J Korean Med Sci. 2023 May 15;38(19):e145
https://doi.org/10.3346/jkms.2023.38.e145
eISSN 1598-6357·pISSN 1011-8934

JKMS

Original Article
Cardiovascular Disorders

Comparison of Newly Proposed LDL-Cholesterol Estimation Equations

Yong Whi Jeong [1*] **Jun Hyuk Koo** [2*†] **Ji Hye Huh** [3] **Young-Jin Kim** [4] **Hoyeon Jeong** [1] **Eun Young Kim** [5] **and Dae Ryong Kang** [6]

[1]Department of Biostatistics, Graduate School, Yonsei University, Seoul, Korea
[2]Yonsei University Wonju Industry-Academic Cooperation Foundation, Wonju, Korea
[3]Division of Endocrinology and Metabolism, Department of Internal Medicine, Hallym University Sacred Heart Hospital, Anyang, Korea
[4]Seoul Clinical Laboratories Biobank, Yongin, Korea
[5]Department of Biomedical Laboratory Science, Songho University, Hoengseong, Korea
[6]Department of Precision Medicine, Wonju College of Medicine, Yonsei University, Wonju, Korea

Received: Nov 11, 2022
Accepted: Feb 1, 2023
Published online: Apr 20, 2023

Address for Correspondence:
Dae Ryong Kang, PhD
Department of Precision Medicine, Wonju College of Medicine, Yonsei University, 20 Ilsan-ro, Wonju 26426, Republic of Korea.
Email: dr.kang@yonsei.ac.kr

*Yong Whi Jeong and Jun Hyuk Koo contributed equally to this work as first authors.
†Current affiliation: HIRA Research Institute, Health Insurance Review & Assessment Service (HIRA), Wonju, Korea.

ORCID iDs
Yong Whi Jeong
https://orcid.org/0000-0003-4746-5764
Jun Hyuk Koo
https://orcid.org/0000-0002-5743-9271
Ji Hye Huh
https://orcid.org/0000-0001-5445-8007
Young-Jin Kim
https://orcid.org/0000-0002-4624-2705
Hoyeon Jeong
https://orcid.org/0000-0001-6812-9343

ABSTRACT

Background: Low-density lipoprotein cholesterol is an important marker highly associated with cardiovascular disease. Since the direct measurement of it is inefficient in terms of cost and time, it is common to estimate through the Friedewald equation developed about 50 years ago. However, various limitations exist since the Friedewald equation was not designed for Koreans. This study proposes a new low-density lipoprotein cholesterol estimation equation for South Koreans using nationally approved statistical data.
Methods: This study used data from the Korean National Health and Nutrition Examination Survey from 2009 to 2019. The 18,837 subjects were used to develop the equation for estimating low-density lipoprotein cholesterol. The subjects included individuals with low-density lipoprotein cholesterol levels directly measured among those with high-density lipoprotein cholesterol, triglycerides, and total cholesterol measured. We compared twelve equations developed in the previous studies and the newly proposed equation (model 1) developed in this study with the actual low-density lipoprotein cholesterol value in various ways.
Results: The low-density lipoprotein cholesterol value estimated using the estimation formula and the actual low-density lipoprotein cholesterol value were compared using the root mean squared error. When the triglyceride level was less than 400 mg/dL, the root mean squared of the model 1 was 7.96, the lowest compared to other equations, and the model 2 was 7.82. The degree of misclassification was checked according to the NECP ATP III 6 categories. As a result, the misclassification rate of the model 1 was the lowest at 18.9%, and Weighted Kappa was the highest at 0.919 (0.003), which means it significantly reduced the underestimation rate shown in other existing estimation equations. Root mean square error was also compared according to the change in triglycerides level. As the triglycerides level increased, the root mean square error showed an increasing trend in all equations, but it was confirmed that the model 1 was the lowest compared to other equations.
Conclusion: The newly proposed low-density lipoprotein cholesterol estimation equation showed significantly improved performance compared to the 12 existing estimation equations. The use of representative samples and external verification is required for more sophisticated estimates in the future.

Keywords: Low-Density Lipoprotein Cholesterol; Estimation Equation; Friedewald Equation; Triglycerides

Eun Young Kim
https://orcid.org/0000-0001-8014-7990
Dae Ryong Kang
https://orcid.org/0000-0002-8792-9730

Disclosure
The authors have no potential conflicts of interest to disclose.

Author Contributions
Conceptualization: Jeong YW, Jeong H, Huh JH, Kang DR. Methodology: Jeong YW, Kang DR. Formal analysis: Jeong YW, Koo JH. Data curation: Jeong YW. Software: Jeong YW. Jeong H. Validation: Jeong YW, Koo JH, Kim YJ, Kang DR. Writing - original draft preparation: Jeong YW, Koo JH. Writing - review and editing: Jeong YW, Koo JH, Huh JH, Kim EY, Kang DR.

INTRODUCTION

An increase in low-density lipoprotein cholesterol (LDL-c) has been found to be a significant risk factor for the development of cardiovascular disease,[1,2] making the accurate measurement of LDL-c values important. LDL-c values can be obtained via direct and calculation measurement methods. For direct measurement of LDL-c, lipid ultracentrifugation (beta-quantification procedure) is generally used, which is inefficient in terms of additional cost and time consumption.[3-5] From this point of view, most clinical sites estimate LDL-c values using calculation methods, particularly the Friedewald equation developed in 1972.[6]

Researchers have highlighted several problems with the Friedewald equation in its current form. First, the Friedewald equation becomes inaccurate as triglyceride (TG) values increase above 200 mg/dL, becomes invalid when they exceed 400 mg/dL,[5] and the result is relatively inaccurate when LDL-c values are below 70 mg/dL.[7,8] Additionally, LDL-c estimates calculated by the Friedewald equation generally underestimate actual LDL-c values.[8,9] Accordingly, various equations for indirectly calculating LDL-c have been developed. However, there are still limitations in applying the new equations to South Koreans. Since most of the equations were not designed for South Koreans, a race bias may exist. Compared to Westerners, South Koreans have relatively low total cholesterol (TC), LDL-c, and high-density lipoprotein cholesterol (HDL-c) levels and are known to have high TG.[10] Therefore, bias may occur if an estimation formula developed in Western countries is generalized to South Koreans. Interestingly, a relevant study conducted in Japan, which is geographically close to South Korea, proposed a Japanese-specific equation.[11]

In relevant studies conducted in South Korea, new equations, such as the Martin equation and the Sampson equation, having been found to be superior to the existing Friedewald equation, have been applied to South Koreans.[10,12,13] However, few studies have set out to develop LDL-c equations specifically for South Koreans. Recently, Choi et al.[5] developed a new LDL-c equation based on the cohort data of the Green Cross Research Institute and tested it using data from the Korea National Health and Nutrition Examination Survey (KNHANES). Their study reported that the newly developed equation was more accurate than 11 other existing equations. However, there is a limitation in that the effort to avoid overfitting was insufficient.

Recently, AI-based predictive models have drawn attention in various fields, and machine learning and deep learning methods are being actively used in the development of LDL-c estimation models.[14-16] AI-based prediction models additionally consider invisible characteristics between variables and generally perform better than existing equations.[17] Therefore, we aimed to develop a custom LDL-c equation model and AI-based prediction model for South Koreans using the KNHANES data. The predicted values from the newly developed equation and estimation model were compared with predicted values calculated from several previously developed equations and directly measured LDL-c values to evaluate their accuracy.

METHODS

Study population and design

This study was based on KNHANES data from 2009 to 2019. The National Health and Nutrition Examination Survey is a nationwide cross-sectional survey conducted by the Korea Centers for Disease Control and Prevention to evaluate the health and nutritional status of Koreans. Details on the survey have been published elsewhere.[12] The subjects of this study included individuals who had their LDL-c levels directly measured among those who had HDL-c, TG, and TC measured, which are variables of the Friedewald LDL estimation formula (N = 18,837). Considering that only one database was used, the results were derived and compared using five-fold cross-validation (**Fig. 1**).

A formula for estimating LDL-c was developed through 12 previous studies,[18-25] and the new equation was developed using the National Health and Nutrition Examination Survey dataset (**Table 1**). In KNHANES data, lipid profiles (TC and TG), HDL-c, and LDL-c were measured using the enzymatic method with a Hitachi Automatic Analyzer 7600 (Hitachi, Tokyo, Japan) in 2009–2012 and Labospect008AS (Hitachi) in 2019. From 2013 to 2018, using the Hitachi Automatic Analyzer 7600-210 (Hitachi), lipid profiles were measured by the enzymatic method, and HDL-c and LDL-c were measured by the homogeneous enzymatic colorimetric

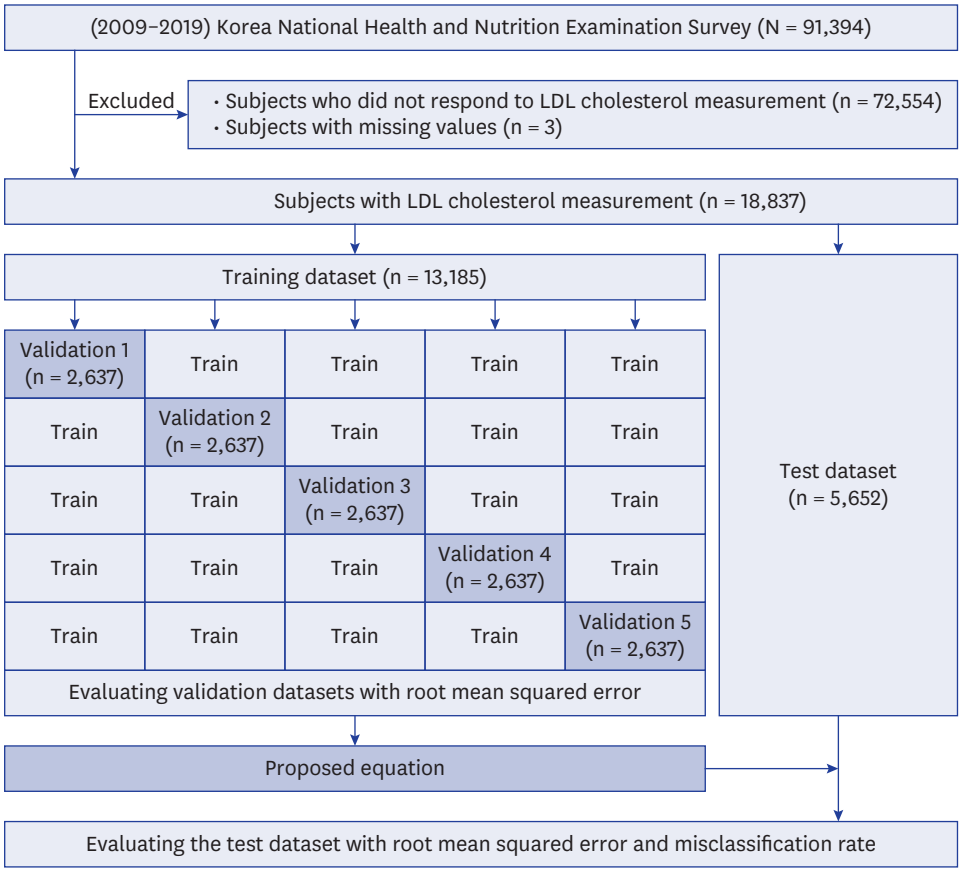

Fig. 1. Study design.
LDL = low-density lipoprotein.

https://doi.org/10.3346/jkms.2023.38.e145

Table 1. Characteristics of the 13 equations for calculating LDL-c

Method for LDL-c estimation	Year	TG range, mg/dL	Measurement method for LDL quantification	Equations
Model 1	The present study	Full range	Automated enzymatic method	0.94 × TC − 0.94 × HDL − 0.12 × TG
Friedewald et al.[6]	1972	Full range	Ultracentrifugation	TC − HDL − TG/5
DeLong et al.[18]	1986	Full range	Ultracentrifugation	TC − (HDL + 0.16 × TG)
Rao et al.[19]	1988	Full range	Ultracentrifugation	TC − HDL − {TG × (0.203 − 0.00011 × TG)}
Hattori et al.[20]	1988	TG < 400	Ultracentrifugation	0.94 × TC − 0.94 × HDL − 0.19 × TG
Anandaraja et al.[21]	2005	Full range	Ultracentrifugation	0.9 × TC − 0.9 × TG/5 − 28
Puavilai et al.[22]	2009	Full range	Automated enzymatic method	TC − HDL − TG/6
Vujovic et al.[23]	2010	TG < 400	Automated enzymatic method	TC − HDL − TG/6.58
Chen and Zhang[24]	2010	Full range	Automated enzymatic method	(TC − HDL) × 0.9 − TG × 0.1
de Cordova and de Cordova[25]	2013	Full range	Automated enzymatic method	0.7516 × (TC − HDL)
Martin et al.[8]	2013	Full range	Ultracentrifugation	TC − HDL − TG/novel factors
Sampson et al.[4]	2021	Full range	Ultracentrifugation	TC/0.948 − HDL/0.971 − (TG/8.56 + TG × Non-HDL/2140 − TG^2/16100) − 9.44
Choi et al.[5]	2021	Full range	Automated enzymatic method	TC − 0.87 × HDL − 0.13 × TG

LDL-c = low-density lipoprotein cholesterol, TG = triglycerides, TC = total cholesterol, HDL = high-density lipoprotein.

method. The new equation was developed using the parameters of TC, HDL-c, and TG and multiple regression analysis from the KNHANES data.

$$\text{Model 1} = \beta_1 * \text{TC} + \beta_2 * \text{HDL-c} + \beta_3 * \text{TG}$$

The primary process in formula development was to derive a simple formula like Friedewald that could be easily used in clinical practice. In addition, optimal β_1, β_2, and β_3 values were derived by minimizing error from a straight line or curve obtained from the actual data using the least squares method. For development and validation, the dataset was randomly divided into a training set (70%) and a test set (30%), and five-fold cross-validation was performed to optimize the prediction model. In addition, overfitting was prevented by repeatedly checking the root mean squared error (RMSE) index values of the training and test sets.

Since the development of an estimation formula alone can introduce bias of good performance for a particular dataset, the results between the prediction model using the machine learning method and the estimation formula were also compared. Among tree-based ensemble methods, XGBoost is one of supervised learning that can perform classification and regression tasks because learning and computation are fast using parallel processing, and XGBoost has its own overfitting regulation function, which enables high-level optimization.[26] This was used as a comparative model, and the optimal hyperparameter was found using a grid search to derive high predictive power.

Statistical analysis

The distribution of variables used to develop the LDL estimation formula for Koreans was confirmed as number (%) for categorical variables and median (interquartile range) for continuous variables. To evaluate the error between actual LDL values and estimated LDL values, RMSE was used for comparison and R^2 was confirmed using a scatter plot and a regression line (**Supplementary Fig. 1**). To verify that the estimated formula can be used in clinical practice, estimated LDL values according to the six NCEP ATP III categories (LDL-c levels: 1) < 70 mg/dL, 2) 70–99 mg/dL, 3) 100–129 mg/dL, 4) 130–159 mg/dL, 5) 160–189 mg/dL, and 6) ≥ 190 mg/dL) and the degree of misclassification according to the actual LDL category was assessed. Weighted kappa coefficients using the Fleiss-Cohen method were used in the analysis to confirm agreement.[27] SAS 9.4 (SAS., Cary, NC, USA) was used to

calculate basic statistics, and Python 3.7.6 (pandas, NumPy, scikit-learn, matplotlib packages, XGBoost) was used to develop the equation and evaluate the results.

Ethics statement

All participants provided written informed consent to participate in this survey, and we received the data in an anonymized form. The study was approved by the Institutional Review Board (IRB) of Wonju Severance Christian Hospital (IRB No. CR321337).

RESULTS

The baseline characteristics and lipid profiles of the subjects are described in **Table 2**. The total number of subjects was 18,837, totaling 13,185 (70.00%) in the training dataset and 5,652 (30.00%) in the test dataset. Overall, 52.38% were male, and 47.62% were female.

The estimation equations developed using previously proposed estimation equations and using the dataset divided according to five-fold cross-validation were compared using RMSE values (**Table 3**). When TG levels were less than 400 mg/dL, the RMSE value of the model 1 (CV1: 7.97; CV2: 8.04; CV3: 8.13; CV4: 8.14; CV5: 8.04) showed the lowest result, and the equation derived from CV1 was used as the final equation. In addition, we confirmed that the error was lowest at 7.97 when compared with actual LDL values in the test dataset using the derived formula. Even when comparing the results of the model 1 with RMSE values (CV1: 7.94; CV2: 7.83; CV3: 7.95; CV4: 8.02; CV5: 7.91) of the model 2 (XGBoost), which underwent a complex calculation process, RMSE values showed no significant difference. Additionally, the scatter plot results were compared when TG levels were less than 400 mg/dL and when there was no range limitation (**Supplementary Fig. 1**). When the TG levels were less than 400 mg/dL, the new equation showed the strongest linearity (R^2 = 0.94), compared to other estimation equations. In the full range results, the coefficient of determination of the Sampson equation, which has an advantage when TG levels increase, was 0.91. The coefficient of determination of the model 1 was 0.90, and the model 2 was the highest at 0.93.

Fig. 2 confirms the degree of misclassification and agreement according to six NCEP ATP III categories when TG levels were less than 400 mg/dL for the estimated LDL values. The misclassification rate was calculated using a confusion matrix derived from the test dataset

Table 2. Baseline characteristics of the study datasets

Variables	Dataset
Total	18,837
Dataset	
Training, validation	13,185 (70.00)
Test	5,652 (30.00)
Sex	
Male	9,867 (52.38)
Female	8,970 (47.62)
Age, yr	48 (34–60)
Triglyceride, mg/dL	156 (85–247)
Cholesterol, mg/dL	
Total	190 (165–217)
HDL	45.84 (39.08–54.40)
Measured LDL	111 (90–134)

Values are presented as number (%) or median (interquartile range).
HDL = high-density lipoprotein, LDL = low-density lipoprotein.

https://doi.org/10.3346/jkms.2023.38.e145

Table 3. Comparison of root mean squared error values between measured LDL-c and estimated LDL-c levels

LDL-c estimation	Samples	Training, validation dataset (n = 13,185)					Test dataset (n = 5,652)
		CV1	CV2	CV3	CV4	CV5	
Model 1	TG < 400	7.97	8.04	8.13	8.14	8.04	7.96
	Full range	9.61	9.99	10.20	9.78	10.43	10.40
Model 2	TG < 400	7.94	7.83	7.95	8.02	7.91	7.82
	Full range	9.35	8.74	8.87	8.69	8.88	8.66
Friedewald et al.[6]	TG < 400	10.97	10.74	10.84	10.88	11.20	10.93
	Full range	17.19	15.65	17.55	16.65	17.74	16.69
DeLong et al.[18]	TG < 400	8.82	9.13	8.90	9.16	9.03	8.93
	Full range	11.82	11.41	12.18	11.82	12.45	12.14
Rao et al.[19]	TG < 400	8.70	8.75	8.70	8.87	8.89	8.72
	Full range	15.98	15.30	16.35	12.58	16.07	13.54
Hattori et al.[20]	TG < 400	15.00	14.38	14.82	14.64	15.14	14.82
	Full range	20.17	18.64	20.43	19.54	20.58	19.60
Anandaraja et al.[21]	TG < 400	16.26	16.25	16.60	16.51	16.57	16.64
	Full range	20.39	19.43	20.93	20.32	20.99	20.33
Puavilai et al.[22]	TG < 400	8.78	9.01	8.81	9.06	9.00	8.87
	Full range	12.32	11.66	12.66	12.21	12.92	12.49
Vujovic et al.[23]	TG < 400	9.11	9.48	9.22	9.49	9.30	9.23
	Full range	11.51	11.40	11.91	11.64	12.20	12.00
Chen and Zhang[24]	TG < 400	8.47	8.35	8.63	8.49	8.46	8.38
	Full range	9.80	10.35	10.47	9.91	10.64	10.60
de Cordova and de Cordova[25]	TG < 400	14.76	14.42	14.90	14.56	14.50	14.46
	Full range	18.93	19.55	20.20	18.78	20.18	19.14
Martin et al.[8]	TG < 400	8.16	8.26	8.35	8.35	8.19	8.16
	Full range	10.64	10.39	10.75	10.27	11.11	10.66
Sampson et al.[4]	TG < 400	8.52	8.63	8.54	8.69	8.66	8.56
	Full range	9.90	9.92	9.98	9.97	10.06	9.99
Choi et al.[5]	TG < 400	15.91	16.38	16.12	16.34	16.05	16.07
	Full range	16.81	17.47	17.33	17.26	17.56	17.54

TG < 400 samples were obtained from 2,474 individuals in cv1-cv5, respectively.
LDL-c = low-density lipoprotein cholesterol, CV = cross validation, TG = triglycerides.

(**Supplementary Fig. 2**), and the degree of concordance was derived using weighted kappa (SE). The misclassification rate of the Sampson equation was 20.7%, the developed equation by Choi in Korea was 43.0%, that of the Martin equation was 19.4%, and that of the model 1 was 18.9%, which was the lowest. The model 1 greatly reduced the tendency of the Friedewald equation to be underestimated and showed a lower misclassification rate than the model 2.

Table 4 and **Fig. 3** compare the RMSE between the predicted and actual values according to the TG level. As a result of the model 1 and the model 2, the RMSE value also tended to increase as the TG level increased, but the RMSE was the lowest at ten or less. The RMSE of the Martin equation was almost similar to the model 1 until the TG level was less than 200. However, when the TG level was 200 or higher, the RMSE value exceeded ten and significantly increased compared to the model 1.

DISCUSSION

We developed a customized LDL-c equation for Koreans based on TC, HDL-c, and TG values using KNHANES data. To verify the consistency and accuracy of the newly developed equation, we compared measured LDL-c values and values predicted from a total of 12 previously developed equations. The LDL-c predicted values calculated from the equations were compared with the directly measured LDL-c value through R^2, RMSE, and scatter plots, and the degree of misclassification in LDL-c category classification according to the NCEP

https://doi.org/10.3346/jkms.2023.38.e145

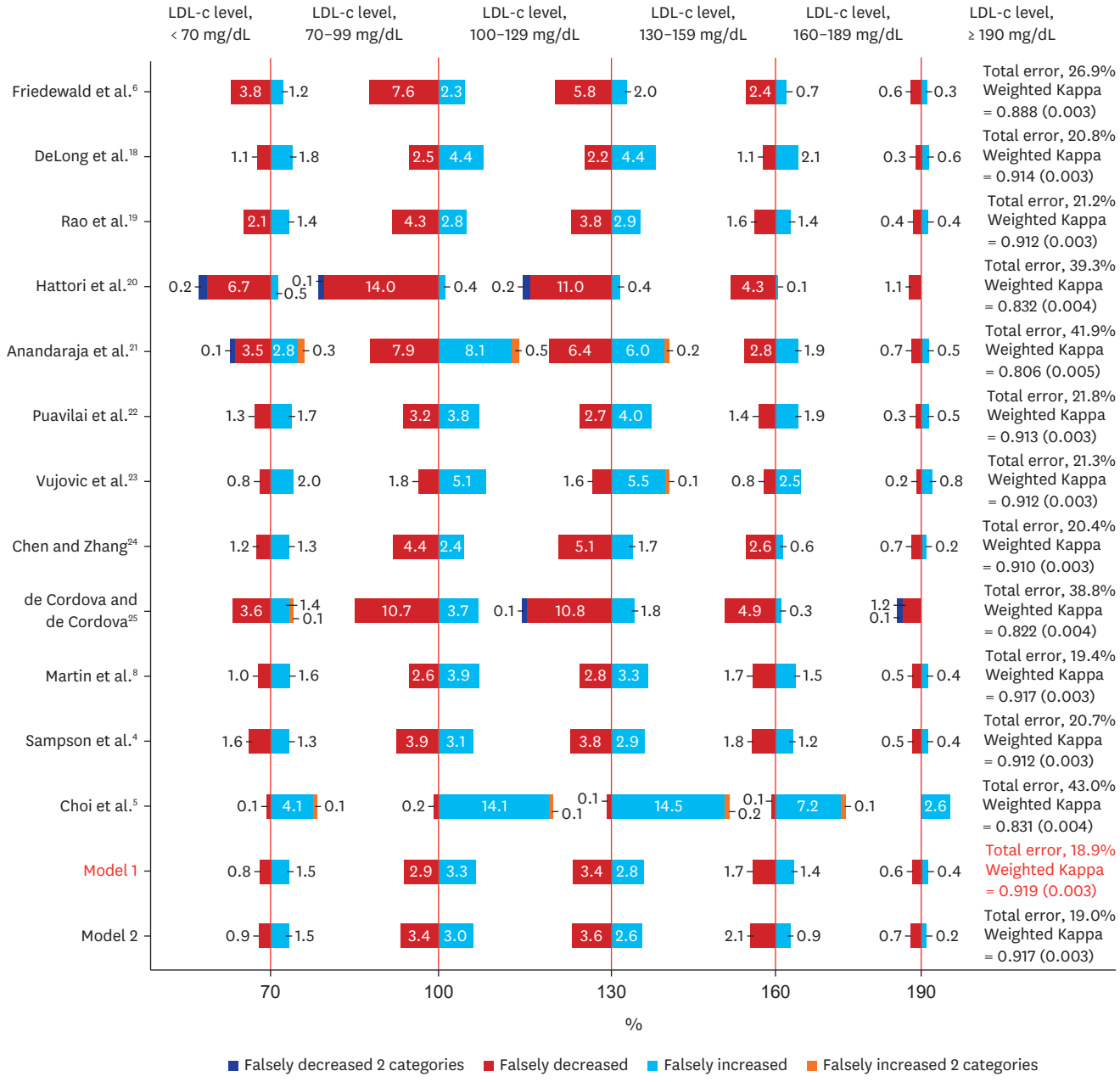

Fig. 2. Misclassification of patients with LDL-c levels using NCEP ATP III criteria in the test dataset. (triglycerides level < 400 mg/dL).
LDD-c = low-density lipoprotein cholesterol.

ATP III standard was also checked. In this process, a machine learning model using XGBoost was also compared. Our results confirmed that the newly developed LDL-c equation shows superior or at least equivalent performance relative to other previously developed equations and the XGBoost model.

Comparing the Friedewald equation, which is currently most used in clinical practice, with our new equation, we found that the performance of the equation developed by this research was superior in estimating LDL-c values for Koreans. Like previous studies,[8,9] we noted that LDL-c estimates calculated by the Friedewald equation tended to underestimate

https://doi.org/10.3346/jkms.2023.38.e145

Table 4. Comparison of root mean squared error values between measured LDL-c and estimated LDL-c levels according to TG levels (TG level < 400 mg/dL)

LDL-c estimation	TG levels			
	< 100 (n = 1,798)	100–149 (n = 961)	150–199 (n = 9.05)	200–400 (n = 2,066)
Model 1	5.97	6.40	7.43	10.00
Model 2	5.73	6.44	7.49	9.91
Friedewald et al.[6]	6.72	6.72	8.85	15.08
DeLong et al.[18]	8.00	7.34	8.20	10.40
Rao et al.[19]	6.80	6.53	7.88	10.97
Hattori et al.[20]	7.32	10.89	13.83	20.41
Anandaraja et al.[21]	19.35	12.32	12.46	16.66
Puavilai et al.[22]	7.73	7.00	7.90	10.60
Vujovic et al.[23]	8.34	7.83	8.75	10.57
Chen and Zhang[24]	6.64	7.50	7.81	10.07
de Cordova and de Cordova[25]	15.83	15.42	12.25	13.17
Martin et al.[8]	5.97	6.40	7.52	10.38
Sampson et al.[4]	15.83	14.93	15.84	16.81
Choi et al.[5]	6.99	6.61	7.55	10.58

LDL-c = low-density lipoprotein cholesterol, TG = triglycerides.

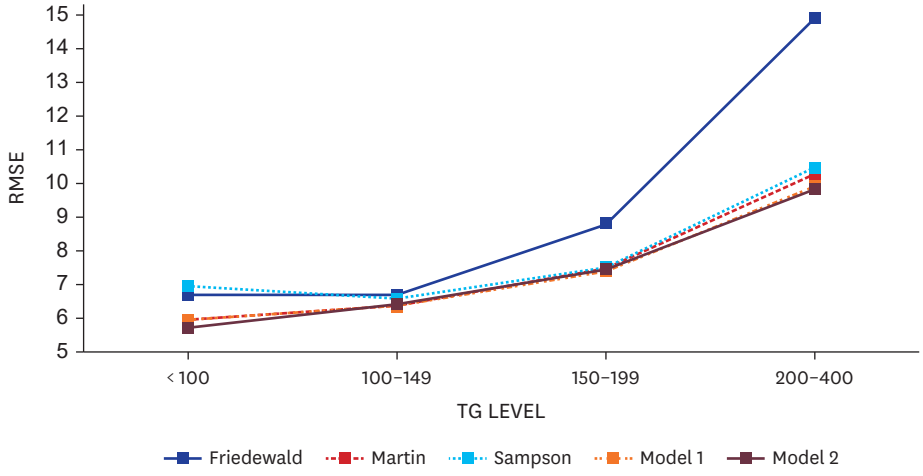

Fig. 3. Comparison of RMSE values between measured LDL-c and estimated LDL-c levels according to TG levels (TG level < 400 mg/dL).
RMSE = root mean squared error, LDL-c = low-density lipoprotein cholesterol, TG = triglycerides.

actual LDL-c values overall. In this study, predicted values from the Friedewald equation were underestimated by 20.2% and overestimated by 6.5% in classification according to the NCEP ATP III standard. The new equation showed values of 9.4% and 9.4%, respectively. Although there were a few cases of overestimation in the new estimation formula, it is more meaningful to reduce underestimation, which can lead to undertreatment for high-risk patients,[28] in terms of conservative clinical judgment.

Among the previously developed equations, the one with the most similar performance to the new equation was the Martin equation. In previous studies conducted in Korea, the Martin equation has been proven to show superior performance over other estimation equations, such as the Friedewald equation.[10,12,13] Lee et al.[13] reported that the Martin equation tends to relatively overestimate LDL-c values, and our study also showed similar results. The key strength of the Martin equation would be the ratio of TG to very low-density lipoprotein cholesterol (VLDL-c) values classified into 180 cells according to levels of TG and non-HDL-c. In this study, the Martin equation had a high misclassification rate with a slight difference

https://doi.org/10.3346/jkms.2023.38.e145

compared to the new equation, but there were relatively few underestimations and many overestimations. This point can be seen as a strength of the Martin equation, which was able to classify median values of the ratio of TG to VLDL-c values through a large sample. Since the median value of the ratio of TG to VLDL-c in Koreans is significantly different from the value suggested by Martin et al.,[28] if additional research with a large sample is conducted, it will be possible to develop an equation that is more suitable for Koreans.

The limitations of this study are as follows. First, external verification using data other than KNHANES data could not be conducted. Although we used five-fold cross-validation to prevent overfitting as much as possible, there is still a limitation with the absence of external validation. It is necessary to supplement this part by using more data, such as clinical data from hospitals, in the future. However, the possibility of bias due to overfitting to unrepresentative data is expected to be low, considering that the model developed in the previous study using the same data as our study maintained excellent performance during external verification.[17]

Second, there is a limitation in the data of the KNHANES itself. Although KNHANES uses a multi-stage stratified cluster sampling approach to ensure representative data, the methods (**Supplementary Data 1**) for selecting people with direct LDL-c measurements selected as subjects in this study varied by year.[29] In addition, it did not consider the administration of lipid-lowering drugs, such as statins and ezetimibe, which may affect differences in the accuracy of measured and estimated values or whether there was any preceding disease, such as hypertension, diabetes, or cardiovascular disease. Moreover, in general, it is necessary to fast for body blood tests in the KNHANES; therefore, the lipid test is conducted with most of the survey subjects maintaining a fasting state for more than 8 hours.[30] This strays from the global trend in which non-fasting tests are being increasingly recommended.[31,32] For more accurate research, research using data that can compensate for these limitations should be conducted.

One study reported misclassification rates of estimating LDL-c using the Friedewald equation and Martin equation for US residents of 14.6% and 8.3%, respectively.[8] Another study reported misclassification rates for the Friedewald, Martin, and Sampson equations of 12.6%, 11.0%, and 10.4%, respectively, in patients at the National Institutes of Health Clinical Center in the United States.[4] The fact that the misclassification rates of previous studies are significantly lower than the results of this study suggests that the equation developed in this study does not fully satisfy the function of being customized for Koreans. For example, the Choi equation is an estimation equation developed for Koreans. But the misclassification rate in the study exceeded 40%. Given this, research on developing an LDL-c equation suitable for Koreans should be continuously conducted, and it is necessary to conduct research based on representative samples in the future.

Recently, many studies have been conducted to estimate LDL-c using AI methods.[14,16,17] This study also developed an estimation model using XGBoost, one of the AI-based methods, in consideration of this aspect. As a result of comparing the formula developed through the conventional statistical method with the AI-based estimation model, the performance was similar. Due to the characteristics of the clinical field, a formula with a simple calculation process is preferred from a conservative point of view, so we expect that the formula developed based on regression analysis will be more appropriate for practical application.

The development of an LDL-c estimation formula for Koreans will have great implications in the aspects of public health, such as reducing the financial burden of national health insurance and contributing to the prevention and treatment of cardiovascular diseases.

SUPPLEMENTARY MATERIALS

Supplementary Fig. 1

Scatter plots of correlations between ground truth LDL-c values (direct LDL-c) and estimated LDL-c values in the cross validation 1 dataset using the (A) Friedewald, (B) DeLong, (C) Rao, (D) Hattori, (E) Anandaraja, (F) Puaviai, (G) Vujovic, (H) Chen and Zhang, (I) deCordova, (J) Martin, (K) Sampson, (L) Choi, (M) Model 1, and (N) Model 2.

Click here to view

Supplementary Fig. 2

Confusion matrix for estimated LDL-c values from the (A) Friedewald, (B) DeLong, (C) Rao, (D) Hattori, I Anandaraja, (F) Puaviai, (G) Vujovic, (H) Chen and Zhang, (I) de Cordova, (J) Martin, (K) Sampson, (L) Choi, (M) Model 1, and (N) Model 2.

Click here to view

Supplementary Data 1

Supplementary Methods

Click here to view

REFERENCES

1. 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* 2016;316(12):1289-97.
PUBMED | CROSSREF
2. Soran H, Dent R, Durrington P. Evidence-based goals in LDL-C reduction. *Clin Res Cardiol* 2017;106(4):237-48.
PUBMED | CROSSREF
3. Tremblay AJ, Morrissette H, Gagné JM, Bergeron J, Gagné C, Couture P. Validation of the Friedewald formula for the determination of low-density lipoprotein cholesterol compared with β-quantification in a large population. *Clin Biochem* 2004;37(9):785-90.
PUBMED | CROSSREF
4. 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(5):540-8.
PUBMED | CROSSREF
5. Choi R, Park MJ, Oh Y, Kim SH, Lee SG, Lee EH. Validation of multiple equations for estimating low-density lipoprotein cholesterol levels in Korean adults. *Lipids Health Dis* 2021;20(1):111.
PUBMED | CROSSREF
6. 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(6):499-502.
PUBMED | CROSSREF
7. 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:

executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol* 2019;73(24):3168-209.
PUBMED | CROSSREF

8. 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(19):2061-8.
PUBMED | CROSSREF
9. Meeusen JW, Lueke AJ, Jaffe AS, Saenger AK. Validation of a proposed novel equation for estimating LDL cholesterol. *Clin Chem* 2014;60(12):1519-23.
PUBMED | CROSSREF
10. Song Y, Lee HS, Baik SJ, Jeon S, Han D, Choi SY, et al. Comparison of the effectiveness of Martin's equation, Friedewald's equation, and a Novel equation in low-density lipoprotein cholesterol estimation. *Sci Rep* 2021;11(1):13545.
PUBMED | CROSSREF
11. Hata Y, Nakajima K. Application of Friedewald's LDL-cholesterol estimation formula to serum lipids in the Japanese population. *Jpn Circ J* 1986;50(12):1191-200.
PUBMED | CROSSREF
12. Kang M, Kim J, Lee SY, Kim K, Yoon J, Ki H. Martin's equation as the most suitable method for estimation of low-density lipoprotein cholesterol levels in Korean adults. *Korean J Fam Med* 2017;38(5):263-9.
PUBMED | CROSSREF
13. Lee J, Jang S, Son H. Validation of the martin method for estimating low-density lipoprotein cholesterol levels in Korean adults: findings from the Korea National Health and Nutrition Examination Survey, 2009–2011. *PLoS One* 2016;11(1):e0148147.
PUBMED | CROSSREF
14. 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.
PUBMED | CROSSREF
15. 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(9):e0239934.
PUBMED | CROSSREF
16. 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;517:108-16.
PUBMED | CROSSREF
17. Lee T, Kim J, Uh Y, Lee H. Deep neural network for estimating low density lipoprotein cholesterol. *Clin Chim Acta* 2019;489:35-40.
PUBMED | CROSSREF
18. DeLong DM, DeLong ER, Wood PD, Lippel K, Rifkind BM. A comparison of methods for the estimation of plasma low- and very low-density lipoprotein cholesterol. The Lipid Research Clinics Prevalence Study. *JAMA* 1986;256(17):2372-7.
PUBMED | CROSSREF
19. Rao A, Parker AH, el-Sheroni NA, Babelly MM. Calculation of low-density lipoprotein cholesterol with use of triglyceride/cholesterol ratios in lipoproteins compared with other calculation methods. *Clin Chem* 1988;34(12):2532-4.
PUBMED | CROSSREF
20. 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):289-99.
PUBMED | CROSSREF
21. 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):117-20.
PUBMED | CROSSREF
22. Puavilai W, Laorugpongse D, Deerochanawong C, Muthapongthavorn N, Srilert P. The accuracy in using modified Friedewald equation to calculate LDL from non-fast triglyceride: a pilot study. *J Med Assoc Thai* 2009;92(2):182-7.
PUBMED
23. Vujovic A, Kotur-Stevuljevic J, Spasic S, Bujisic N, Martinovic J, Vujovic M, et al. Evaluation of different formulas for LDL-C calculation. *Lipids Health Dis* 2010;9(1):27.
PUBMED | CROSSREF

24. 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):52.
PUBMED | CROSSREF

25. de Cordova CM, 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(Pt 1):13-9.
PUBMED | CROSSREF

26. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; Long Beach, CA, USA. San Francisco, CA, USA: Association for Computing Machinery; 2016, 785-94.

27. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd ed. New York, NY, USA: John Wiley & Sons; 2003.

28. Martin SS, Giugliano RP, Murphy SA, Wasserman SM, Stein EA, Ceška R, et al. Comparison of low-density lipoprotein cholesterol assessment by Martin/Hopkins estimation, Friedewald estimation, and preparative ultracentrifugation: insights from the FOURIER trial. *JAMA Cardiol* 2018;3(8):749-53.
PUBMED | CROSSREF

29. KDCA. Korea National Health and Nutrition Examination Survey items by year. https://knhanes.kdca.go.kr/knhanes/sub02/sub02_03.do. Updated 2022. Accessed June 7, 2022.

30. Kim S, Park E. Differences in height, weight, BMI, and obesity rate between 2018 Community Health and Korea National Health and Nutrition Examination Surveys. *J Health Inform Stat* 2020;45(3):281-7.
CROSSREF

31. Nordestgaard BG, Langsted A, Mora S, Kolovou G, Baum H, Bruckert E, et al. Fasting is not routinely required for determination of a lipid profile: clinical and laboratory implications including flagging at desirable concentration cut-points-a joint consensus statement from the European Atherosclerosis Society and European Federation of Clinical Chemistry and Laboratory Medicine. *Eur Heart J* 2016;37(25):1944-58.
PUBMED | CROSSREF

32. Pallazola VA, Quispe R, Elshazly MB, Vakil R, Sathiyakumar V, Jones SR, et al. Time to make a change: assessing LDL-C accurately in the era of modern pharmacotherapeutics and precision medicine. *Curr Cardiovasc Risk Rep* 2018;12(11):1-7.
CROSSREF