaştırmayı amaçladık.

Gereç ve Yöntemler: Bu çalışmaya Sivas Cumhuriyet Üniversitesi Hastanesi'nde tedavi gören 18 yaş altı 2,563 çocuk dahil edildi. LDL-K değerleri Roche direkt yöntemi kullanılarak ölçüldü ve Friedewald, Martin/Hopkins, Chen, Anandaraja ve Hattori formülleri ile makine öğrenmesi modelleri (Ridge, Lasso, elastic net, destek vektör regresyonu, rastgele orman, gradyan artırma ve aşırı gradyan artırma) kullanılarak tahmin edildi. Tahminler ve direkt ölçümler arasındaki uyum hem genel olarak hem de LDL-K ve TG alt seviyeleri için ayrı ayrı değerlendirildi. Ayrıca, doğrusal regresyon analizleri gerçekleştirilmiş olup her bir LDL-K tahmini ile direkt ölçüm yöntemi arasındaki fark artık hata grafikleri ile gösterilmiştir.

Bulgular: Tahminlenen LDL-K değerleri ile Roche direkt metodu ile ölçülen LDL-K değerleri arasındaki uyum, makine öğrenmesi modelleri için yaklaşık yüzde 0,92-0,93 ve LDL-K tahmin formülleri için yaklaşık %0,85 idi. Destek vektör regresyonu en uyumlu sonuçları (uyum=0,938) verirken, Hattori ve Martin-Hopkins formülleri en az uyumlu sonuçları (uyum=0,851) vermiştir.

Sonuç: Makine öğrenmesi modelleri, LDL-K tahmin formüllerine kıyasla daha uyumlu LDL-K tahmini yaptığından, makine öğrenmesi modelleri, geleneksel LDL-K tahmin formülleri ve doğrudan analizlerin yerine kullanılabilir.

Anahtar Kelimeler: Kardiyovasküler Hastalıklar, Kolesterol, Lipoproteinler, Düşük-Yoğunluklu Lipoprotein, Makine Öğrenimi

Atıf/Cite as: Koçhan, N. (2023). Estimation of LDL-C using machine learning models and its comparison with directly measured and calculated LDL-C in Turkish pediatric population. Abant Tıp Dergisi, 12 (1) , 63-75 . DOI: 10.47493/abantmedj.1217478

[*]**Sorumlu Yazar (Corresponding Author):** Necla Kochan, **e-mail**: necla.kayaalp@gmail.com

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide. Dyslipidemia, a condition caused by abnormal lipoprotein metabolism, is a significant risk factor for CVDs. It produces atherosclerotic lesions in children ages 2 to 16, leading to an elevated risk of CVDs later in their lives (1,2). Indeed, fatty lines and other early symptoms of atherosclerosis have been observed in children as young as two years old (1). Therefore, if the symptoms of CVD development are detected in childhood, CVDs can be prevented later in their adulthoods with earlier interventions, such as lowering lipid levels or regulating lipid status, which have been shown as effective interventions not only for primary but also for secondary preventions (3–5).

The serum level of low-density lipoprotein cholesterol (LDL-C) is one of the most critical markers used for CVD risk assessment (6,7). The gold standard for LDL-C measurement is β-quantification, where lipoprotein particles are separated by ultracentrifugation (8). However, β-quantification is inappropriate for routine use because it is costly, time-consuming, necessitates a large cohort and some specialized tools (9–12). Therefore, except in a few laboratories, the use of this method has remained limited (12) and other direct methods or various equations such as Friedewald and Martin/Hopkins equations are mostly preferred to measure LDL-C levels.

LDL-C estimate using the Friedewald formula was a brand-new method that was introduced in 1972. Due to its benefits over β-quantification methods, including cost effectiveness and time savings, it established a new norm in clinical guidelines worldwide (9). This method, despite its extensive use in clinical laboratories, has several limitations. First, Friedewald uses TG/5 formula (TG: triglycerides) to calculate very-low-density lipoprotein cholesterol (VLDL-C), which does not always provide an accurate estimate. Second, the Friedewald formula necessitates fasting serum for accurate LDL-C estimation. The chylomicrons present in the fed state lead to an overestimated VLDL-C (8). The use of Friedewald's equation is not recommended in the presence of high TG concentrations (TG>400) and type III hyperlipidemia. The third limitation is that in cases where LDL-C is <70 and TG ≥150 mg/dL, it may lead to underestimation of LDL-C and inadequate treatment of patients (13). To overcome such limitations, others have developed new formulas. Martin/Hopkins formula is one of those formulas that estimates an adjustable ratio using triglyceride and non-high-density lipoprotein (non-HDL-C) levels (14). Thus, with a more accurate estimate of VLDL-C, the LDL-C calculation is assumed to be more reliable. The use of the Martin/Hopkins formula was recommended in 2018 by the American College of Cardiology and American Heart Association guidelines for those with low LDL-C due to the advantages of the formula (15). Additionally, other formulas such as Chen (16), Anandaraja (17), and Hattori (18) have been developed. Even though these formulas have been validated in various populations, additional research is still required (19). Recently, machine learning (ML) methods such as gradient boosting, random forest, and support vector machines, etc., have been investigated for LDL-C estimation in different populations (19–22). However, the literature on the pediatric population particularly in Turkish pediatric population regarding LDL-C estimation is limited, and while ML approaches are employed, their application to the pediatric population is limited, as well.

Therefore, the purpose of this study was to examine the validity of the LDL-C levels estimated by various LDL-C estimating formulas (i.e., Friedewald, Martin-Hopkins, Chen, Anandaraja, Hattori) and powerful ML algorithms with the directly measured LDL-C levels by Roche commercial kits in the pediatric population of Türkiye.

Materials and Methods

Study Population

The study was conducted in a retrospective design and the study cohort consisted of 2,356 lipid profile samples collected between March 3, 2011, and December 31, 2019. The LDL-C, HDL-C, TG, and total

cholesterol (TC) lipid profiles that were directly measured between these dates were analyzed. The study was carried out in compliance with the Declaration of Helsinki and followed Good Clinical Practice guidelines. This study was approved by Sivas Cumhuriyet University of medical sciences with document no: 2022-06/02 (date: 22.06.2022).

Lipid Measurements

Roche (Mannheim-Germany) Cobas 8000, c-702 and c-501 modules were utilized to directly measure HDL-C, LDL-C, TG and TC. The colorimetric enzymatic reaction is used for all other measures. The analytical coefficient of variation (CV) for each parameter on each module was calculated by using two-level internal quality control results over a one-year period and pooled CV results were established using these data. The bias for each parameter on each module was given as the average absolute % deviation determined from the external quality assessment over the same year period. Bias and CV values for Cobas c501 were found as follows, respectively; 3.34 and 3.49 for TC; 2.77 and 3.92 for HDL-C; 3.51 and 2.85 for LDL-C; 3.25 and 2.69 for TG. Bias and CV values for Cobas c702 were found as follows, respectively; 4.0 and 3.29 for TC; 3.05 and 2.90 for HDL-C; 3.63 and 3.58 for LDL-C; 3.30 and 2.86 for TG.

Lipid Estimations by Formulas and Machine Learning Models

To estimate LDL-C concentration, we used both traditional LDL-C estimating formulas and ML algorithms. For traditional formulas, we used Friedewald (9), Martin/Hopkins (14), Chen (16), Anandaraja (17) and Hattori (18). These formulas are listed below.

Friedewald's LDL-C estimation formula, which is denoted by LDL-C_F, is defined as follows:

$$LDL\text{-}C\_F = TC - HDL\text{-}C - (TG/5)$$

Martin/Hopkins LDL-C estimation formula denoted by LDL-C_M is defined as follows:

$$LDL\text{-}C\_M = TC - HDL\text{-}C - (TG/C)$$

In this formula, C denotes the adjustable factor that is used to estimate the TG/VLDL-C ratio. As stated in (14), 180-cell strata-specific median TG/VLDL-C ratio table was generated to estimate C. These median ratios range between 3 to 12 and are calculated using the TG and non-HDL-C sublevels. For more details and the 180-cell strata table, the readers are referred to see (14).

Chen's LDL-C estimation formula, which is denoted by LDL-C_C, is defined as

$$LDL\text{-}C\_C = (0.9*TC) - (0.9*HDL\text{-}C) - (0.1*TG)$$

Anandaraja's LDL-C estimation formula, which is denoted by LDL-C_A, is defined as

$$LDL\text{-}C\_A = (0.9*TC) - (0.9*TG/5) - 28$$

Hattori's LDL-C estimation formula, which is denoted by LDL-C_H, is defined as

$$LDL\text{-}C\_H = (0.94*TC) - (0.94*HDL\text{-}C) - (0.19*TG)$$

For ML algorithms, we implemented Ridge, Lasso, elastic net, random forest (RF), gradient boosting (GBM), extreme gradient boosting (XGBoost), and support vector regression (SVR), algorithms using TC, TG and HDL-C to predict the LDL-C concentrations. Ridge, Lasso and elastic net are three widely used penalized regression models, each of which aim to decrease the number of regression coefficients by shrinking the coefficients towards zero. Ridge regression penalizes the regression model with a penalty term called L2-norm, whereas Lasso penalizes the regression model with a penalty term called L1-norm. The penalty can be fine-tuned using a constant known as lambda (λ). Elastic net, on the other hand, is a convex combination of Ridge and Lasso models. For more details, readers are referred to (23,24). RF is an ensemble learning model that combines multiple decision trees to produce more accurate predictions or results. It has numerous advantages, including the avoidance of overfitting (23). Gradient boosting is a

boosting learning algorithm that learns from the previous mistakes-residual errors and XGBoost is an optimized GBM with high speed and performance. Support Vector Regression (23) operates on the same principle as the support vector machines. The main goal is to find the line that fits the data the best. This line is called as hyperplane.

For ML algorithms, the original dataset was randomly split into training (70%) and test (30%) sets. The model was trained using the training set, and the model's performance was evaluated using the test set. The models' hyperparameters were fine-tuned using 10-fold- cross-validation. The caret package in R (version 4.0.4) was used to implement ML models (25).

Statistical Analysis

To assess the performance of the regression models, standard error of the estimate (SEE) was calculated using the following formula:

$$\text{SEE} = \sqrt{\frac{(\text{Observed} - \text{Predicted})^2}{N}}$$

where N is the number of observations in the data. The 95% confidence interval for each model was also calculated using chi-square distribution with N-1 degrees of freedom:

$$\sqrt{\frac{\sum(N-1)\text{SEE}^2}{\chi^2_{0.025}}} < \text{SEE} < \sqrt{\frac{\sum(N-1)\text{SEE}^2}{\chi^2_{0.975}}}$$

The overall concordance statistic was used to determine the consistency between estimated LDL-C and directly measured LDL-C. The concordance statistic, which was calculated as the ratio of direct LDL-C within a specific range as estimated LDL-C levels in the same range as direct LDL-C, was mathematically defined as follows:

$$\text{Concordance} = \frac{\#\ of\ A \cap B}{\#\ of\ B}$$

where A and B denote the samples with estimated LDL-C within a specific range and the samples with directly measured LDL-C in the same range as A, respectively. The concordance statistics for TG and non-HDL-C sublevels were also computed. In order to compare the estimated and directly measured LDL-C levels, ordinary least squares regression analyses were employed for each method and Roche direct assay. Moreover, residual error plots were produced to demonstrate the difference between measured and estimated LDL-C levels based on TG levels. R 4.0.4 (*www.r-project.org*) was used to perform all statistical analyses.

Results

Patient Characteristics

Table 1 summarizes the patient characteristics. The study included a total of 2,356 children, with 57.85% being females and 42.15% being males. The median age of all individuals participated in this study was 13(9-16). The mean values of directly LDL-C, TC, TG, and HDL-C levels were 94.36±32.26 mg/dL,

152.96±36.80 mg/dL, 118.05±67.69 mg/dL, and 45.64±12.39 mg/dL, respectively. Patient characteristics for both the training and the test sets were also given (see Table 2).

Table 1

Characteristics of the study population (N = 2,356)

Characteristics	Values
Age (years)	13(9-16)
Sex	
Female	1363 (57.85)
Male	993 (42.15)
Lipid values	
TC (mg/dL)	152.96±36.80
TG (mg/dL)	118.05±67.69
HDL-C (mg/dL)	45.64±12.39
Non-HDL-C (mg/dL)	107.32±34.89
LDL-CD (mg/dL)	94.36±32.26

Values are presented as N (%), mean±SD or median (1st – 3rd quartiles). TC: total cholesterol; TG: triglycerides; HDL-C: high-density lipoprotein cholesterol; LDL-CD: low-density lipoprotein cholesterol measured by Roche direct assay

Table 2

Characteristics Of the Training and Test Sets

Characteristics	Values (Training)	Values (Test)
Age (years)	13(9-16)	13(10-15.5)
Sex		
Female	709 (43)	284 (40.17)
Male	940 (57)	423 (59.83)
Lipid values		
TC (mg/dL)	152.9±37.34	153.20±35.52
TG (mg/dL)	118.40±67.69	117.30±67.73
HDL-C (mg/dL)	45.46±12.35	46.08±12.50
Non-HDL-C (mg/dL)	107.40±35.62	107.10±33.15
LDL-CD (mg/dL)	94.30±33.06	94.510±30.32

Values are presented as N (%), mean±SD or median (1st – 3rd quartiles). TC: total cholesterol; TG: triglycerides; HDL-C: high-density lipoprotein cholesterol; LDL-CD: low-density lipoprotein cholesterol measured by Roche direct assay

The Results of Hyperparameter Tuning

All ML methods were applied as recommended in the documentation and hyperparameters were tuned using 10-fold cross validation and grid search. In Ridge, Lasso, and elastic net models, the parameter that was tuned was lambda (range 0-1). The optimal lambda values for Ridge, Lasso, and elastic net were 0, 1, and 1e-08, respectively. Two parameters, sigma (range 0-1) and C (tradeoff between decision boundary and misclassification error, range 0.25-1), were fine-tuned in SVR. The best values for sigma and C were 0.001 and 1, respectively. We note here that radial basis kernel was chosen for SVR. The hyperparameters that were tuned in RF were number of trees (found 2000) and mtry, which is the number of variables randomly sampled as candidates at each split (found 3). For GBM, interaction depth (range 1-7), number of trees (range 500-5000), shrinkage or eta (learning rate, 0.01-0.1) and the minimum number of observations in the terminal nodes of trees (n.minobsinnode, range 10-15) were tuned. The optimal values for interaction depth, number of trees, learning rate and n.minobsinnode were 1, 5000, 0.01 and 10, respectively. For XGBoost, the number of rounds was set to 1000, gamma was set to 0 and maximum depth of tree (max.depth, range 1-5), the percentage of columns to be randomly sampled for each tree (colsample_bytree, range 1-3), learning rate (eta, range 0.025-0.1) and subsample (range 0.5-1) were fine-tuned. The optimal parameters for XGBoost were found to be 1 for max.depth, 0.1 for eta, and 1 for subsample.

Comparison of LDL-C Concentrations Calculated by Various Formulas and ML Models versus Roche Direct Assay

To assess and compare the performance of the formulas and predictive ML models, standard error of estimate was used. The SEE and 95% confidence intervals for the LDL-C estimating formulas and ML models are provided in Figure 1. While ML models performed comparably, they all outperformed the LDL-C estimating formulas. Ridge, Lasso, and elastic net models, on the other hand, outperformed all other ML models, with the lowest SEE (Figure 1). Furthermore, Chen's formula has the lowest SEE when compared to other LDL-C estimating formulas.

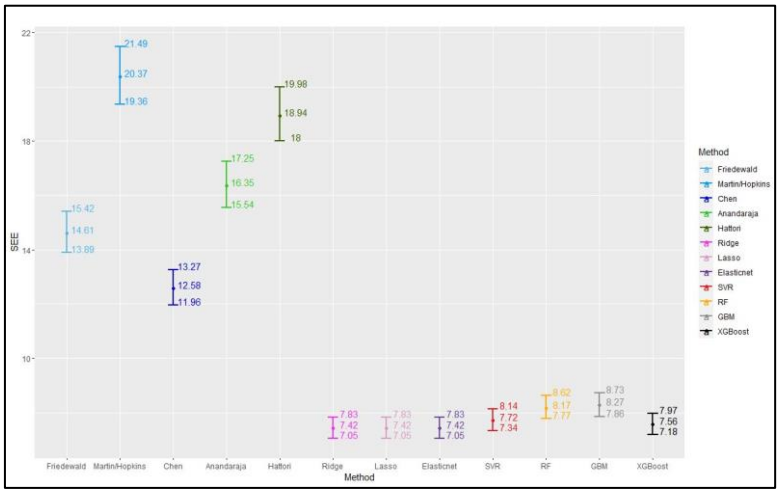

Figure 1. Standard error of estimate and 95% confidence intervals for each LDL-C estimating model. Each bar in the graph shows the lower confidence interval for SEE, SEE and the upper confidence for SEE from bottom to top, respectively. While x axis represents each model used for estimating LDL-C levels, y axis represents the SEE for each method (i.e., LDL-C estimating formulas and ML models).

Overall Concordances of Each Method Used for LDL-C Estimation

The concordance statistic was used to assess the consistency between estimated LDL-C and directly measured LDL-C. Figure 2 shows the overall concordances for each formula and ML models. The concordance was about 0.92-0.93 percent for ML models, while it was around 0.85 percent for LDL-C estimating formulas. The SVR formula generated the most concordant results (concordance=0.938), while the Hattori and Martin-Hopkins formulas produced the least concordant results (concordance=0.851), as shown in Figure 2.

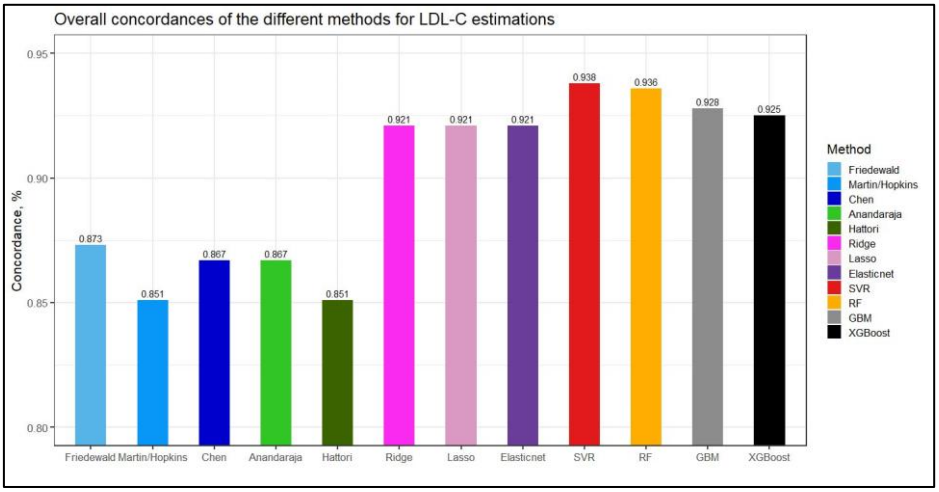

Figure 2. Overall concordance of each method used for LDL-C estimation. Overall concordances of each method (i.e., LDL-C estimating formulas and ML models) used for LDL-C estimation are shown in a bar chart, with each bar displaying the concordance of LDL-C concentrations by each LDL-C estimating formulas or predictive ML models with the Roche direct assay.

Concordances of Each Method Used for LDL-C Estimation by LDL-C Sublevels

The concordances of each method (i.e., LDL-C estimating formulas and ML models) used for LDL-C estimation by LDL-C sublevels (< 110 mg/dL, 110 to 129 mg/dL and ≥ 130 mg/dL) are provided in Figure 3 and Supplementary File (Results.xlsx). When LDL-C was less than 110 mg/dL, the overall performance of LDL-C estimating formulas was superior to ML models. However, Anandaraja's formulas showed the least concordant result compared to LDL-C estimating formulas. Even though the Friedewald's formula was slightly better than Anandaraja's formula, the most concordant results were achieved with Hattori's formula. When LDL-C levels were between 110 and 129 mg/dL, each method's overall concordance decreased. However, the ML models gave the most concordant results. In contrast to the LDL-C<110 sublevel, Hattori had the lowest concordancy and Martin-Hopkins had the second lowest when LDL-C was between 110 and 129 mg/dL. It can be seen from figure that ML models outperformed the LDL-C estimating formulas and for ML models, the highest performances were achieved with SVR and RF models. When LDL-C was higher than 129 mg/dL, ML models performed better than LDL-C estimating formulas. Among LDL-C estimating formulas, Anandaraja's formula performed best, while SVR performed best among ML models.

Concordances of Each Method Used for LDL-C Estimation by Triglycerides Sublevels

The concordances of each method (i.e., LDL-C estimating formulas and ML models) used for LDL-C estimation by triglycerides sublevels (<75 mg/dL, 75 to 129 mg/dL and ≥ 130 mg/dL) are given in Figure 4 and Supplementary File (Results.xlsx). The results showed that ML models gave higher concordant results overall for each TG sublevels. The performance of ML models varies depending on TG sublevel. When TG<75 mg/dL, RF produced the most concordant results, whereas Anandaraja produced the least concordant results. The results of ML models such as SVR, Ridge, Lasso, and elastic net were quite similar. When the TG was between 75 mg/dL and 129 md/dL, ML models outperformed the LDL-C estimating formulas in terms of concordance. SVR and RF performed the best among all ML models, and Anandaraja performed the best among all LDL-C estimating formulas. When TG was above 129 mg/dL, Ml models produced the most concordant results, with SVR being the best. However, the overall concordance for each model decreased when TG≥130.

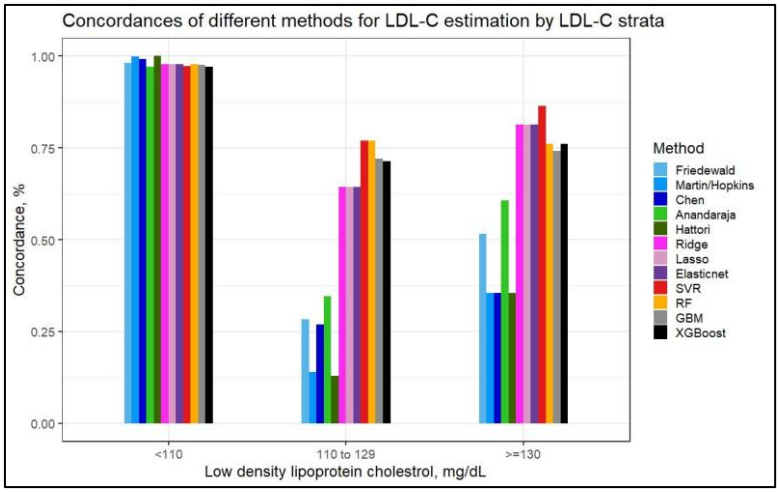

Figure 3. Concordance of each method used for LDL-C estimation by LDL-C sublevels. Concordances of each method (i.e., LDL-C estimating formulas and ML models) for LDL-C estimation by LDL-C sublevels are shown in a bar chart, with each bar displaying the concordance of LDL-C concentrations by each LDL-C estimating formulas or predictive ML models with the Roche direct assay.

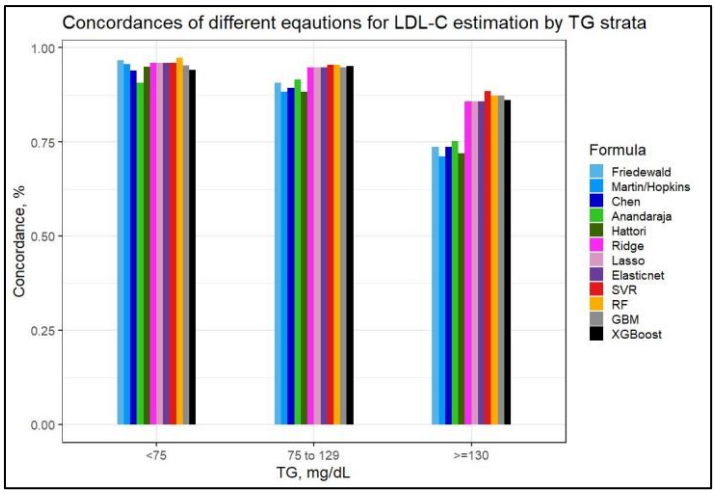

Figure 4. Concordance of each method used for LDL-C estimation by triglycerides sublevels. Concordances of each method (i.e., LDL-C estimating formulas and ML models) for LDL-C estimation by triglycerides sublevels are shown in a bar chart, with each bar displaying the concordance of LDL-C concentrations by each LDL-C estimating formulas or predictive ML models with the Roche direct assay for each triglycerides sublevel.

Regression Analysis Between Estimated and Directly Measured LDL-C Levels

Linear regression analyses were performed to investigate the relationship between estimated LDL-C levels and directly measured LDL-C levels. Figure 5 shows the regression plots between estimated and directly measured LDL-C levels for each formula and ML model. It is obvious to see that ML models display a better correlation with directly measured LDL-C levels overall.

The residual error plots, showing the difference between LDL-C estimations and direct measurements varies according to triglyceride levels, are given in Figure 6. It can be seen from the figure that LDL-C estimating formulas underestimated the LDL-C levels when TG level elevated. It was observed that the difference for ML models, particularly for Ridge, Lasso and elastic net was close to zero.

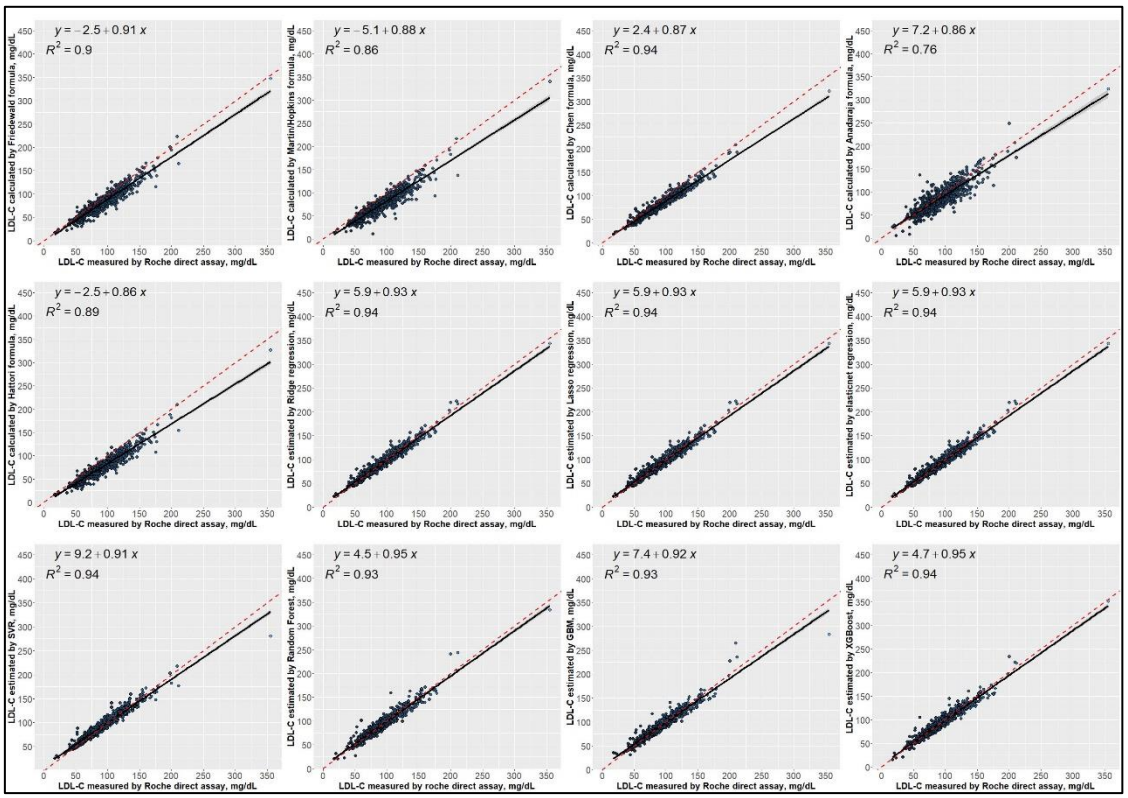

Figure 5. Regression analysis between estimated and directly measured LDL-C levels. Correlations between LDL-C levels estimated by different formulas or ML predictive models with LDL-C levels measured by Roche direct assay.

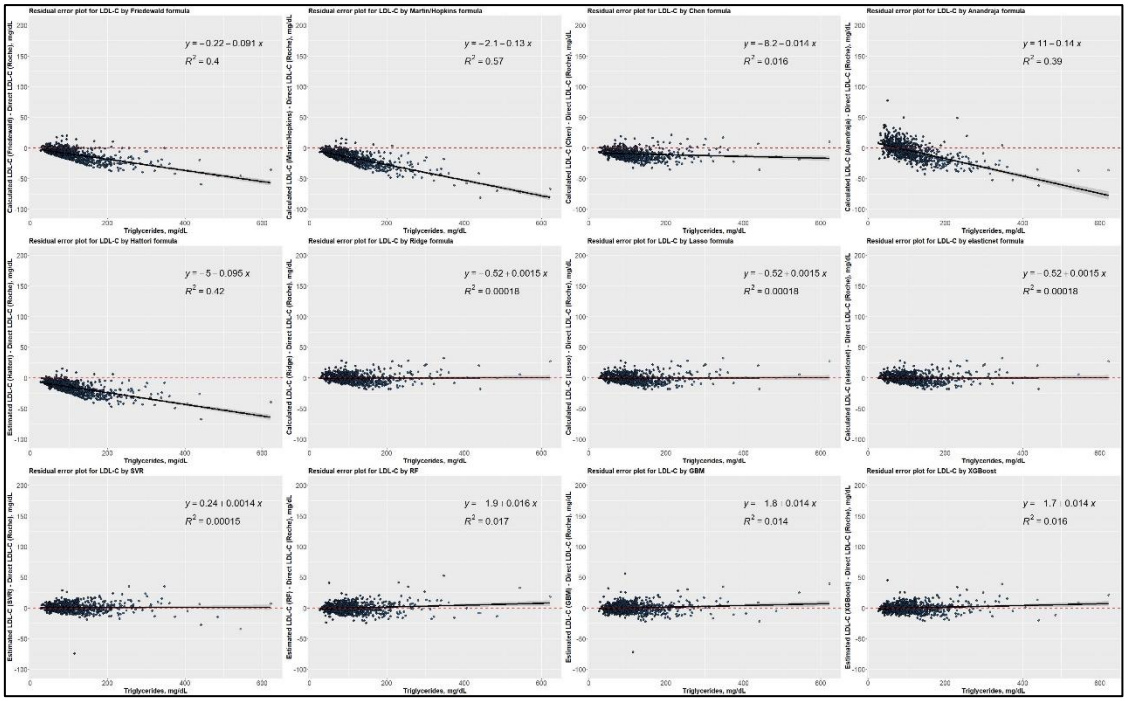

Figure 6. Residual error plots for LDL-C by either different formulas or ML predictive models with respect to Roche direct assay. The values on the x-axis demonstrate TG levels, while the values on the y-axis demonstrate the difference between estimated LDL-C (i.e., LDL-C estimating formulas and ML predictive models) and directly measured LDL-C.

Discussion

The standard error of estimate, which assesses the predictive power of the model, is a statistical measure used to estimate the accuracy of predictions or estimates made by a statistical model (26). It represents the average amount by which the observed values deviate from the predicted values, which helps us understand the variability of the data points around the regression line. We calculated the SEE for each model and the results showed that ML models outperformed the LDL-C estimating formulas in terms of SEE (Figure 1).

While studies investigating the association between LDL-C levels and cardiovascular risk in children and adolescents are relatively common, studies using machine learning algorithms for LDL-C estimation are limited. However, studies in adult populations have shown that ML algorithms can provide more accurate and reliable estimates of LDL-C levels compared to traditional equations (19,20,21,22), and our results suggest that this may also be true in the pediatric population. Our study demonstrated that ML algorithms outperformed the traditional LDL-C estimating equations in estimating LDL-C levels in a pediatric population, suggesting that ML algorithms may provide a more accurate and reliable approach for LDL-C estimation in the Turkish pediatric population. These findings highlight the potential utility of ML algorithms in improving cardiovascular risk assessment and management in children and adolescents. However, further studies are needed to confirm these findings and to evaluate the clinical utility of ML-based LDL-C estimation in the pediatric population.

Due to the Friedewald formula's limitations, such as the requirement for fasting serum and an underestimation of LDL-C values less than 70 and greater than 150 mg/dL, new formulas such as Chen, Anandaraja, Hattori, and Martin-Hopkins formulas were developed. Recently, researchers have started to explore ML algorithms for the purpose of LDL-C estimation due to the fact that they can learn the natural structure of the data as well as the relationship between variables that produces more accurate predictions (19,21,22,27). However, very little research has been conducted in the pediatric population to assess the validity of traditional LDL-C estimating equations and powerful ML algorithms. In this study, we investigated the validity of ML algorithms, as well as several LDL-C estimating equations for the Turkish pediatric population. The performance of the ML algorithms was also compared with these conventional LDL-C estimating equations. The results showed that ML algorithms predict LDL-C levels more accurately than the LDL-C estimating formulas. However, there is no single model that we can call the best among ML algorithms. The performance of ML models varies according to LDL-C levels.

Recently, it has been demonstrated that ML techniques can be used to replace existing LDL-C estimation models, particularly for higher TG levels. In the presence of high TG levels, chylomicrons accumulate at high levels, which may potentially alter the association between TG and cholesterol levels. In order to assess the impact of high TG levels on the precision of each method developed for LDL-C estimation in the children and adolescence, we used the cut-off value of ≥130 mg/dL, abnormally high TG levels in children and adolescents (28). The results showed that SVR gave the highest concordance for TG≥130 mg/dL. Given that SVR and RF had the most concordant results overall, and SVR performed best at higher TG levels, we believe that ML models can be used as an alternative method to the conventional LDL-C estimating equations for accurate LDL-C estimation.

There are three limitations of our study. First, we did not use beta quantification as a reference method because it is an expensive and labor-intensive manual technique. However, using this method instead of the Roche direct assay could yield more accurate results. Second, we did not analyze the results for fasting and non-fasting subjects separately. The concentrations of TG-enriched chylomicrons were shown to be higher in non-fasting samples than in fasting samples, which may result in an increased ratio of TG to cholesterol in VLDL (20,29). As a result, Friedewald's underestimation of LDL-C may have been overstated. Third, the study scope is limited due to the absence of external validation, and to enhance the applicability of the findings, further validation is necessary. It is crucial to validate the study results not only in other

Turkish pediatric populations but also across diverse pediatric populations from different ethnicities. This will enable researchers to verify whether the results are consistent across a broader population and promote the external validity of the study.

Conclusion

In conclusion, it has been demonstrated that ML algorithms outperformed traditional LDL-C estimating equations in the Turkish pediatric population. SVR and RF models, among ML algorithms, provided more accurate LDL-C estimates in the Turkish pediatric population.

Ethics Committee Approval: This study was approved by Sivas Cumhuriyet University of medical sciences with document no: 2022-06/02 (date: 22.06.2022).

Informed Consent: Consent was not obtained as it was a retrospective study.

Conflict of Interest: Authors declared no conflict of interest.

Financial Disclosure: Authors declared no financial support.

Acknowledgement: We would like to express our gratitude to Dr. Halef Okan Doğan for generously sharing the data used in this research. Additionally, we extend our thanks to the referees and editors for their insightful comments, which greatly contributed to the refinement and overall quality of the paper.

References

1. Berenson GS, Srinivasan SR, Bao W, Newman W, Tracy RE, Wattigney WA, et al. Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults. The Bogalusa Heart Study. N Engl J Med. 1998; 338 (23): 1650-1656. doi: 10.1056/NEJM199806043382302.
2. Daniels SR, Greer FR. Lipid screening and cardiovascular health in childhood. Pediatrics. 2008; 122 (1): 198-208. doi: 10.1542/peds.2008-1349.
3. McGill HC, McMahan CA, Zieske AW, Malcom GT, Tracy RE, Strong JP. Effects of nonlipid risk factors on atherosclerosis in youth with a favorable lipoprotein profile. Circulation. 2001; 103 (11): 1546-1550. doi: 10.1161/01.cir.103.11.1546.
4. Gidding SS. Cholesterol Guidelines Debate. Pediatrics. 2001; 107 (5): 1229–1230. https://doi.org/10.1542/peds.107.5.1229.
5. Fox KM, Wang L, Gandra SR, Quek RGW, Li L, Baser O. Clinical and economic burden associated with cardiovascular events among patients with hyperlipidemia: A retrospective cohort study. BMC Cardiovasc Disord. 2016; 16 (1). doi: 10.1186/s12872-016-0190-x.
6. Molavi F, Namazi N, Asadi M, Sanjari M, Motlagh ME, Shafiee G, et al. Comparison common equations for LDL-C calculation with direct assay and developing a novel formula in Iranian children and adolescents: The CASPIAN v study. Lipids Health Dis. 2020; 19 (1). https://doi.org/10.1186/s12944-020-01306-7.
7. Silverman MG, Ference BA, Im K, Wiviott SD, Giugliano RP, Grundy SM, et al. Association between lowering LDL-C and cardiovascular risk reduction among different therapeutic interventions: A systematic review and meta-analysis. JAMA - J Am Med Assoc. 2016; 316 (12): 1289-1297. doi: 10.1001/jama.2016.13985.
8. CDC. Centers for Disease Control and Prevention National Reference System for Cholesterol - Cholesterol Reference Method Laboratory Network - Total Cholesterol - Certification Protocol for Manufacturers-Revised. 2004;(cited 2022 April 20). Available from: http://www.cdc.gov/labstandards/pdf/crmln/FrozVsFreshProtocolOct04.pdf

9. Friedewald Wt Fau - Levy RI, Levy Ri Fau - Fredrickson DS, Fredrickson DS, Clin C. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem. 1972 Jun; 18 (6): 499-502.
10. De Cordova CMM, Schneider CR, Juttel ID, De Cordova MM. Comparison of LDL-cholesterol direct measurement with the estimate using the Friedewald formula in a sample of 10,664 patients. Arq Bras Cardiol. 2004; 83 (6). doi: 10.1590/s0066-782x2004001800006.
11. de Cordova CMM, de Cordova MM. A new accurate, simple formula for LDL-cholesterol estimation based on directly measured blood lipids from a large cohort. Ann Clin Biochem. 2013; 50 (1): 13-9. doi: 10.1258/acb.2012.011259.
12. Türkalp I, Çil Z, Özkazanç D. Analytical performance of a direct assay for LDL-cholesterol: A comparative assessment versus Friedewald's formula. Anadolu Kardiyol Derg. 2005; 5 (1):13-17. https://pubmed.ncbi.nlm.nih.gov/15755695/.
13. Hermans MP, Ahn SA, Rousseau MF. Novel unbiased equations to calculate triglyceride-rich lipoprotein cholesterol from routine non-fasting lipids. Cardiovasc Diabetol. 2014; 13 (1). https://doi.org/10.1186/1475-2840-13-56.
14. 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 - J Am Med Assoc. 2013; 310 (19): 2061-2068. doi: 10.1001/jama.2013.280532.
15. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Vol. 139, Circulation. 2019.
16. Chen Y, Zhang X, Pan B, Jin X, Yao H, Chen B, et al. A modified formula for calculating low-density lipoprotein cholesterol values. Lipids Health Dis. 2010; 9 (1). doi: 10.1186/1476-511X-9-52.
17. Anandaraja S, Narang R, Godeswar R, Laksmy R, Talwar KK. Low-density lipoprotein cholesterol estimation by a new formula in Indian population. Int J Cardiol. 2005; 102 (1): 117-120. doi: 10.1016/j.ijcard.2004.05.009.
18. Hattori Y, Suzuki M, Tsushima M, Yoshida M, Tokunaga Y, Wang Y, et al. Development of approximate formula for LDL-chol, LDL-apo B and LDL- chol/LDL-apo B as indices of hyperapobetalipoproteinemia and small dense LDL. Atherosclerosis. 1998; 138 (2): 289-299. doi: 10.1016/s0021-9150(98)00034-3.
19. 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 (1): 76-86. doi: 10.1177/00045632211046805.
20. Çubukçu HC, Topcu Dİ. Estimation of Low-Density Lipoprotein Cholesterol Concentration Using Machine Learning. Lab Med. 2022; 53 (2): 161-171. doi: 10.1093/labmed/lmab065.
21. Tsigalou C, Panopoulou M, Papadopoulos C, Karvelas A, Tsairidis D, Anagnostopoulos K. Estimation of low-density lipoprotein cholesterol by machine learning methods. Clin Chim Acta. 2021; 517: 108-116. doi: 10.1016/j.cca.2021.02.020.
22. Kwon Y-J, Lee H, Baik SJ, Chang H-J, Lee J-W. Comparison of a Machine Learning Method and Various Equations for Estimating Low-Density Lipoprotein Cholesterol in Korean Populations. Front Cardiovasc Med 2022 Feb 10; 9(February): 824574. doi: 10.3389/fcvm.2022.824574.
23. Hastie, Trevor, Tibshirani, Robert, Friedman J. The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition. Springer series in statistics. 2009.
24. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: The lasso and generalizations. Statistical Learning with Sparsity: The Lasso and Generalizations. 2015.
25. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008; 28 (5): doi: 10.18637/jss.v028.i05.
26. Kuhn, Max. Applied Predictive Modeling. Springer. 2013.

27. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. Vol. 76, Journal of the American College of Cardiology. 2020.
28. Lim JS, Kim EY, Kim JH, Yoo JH, Yi KH, Chae HW, et al. 2017 clinical practice guidelines for dyslipidemia of korean children and adolescents. Ann Pediatr Endocrinol Metab. 2020; 25 (4): 454-462. doi: 10.3345/cep.2020.01340.
29. Lund SS, Petersen M, Frandsen M, Smidt UM, Parving HH, Vaag AA, et al. Agreement between fasting and postprandial LDL cholesterol measured with 3 methods in patients with type 2 diabetes mellitus. Clin Chem. 2011; 57 (2): 298–308. https://doi.org/10.1373/clinchem.2009.133868.