Original Article

A New Low-Density Lipoprotein Cholesterol Estimation Model from a Linear Regression Model and an Artificial Neural Network

Thaisiam P, MD[1], Sothornwit J, MD[2], Charoensri S, MD[2], Pattanapairoj S, BE[3], Kotruchin P, MD, PhD[4], Pongchaiyakul C, MD[2]

[1] Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[2] Division of Endocrinology and Metabolism, Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[3] Department of Industrial Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand

[4] Department of Emergency Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

Background: Low-density lipoprotein cholesterol (LDL-C) estimation from Friedewald equation is frequently used in clinical practice. However, limitations have emerged regarding its use, including patients with triglyceride (TG) levels of more than 400 milligrams per deciliters (mg/dL), or LDL-C level of less than 70 mg/dL. Despite that many new LDL-C equation models derived from linear regression analysis have been proposed, the accuracy of these generated formulas is still questionable. The authors developed a new LDL-C prediction model constructed by an artificial neural network (ANN), an information processing and computational system modeled after a biological nervous system, with expected better accuracy than both the Friedewald equation and the linear regression models.

Materials and Methods: A cross-sectional study was conducted. Serum lipid profiles (total cholesterol [TC], TG, high-density lipoprotein cholesterol [HDL-C], and LDL-C) were collected from 10,949 participants irrespective of specimen collection time, time since last caloric intake, comorbidities, and current medications. Direct LDL-C measurement determined by homogeneous assay was considered as the gold standard. Data were randomly divided into two cohorts, one for developing an equation from a linear regression model and ANN model, and another for validation and analyzing the predictive accuracy among the Friedewald equation, linear regression, and ANN model.

Results: The new simple equation derived from the linear regression model was 0.9 TC – 0.1 TG – 0.8 HDL-C. The correlation coefficient between direct LDL-C measurement and Friedewald-calculated LDL-C, LDL-C calculated using linear regression, and ANN-calculated LDL-C were 0.966 ($p<0.001$), 0.977 ($p<0.001$), and 0.978 ($p<0.001$), respectively. The ANN model demonstrated less root mean square error (RMSE) than the Friedewald equation or the linear regression model, which implied better accuracy, even when TG levels were more than 400 mg/dL or direct LDL-C levels were less than 70 mg/dL.

Conclusion: The ANN model is a highly accurate and a better LDL-C estimating tool, even in patients with TG level greater than 400 mg/dL and LDL-C level less than 70 mg/dL.

Keywords: Artificial neural network, Friedewald, Equation, Lipid, Low-density lipoprotein cholesterol

Received 26 Nov 2019 | Revised 4 Feb 2020 | Accepted 11 Feb 2020

J Med Assoc Thai 2020;103(4):346-52
Website: http://www.jmatonline.com

Dyslipidemia, particularly elevation of low-density lipoprotein cholesterol (LDL-C) level is a major risk factor for coronary heart disease and cardiovascular diseases (CVD)[1,2]. There is good evidence that screening for lipid disorders can identify asymptomatic adult eligible for preventive therapy[3]. Indeed, LDL-C can be directly measured by different methods; however, due to limited availability and high cost of direct LDL-C measurement, LDL-C estimated by the Friedewald equation as total cholesterol (TC) – high-density lipoprotein cholesterol (HDL-C) – [triglyceride (TG)/5] in milligrams per deciliters (mg/dL), is typically used in clinical and research setting worldwide[4]. It is known that the Friedewald

Correspondence to:

Sothornwit J.

Division of Endocrinology and Metabolism, Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand.

Phone: +66-43-363664

Email: jinso@kku.ac.th

How to cite this article: Thaisiam P, Sothornwit J, Charoensri S, Pattanapairoj S, Kotruchin P, Pongchaiyakul C. A New Low-Density Lipoprotein Cholesterol Estimation Model from a Linear Regression Model and an Artificial Neural Network. J Med Assoc Thai 2020;103:346-52.

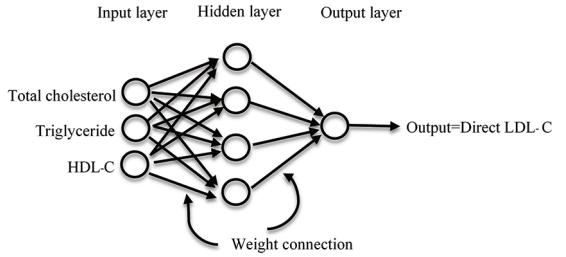

Figure 1. The single-hidden-layer feed-forward neural network. The network consists of one input layer, one hidden layer, and one output layer. Each node is linked by weighted connections, which the ANN uses to map the relationship between the input and output. During the training process, the weighted connections were adjusted until the relationship between the input and output was well defined.

formula has some limitations and is less accurate in patients with TG level greater than 400 mg/dL or LDL-C level of less than 70 mg/dL, and in those with type 3 dyslipoproteinemia[5,6]. Previous studies demonstrated considerable differences between direct measurement of LDL-C and the estimation in many other conditions[7-9]. Furthermore, the Friedewald equation was derived based on samples obtained in a fasting state, so is not suitable for use in those who did not fasted[4].

Although, several new LDL-C equations including the equations from Thai population[10-12] have been developed, most of which were derived from linear regression analysis[10-19], the accuracy of LDL-C prediction with each equation varies among populations[18]. In a previous study in Thailand, Rungtanapirom et al[11] reported LDL-C level from direct measurement was about 13.4% higher than calculated LDL-C level, while Puavilai et al[10], found that the differences of LDL-C levels between direct measurement and a new equation were 8.4% and 16.7% (more than 10 mg/dL) in subjects with TG levels of 200 to 299 and 300 to 399 mg/dL, respectively.

Recently, computer-based diagnostic systems, i.e., artificial neural network (ANN) have been applied to simulate the combined-diagnostic models or equations from the measured biological markers. ANN is the data processing and computational systems that was inspired by the study of biological neural processing. It can process, learn, and remember information[20,21]. The authors constructed an LDL-C prediction model using an ANN with expected better accuracy than Friedewald equation in predicting LDL-C using direct LDL-C measurement as a gold standard. In the present study, the authors also compared the accuracy of LDL-C prediction model between ANN and the equation model derived from linear regression analysis.

Materials and Methods

The present study was carried out in accordance with the recommendations of the Center for Ethics in Human Research, Khon Kaen University (HE541027). This was a cross-sectional study conducted in Srinagarind Hospital, a tertiary care setting in Khon Kaen, Thailand, between November 2010 and January 2011. The data were collected from all subjects who participated and gave a written informed consent prior to lipid measurement. A single measurement of all four lipid profiles (TC, HDL-C, TG, LDL-C) was collected irrespective of specimen collection time, time since last caloric intake, comorbidities, and current medications. Demographic characteristics, including sex and age, and results of lipid profiles were recorded.

TC, TG, and HDL-C were measured using enzymatic methods with an automatic autoanalyzer (Cobas Integra 800; Roche Diagnostics, Mannheim, Germany). LDL-C was measured directly in mg/dL by homogeneous assay using a Hitachi c6000 chemical analyzer as the gold standard in the present study. Friedewald LDL-C was calculated based on the Friedewald equation: TC – HDL-C – (TG/5). The participants (n=10,949) were randomly divided into two cohorts at a ratio 1:1 using a computerized random number generator (development cohort: n=5,477; validation cohort: n=5,472). In the development cohort, a new equation was developed using linear regression based on three lipid profiles (TC, TG, and HDL-C), and the training set for the LDL-C prediction model was constructed using an ANN with 10-fold cross validation. A single-hidden-layer feed-forward neural network, which was the simplest type of ANN, was used in the present study. This neural network consisted of one input layer (TC, TG, HDL-C), one hidden layer, and one output layer (direct LDL-C). Each node (input node, hidden node, and output node) was linked by weighted connections (Figure 1), which the ANN used to map the relationship between the input and output. During the training process, the weighted connections were adjusted until the relationship between the input and output was well defined. In the validation cohort, the associations and differences between the Friedewald LDL-C equation, linear regression model, and ANN model were analyzed, and the predictive accuracy of each was evaluated.

Table 1. Participants' baseline characteristics

	Development group (n=5,477) n (%)	Validation group (n=5472) n (%)	p-value
Sex distribution			
Male	2,553 (46.6)	2,510 (45.9)	0.44
Female	2,924 (53.4)	2,962 (54.1)	0.44
Age (years); mean±SD	54.8±13.9	54.9±14.1	0.66
Lipid profile; mean±SD			
Total cholesterol (mg/dL)	203.4±49.4	204.2±49.5	0.40
HDL-C (mg/dL)	52.5±16.4	52.6±16.3	0.75
Triglyceride (mg/dL)	154.4±109.4	152.8±99.7	0.42
Direct LDL-C (mg/dL)	126.8±41.4	127.5±42.1	0.38
Triglyceride levels (mg/dL)			
<100	1,711 (31.2)	1,778 (32.5)	0.16
101 to 200	2,641 (48.2)	2,597 (47.5)	0.43
201 to 300	766 (14.0)	745 (13.6)	0.57
301 to 400	206 (3.8)	210 (3.8)	0.83
>401	153 (2.8)	142 (2.6)	0.52

SD=standard deviation; HDL-C=high-density lipoprotein cholesterol; LDL-C=low-density lipoprotein cholesterol

Statistical analysis

Statistical analysis was performed using SPSS Statistics for Windows, version 17.0 (SPSS Inc., Chicago, Ill, USA). Pearson's correlation coefficient was used to evaluate the relationship among the LDL-C levels derived from the Friedewald equation, linear regression model, and ANN model. Differences were demonstrated using the Bland-Altman method and expressed as mean differences and limited of agreement. The predictive accuracy of each model was determined using the root mean square error (RMSE).

Results

Ten thousand nine hundred forty-nine participants underwent single lipid profile measurements during the present study period and 5,477 were randomly assigned to the development cohort while 5,472 were assigned to the validation cohort. There were no significant differences in demographic characteristics (age, sex, and lipid profile) between the development and validation groups (Table 1). One hundred fifty-three (2.8%) and 142 (2.6%) participants had TG levels of at least 400 mg/dL in the development and validation group, respectively.

Developing a model using linear regression analysis

A new equation was constructed based on linear regression analysis in the development group: 0.917 cholesterol – 0.108 TG – 0.794 HDL-C – 1.301. The equation was simplified as 0.9 cholesterol – 0.1 TG – 0.8 HDL-C for more convenient use in clinical practice.

Developing a model using an artificial neural network

An LDL-C prediction model using an ANN was constructed based on the following settings, four hidden units, sigmoid function, and 5,000 learning cycles. This model was built using the open-source machine-learning software, Rapid Miner Studio version 9.2. The results were evaluated with validation models constructed using k-fold cross validation, which is widely used in medical field. As k=10 in the present study model, all data samples were randomly divided into ten subsets. Each subset was used as a validation set and the remaining nine subsets were put as the training sets. The validation process was repeated ten times, and each of the subsets was used as the validation set only once (Figure 2). This method was used to reduce the bias between the training and

	All Data									
	Training Set									Validation Set
Fold 1	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 2	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 3	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 4	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 5	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 6	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 7	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 8	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 9	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Fold 10	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10

Figure 2. 10-fold cross validation. To reduce the bias between the training and validation sets, all data samples were randomly divided into ten subsets. Each subset was used as a validation set (subset in the grey area) and the remaining nine subsets were put as the training sets. The validation process was repeated ten times, and each of the subsets was used as the validation set only once.

Table 2. Diagnostic accuracy among models

Methods	Equation	RMSE	Correlation coefficient	p-value
Friedewald	TC – HDL – (TG/5)	12.80	0.966	<0.001
Linear regression	0.9 TC – 0.1 TG – 0.8 HDL	9.20	0.977	<0.001
Artificial neural network	None	8.96	0.978	<0.001

TC=total cholesterol; HDL=high-density lipoprotein cholesterol; TG=triglyceride; RMSE=root mean square error

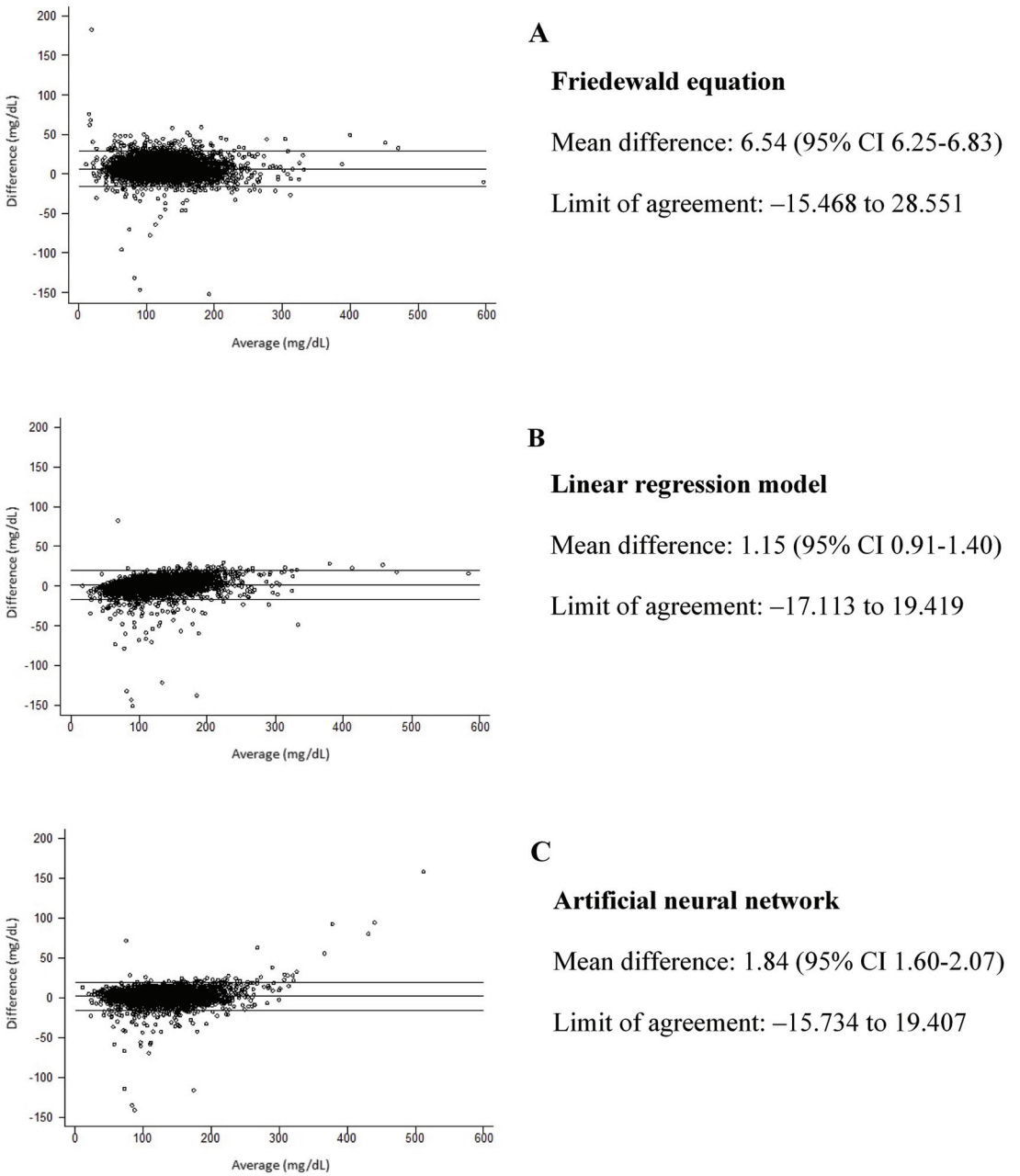

Figure 3. Comparison among models using the Bland-Altman method; Friedewald LDL-C equation, and direct LDL-C measurement (A), linear regression model and direct LDL-C measurement (B), ANN model and direct LDL-C measurement (C).

Table 3. Diagnostic accuracy among models based by triglyceride levels

Triglyceride levels (mg/dL)	Friedewald RMSE	Linear regression RMSE	ANN RMSE	p-value
<200	9.49	7.90	7.55	<0.001
200 to 400	18.18	9.20	10.12	<0.001
>400	36.27	27.83	25.20	<0.001

ANN=artificial neural network; RMSE=root mean square error

Table 4. Diagnostic accuracy among models based by LDL levels

Direct LDL levels (mg/dL)	Friedewald RMSE	Linear regression RMSE	ANN RMSE	p-value
<70	18.91	19.29	16.97	<0.001
70 to 99	10.97	8.34	7.37	<0.001
100 to 129	16.43	10.42	7.33	<0.001
130 to 159	11.91	7.76	7.27	<0.001
160 to 189	16.72	11.03	7.12	<0.001
>190	16.11	16.33	16.11	<0.001

LDL=low-density lipoprotein cholesterol; ANN=artificial neural network; RMSE=root mean square error

validation sets, even though the samples size was small.

The correlation coefficients between the Friedewald LDL-C equation, linear regression model, and the ANN model with direct LDL-C were 0.966 ($p<0.001$), 0.977 ($p<0.001$), and 0.978 ($p<0.001$), respectively (Table 2). Using the Bland-Altman method, the authors determined that the mean differences between the reference method and the Friedewald LDL-C equation, linear regression model, and ANN model were 6.54 (95% CI 6.25 to 6.83), 1.15 (95% CI 0.91 to 1.40), and 1.84 (95% CI 1.60 to 2.07), respectively (Figure 3). The diagnostic accuracy of the Friedewald equation and developed models are compared in Table 2, and the diagnostic accuracy among models based by TG levels and LDL-C levels are compared in Table 3 and 4, respectively.

Discussion

Several disadvantages of the Friedewald equation in LDL-C estimation limit its use in clinical practice. For instance, the Friedewald equation is underestimated and has lower accuracy in the setting of TG levels of more than 400 mg/dL or LDL-C level of less than 70 mg/dL[4]. Although, several equations for more valid LDL-C estimation have been modified from the conventional Friedewald formula using linear models, the accuracy of LDL-C prediction with each equation varies among populations and are complicated[10-19]. In Thailand, Rungtanapirom et al proposed two new equations in 2008, including age, sex, and body mass index (BMI) as the explanatory variables in the equation [0.98 TC – 0.84 HDL – 0.12 TG + 0.056 age + 0.071 BMI and 0.98 (TC – HDL) – 0.12 TG + 0.1 age + 2.4 sex + 0.2 BMI; values for sex are male=1, female=2][11]. In 2009, Puavilai et al showed that a modified Friedewald equation (TC – HDL – 1/6 TG) had 83.8% accuracy when compared to direct measured LDL ±10 mg[10]. However, these equations still pose a limitation in patients with severe hypertriglyceridemia[10,12]. In the present study, the authors proposed a new method for generating LDL-C estimation model, the ANN model, with the highest accuracy (r=0.978, RMSE=8.96) compared with the Friedewald equation (r=0.966, RMSE=12.80) and a linear regression equation (r=0.977, RMSE=9.20).

In the present study, ANN is a set of algorithms designed to recognize patterns by interpreting sensory data through a kind of machine perception. During the learning process, the neural network finds the right algorithm for transforming input (TC, TG, and HDL-C levels) into output (calculated LDL-C levels) so that as new cases and datasets are collected, the model adjusts itself and becomes more accurate and generalizable. An ANN can model both linear and non-linear relationships among variables and thus, usually outperforms multiple regression and related techniques[21]. Therefore, an ANN model can be adopted by any laboratories to predict LDL-C in different ethnics using lipid data from their subjects and results in more accurate LDL-C estimation in those specific populations.

Because the data in the present study were collected from participants regardless of specimen collection time, time since last caloric intake, comorbidities, and current medications, the present study new LDL-C equation and model can be used in patients with diabetes or metabolic syndrome, those using lipid-lowering medication, and in both fasting and non-fasting samples. The authors also found that the ANN model was more accurate than the Friedewald equation for predicting LDL-C levels in patients with TG levels greater than 400 mg/dL and LDL-C less than 70 mg/dL (RMSE 25.2 versus 36.3 and 16.9 versus 18.9, respectively). The present findings must be interpreted in the context of several potential strengths and weaknesses. The strengths of the present study include the random selection

process, large sample size, the use of direct LDL-C measurement as a gold standard, and the applicability of the model in patients with comorbidities or concurrent medication use and in both fasting and non-fasting specimens. However, participants were all Thai, and so care should be taken when extrapolating these results to other populations in which genetic, lifestyle, or environmental factors differ. Moreover, lipid profiles were measured at a single time point and may not reflect patients' long-term lipid profiles. In addition, stratification of the model by comorbidities and concurrent medication was not performed. Measurement errors could have also affected the results, but this is a limitation in any study of this type.

Conclusion

The model generated using an ANN had high accuracy and may be more accurate in predicting LDL-C levels in patients with TG levels that are greater than 400 mg/dL and LDL-C levels that are less than 70 mg/dL.

What is already known on this topic?

Direct LDL-C measurement is expensive and not available in some laboratories. LDL-C level estimation from TC, HDL-C, and TG levels is a cost-saving method. However, the Friedewald equation and of those derived from linear regression models are not accurate in patients with TG levels greater than 400 mg/dL and LDL-C levels of less than 70 mg/dL.

What this study adds?

This study developed a new LDL-C prediction model constructed using an ANN with better expected performance than that of both the Friedewald equation and of those derived from linear regression models. This new method has better accuracy in predicting LDL-C levels in patients with TG levels greater than 400 mg/dL and LDL-C levels of less than 70 mg/dL.

Ethics consideration

The present study was carried out with the approval of the Khon Kaen University Center for Ethics in Human Research. Written informed consents were obtained from all participants in accordance with the Declaration of Helsinki.

Acknowledgement

The authors would like to thank Dylan Southard for revising the English-language version of this manuscript.

Conflicts of interest

The authors declare that they have no conflict of interest.

References

1. Nelson RH. Hyperlipidemia as a risk factor for cardiovascular disease. Prim Care 2013;40:195-211.
2. Keevil JG, Cullen MW, Gangnon R, McBride PE, Stein JH. Implications of cardiac risk and low-density lipoprotein cholesterol distributions in the United States for the diagnosis and treatment of dyslipidemia: data from National Health and Nutrition Examination Survey 1999 to 2002. Circulation 2007;115:1363-70.
3. U.S. Preventive Services Task Force. Screening for lipid disorders in adults: recommendation statement. Am Fam Physician 2009;80:1273-4.
4. 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.
5. 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.
6. Roberts WC. The Friedewald-Levy-Fredrickson formula for calculating low-density lipoprotein cholesterol, the basis for lipid-lowering therapy. Am J Cardiol 1988;62:345-6.
7. Mora S, Rifai N, Buring JE, Ridker PM. Comparison of LDL cholesterol concentrations by Friedewald calculation and direct measurement in relation to cardiovascular events in 27,331 women. Clin Chem 2009;55:888-94.
8. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, et al. Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. Circulation 2005;112:2735-52.
9. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). JAMA 2001;285:2486-97.
10. Puavilai W, Laoragpongse D. Is calculated LDL-C by using the new modified Friedewald equation better than the standard Friedewald equation? J Med Assoc Thai 2004;87:589-93.
11. Rungtanapirom S, Ongphiphadhanakul B, Kiat-bamrungpunt T, Bhakdikul J, Archavanuntakun S, Taerakul C, et al. Comparative study of direct-measured and calculated LDL in clinical use. J Med Assoc Thai 2008;91:989-94.
12. 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;92:182-7.
13. 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;256:2372-7.
14. 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;102:117-20.
15. 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;9:52.
16. 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;9:27.
17. 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;50:13-9.
18. Dansethakul P, Thapanathamchai L, Saichanma S, Worachartcheewan A, Pidetcha P. Determining a new formula for calculating low-density lipoprotein cholesterol: data mining approach. EXCLI J 2015;14: 478-83.
19. Ephraim RKD, Acheampong E, Swaray SM, Odame AE, Agbodzakey H, Adoba P, et al. Developing a modified low-density lipoprotein (M-LDL-C) Friedewald's equation as a substitute for direct LDL-C measure in a ghanaian population: a comparative study. J Lipids 2018;2018:7078409.
20. Razi MA, Athappilly K. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Syst Appl 2005;29:65-74.
21. Somers MJ, Casal JC. Using artificial neural networks to model nonlinearity: the case of the job satisfaction--job performance relationship. Organ Res Methods 2009;12:403-17.