Estimation of Mathematical Models Accuracy for Calculation of LDL-Cholesterol Concentration

Vladimir V. Kuz'menko[1], Alexander Yu. Gornov[2], Anton S. Anikin [2]

[1] Department of clinical laboratory diagnostics, ISMAPgE – Branch Campus
of the FSBEI FPE RMACPE MOH Russia,
Yubileynyy, 100
Irkutsk, Russia
E-mail: kw7@mail.ru

[2] Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of the Russian Academy of Sciences
Lermontov str., 134
Irkutsk, Russia
E-mail: gornov@icc.ru

Abstract

During investigation the accuracy of the Friedewald's calculation method in case method of ultracentrifugation was been replaced by the method of photometry. There was investigated the possibility and limitations of using Shepard's method for calculating the concentration of low density lipoprotein (LDL) cholesterol. A comparison of accuracy various mathematical models was given to calculate the concentration of LDL cholesterol. The computational methods for calculating the quantities considered LDL cholesterol in this paper are not inferior in their characteristics to analytical methods.

Keywords: Mathematical model, LDL cholesterol, Shepard's method.

1. Introduction

The selection of the optimal method for modeling biological parameters depends on a number of characteristics of the model described. Such characteristics include the number of parameters which evaluated their limitations, the number of observations, and the presence of drop-out values. Taking these features into consideration allows us to select the most suitable methods for constructing the model, however, for the final choice, a computational experiment is necessary.

Determination of laboratory parameters of lipid metabolism in blood serum, in particular cholesterol fractions, is an important criterion for diagnosis, prognosis assessment and treatment tactics for a number of endocrine and cardiovascular diseases [1, 2]. These studies are included in the standards of diagnosis of patients, which in turn caused an increase in the load on the laboratory, as well as the cost of patient examination. When the lipid fractions in the blood instead of direct determination of low density lipoprotein (LDL) cholesterol fraction are assessed, as a rule the cost of patient examination reduce and a calculation method by W. T. Friedewald [3, 4] is used in the laboratory.

The Friedewald's formula allows you to calculate the concentration of cholesterol of this lipid fraction accurately when you use the technology of separation of cholesterol fractions by ultracentrifugation. At present, the ultracentrifugation method is not applied due to its duration and complexity of automation; however, clinical laboratories often continue to use the calculation method to estimate the concentration of LDL cholesterol.

It is known that Friedwald's technique has a number of significant limitations [5, 6]. One of them is low accuracy in case of high plasma triglyceride concentration (more than four g/l), which prevents dynamic monitoring of the effectiveness of therapy in patients. In addition, the use of the formula is based on the assumption that the ratio of triglycerides and cholesterol of very low density lipoproteins is stable and amounts five to one. Finally, this method can have a significant error if a significant amount of chylomicrons appears in the blood, if the concentration of which in most laboratories is not determined. There is no certainty that the laboratory staffs is aware of these limitations described by the author of the methodology, and you should take them into consideration in their work using the calculation method. It is safe to assume that this technique, like any other indirect method, will have a lower accuracy of the results compared to the direct definition of the same indicator. At the same time, the application of the calculation method is attractive because it makes possible to reduce the cost of a laboratory examination by refusing to analyze one indicator.

We performed a modeling of concentration of LDL cholesterol in conditions of the evaluation of parameters of lipid metabolism in modern laboratory equipment (Modular, Roche Diagnostics).

2. Goal and Tasks

The aim of the study was to search for a method for modeling the concentration of LDL cholesterol, which gives a minimum error in its application.

To achieve this goal, the following tasks were accomplished:

a. to examine the acceptability of the Friedwald's technique under the conditions of replacing the ultracentrifugation method to determine the fractions of lipoproteins by the photometry method,
b. evaluate the linear regression technique for calculating the amount of LDL cholesterol,
c. to examine the acceptability of the Shepard's method for the determination of lipoprotein fractions.

3. Materials and Methods

For the calculations, a sample of research results was prepared, consisting of 4 384 observations. It included the results of a laboratory examination of patients in which a concentration of the following parameters was determined from a single serum sample for a short period of time: triglycerides, total cholesterol, high-density lipoprotein (HDL) cholesterol and LDL cholesterol. Laboratory studies were performed on the Modular P-800/ISE analyzer from Roche Diagnostics (Switzerland) using reagents and standard samples from the same company. In addition, in parallel with laboratory analyzes, concentrations of LDL cholesterol and very low density (VLDL) cholesterol were calculated using the formula proposed by Fridwald. The calculation of the linear regression formula was carried out using the STATISTICA data analysis software version 6.0.

The Shepard's method [7, 8], which is similar to the inverse distance method and the method of radial basis functions, widely used in neuromodulation, has rarely been used in the study of data analysis problems. The main area of its application was the processing of cartographic information, although the capabilities of this method, in our opinion, are much wider [9]. The Shepard's method involves the construction of an interpolant in the form of a ratio of two fractional rational functions based on experimental data - a training sample. The low operational complexity of the corresponding algorithms makes it possible to use it for processing tables of experimental data with a large (thousands and tens of thousands) number of indicators and precedents. The proposed method involves two-time application of the Shepard method, at the first stage to improve the quality of input data (their "cleaning"), and at the second – to build models of the dependence under study. The algorithms of the proposed technique are implemented in C/C++ and operate under Windows and Linux OS.

The effectiveness of models which improves the accuracy of the calculation of LDL cholesterol was evaluated by comparing systematic and random errors in the calculations performed on these models. To compare the significance of differences in mean values and variances of analyzed samples, we used the MS Excel data analysis package.

4. Results

First stage. To estimate the error of the Friedwald's method, we calculated the difference in absolute and relative (%) units between the calculated cholesterol concentration in LDL composition and the values obtained directly during the analysis at the number of observations n=4384. We have found that the average systematic error of the values calculated by the Friedwald's formula is -11.49 %, while the dispersion of the relative errors was 818.36 (table.1).

Such significant random and systematic errors can be explained by the imperfection of the indirect method of determination, as well as by the using of triglycerides, total cholesterol and HDL cholesterol obtained by using other analytical methods for calculations than the authors of the widely used formula. If the original Friedwald's formula reflects the regularities of the distribution of cholesterol in different lipid fractions in the allocation of their method of ultracentrifugation [4] rather accurately, so in order to determine the fractions of cholesterol by the photometric method on the analyzer Modular calculation formula needs to be update.

The validity of the observation is confirmed by the fact that the concentration of total cholesterol in the blood serum was lower by an average of 26.1% than the sum of cholesterol concentrations determined in different lipid fractions (VLDL cholesterol, calculated value and LDL cholesterol, HDL cholesterol) of the same sample, although these values should be equal. Regardless of the analytical methods used, the measured amount of total cholesterol as a whole cannot be less than the total amount in individual fractions. Since in these calculations the amount of VLDL was calculated in accordance with the Friedwald's recommendations [3] as the number of triglycerides in the sample in mmol/l divided by 2.22, it can be assumed that the cause of the revealed discrepancies between the calculated cholesterol concentration in the sample and the determined laboratory route is in the value of this coefficient.

As a next step in improving the accuracy of VLDL cholesterol calculation, we made an attempt to clarify the coefficients in the formula which was used before. The amount of VLDL cholesterol was calculated to do this as the difference between total cholesterol and the amount of cholesterol in the fractions of HDL and LDL. The calculations showed that its average value in the serum of patients was 0.35 mmol/L. This quantity is approximately 6.23 per cent of total cholesterol. Based on these calculations we can assume that in our measurements on the Modular analytical system together with LDL cholesterol and intermediate density lipoproteins we detected a significant part of VLDL cholesterol.

Second stage. The possibility of usage the multiple regression method is analyzed to refine the calculation formula. The calculations are performed using the STATISTICA software package. First of all, an attempt was made to test the formula calculated [10] on a random sample consisting of 212 observations on all 4384 observations. According to the results given in table 1, the average relative error decreased by 11 times in comparison with the Friedwald's method. The paired two-sample t-test for the means demonstrated the presence of significant differences ($p < 0.001$) between the test and the initial methods of calculation. Unfortunately, the variance was significant when during building the model was used the regression method.

Then the coefficients were calculated on a large sample (n=4384), with the coefficient of multiple correlation (R), which reflected that the degree of dependence of the concentration of LDL cholesterol on other variables was equal to 0.951.

Table 1. Parameters of relative error in the calculation of X-LDL, obtained using different calculation methods

Methods of mathematical modeling	Average relative error of calculation, %	Varian ce
Friedwald method (n=4384)	-11.49	818.36
Formula obtained using the linear regression method for n=212	1.05	5546
Formula obtained using the linear regression method for n=4384	0.23	4989
Method based on the shepherd operator (before «cleaning») (n=212)	3.95	452.24
The technique is based on the Shepard operator (after "cleaning") (n=202)	1.16	121.76

A value of R is close to 1.0 indicates so the model explains almost all the variability of the corresponding variables. The application of the presented formula allowed us to reduce the average relative error of calculation of LDL cholesterol to 0.23 % and the variance decreased by 10 %, but it still remained large (table 1). The use of two-sample F – test allowed us to verify the significance of differences in the variance of compared samples at the significance level of 0.001. The presence of large variance in the models was constructed using the multiple regression method which prompted us to search for alternative methods of calculations.

Third stage. To improve the accuracy of the method, we reduced the random systematic errors of the calculation method and the method of creating static models based on the Shepard operator. The quality of the model was checked using the Committee method. In the course of this work we calculated the index of interest (LDL cholesterol) using a multidimensional approximated function constructed by the Shepard method on all other elements of the initial sample. Since the true values of LDL cholesterol for all elements are known, calculations of the relative error of the described model were carried out for all elements of the initial sample. The algorithm of quality evaluation of the obtained model:

Algorithm 1 ("Model Testing").

a. the I-th element is removed from the selection,
b. the remaining part of the sample we construct the model Shepherd,

c. using the resulting model, calculate the value of the function at the point corresponding to the removed element,
d. the relative error of the obtained value is calculated,
e. the deleted item is returned to the selection,
f. the listed operations are repeated for all the selection elements,
g. the resulting errors are used to calculate the average, minimum and maximum errors, as well as the variance.

To reduce the number of calculations in the course of assessing the possibility of using the Shepard's method, we conducted computational experiments on a test sample consisting of 212 observations. At the same time all the calculations were carried out on an ordinary personal computer without using of parallel computing technologies. The results of the calculations according to the following parameters of the relative error are in table 1: the arithmetic mean of the relative error of calculation is 3.94% and the variance is 452.23. The average error of calculations during usage of the Shepherd's method was significantly less than average error than using the Friedewald method ($P < 0.001$).

Unfortunately, these results cannot be considered as satisfactory ones, because there was a high dispersion, which was characterized by the presence of points with a high error. At the same time for some patients there was an unacceptably large deviation of the calculated value of LDL cholesterol from the measured made in the laboratory, and despite the number of such elements is small, their presence dramatically reduces the possibility of practical applicability of the created model.

According to the experience of previous works z to carry out so-called "horizontal cleaning", which consists in the directed removal of the sample elements that make the greatest "noise", thereby increasing the error of the created model. Performing this operation in some cases allows you to improve the quality of the model significantly (Fig.1). The maximum number of elements which is to be removed during the "cleaning" was selected from the "95% rule" ("rule 2 Sigma") - i.e. the model was based on the assumption that the majority of patients have common patterns and relationships and those elements that we remove (5%) are a separate group that needs to be considered in more detail and create a different model taking into consideration its features. With 212 patients the maximum number of removed sample points was 10 observations. The results of testing of the created model are shown in table 2. The algorithm of "cleaning" of the model is next:

Algorithm 2 ("Cleaning the Training Sample»):
a. the I-th element is removed from the selection,
b. Shepard's model is built on the rest of the sample,
c. the model is tested (algorithm 1),
d. test results are saved,
e. the deleted item is returned to the selection,
f. these operations are repeated for all selection items,
g. the element is selected, so the removal makes the necessary characteristic of the model much better.

In this paper, we decided to minimize the dispersion, but the algorithm allows you to choose any other valid parameter, for example, the average error and others.

Table 2. Change of relative error and its dispersion in the process of "cleaning" (n=212)

The number of removed samples	Parameters in the error model	
	Arithmetic mean, %	Variance
0	3.94	452.24
10	1.16	121.75

After analyzing the results of testing performed on 212 cases (table 2) it can be argued that the procedure of "cleaning" has allowed to reduce the arithmetic mean significantly and the dispersion of the error of the model by 3-4 times. The above-described computational experiments have shown the feasibility of calculations on a sample with a large number of observations.

Fourth stage. Further modeling was carried out on a full sample of 4384 observations. At this stage the testing of models and the implementation of the procedure of their "cleaning" required significant computational resources. As it can be seen from the above algorithms, during the procedure of "cleaning" the number of model constructions and calculations of the function values depend quadratically on the sample size:

$$K = N_2 * N_1 = n * (n — 1) = n^2 - n, \quad (1)$$

where n is the sample size, N_2 is the number of iterations of algorithm 2, N_1 is the number of iterations of algorithm 1. If the number of observations is equal to 4384, the number of iterations will be: $4384^2 – 4384$. Thus, with increasing the sample size, is the number of elements in its composition according to several thousand observations, the processor time of the "cleaning" operation can grow to unacceptable values (days and weeks). To speed up the calculations we implemented a modified parallel algorithm, which allows us to use several core processors in SMP-systems

(Symmetric Multiprocessing).

Initial testing of the model showed that its characteristics, in particular, the dispersion of the relative error, cannot be considered as satisfactory ones. Their quality was tested for all models and the subsequent "cleaning" was performed. In addition we tested the hypothesis that errors can be reduced by dividing the initial sample into several parts and then building models on each of them. The first option is considered to divide all patients into 2 groups — men and women. For all models their quality was tested with subsequent "cleaning". 1% of the total number of elements was selected as the "clean" border - "rule 3 Sigma" was used.

Using the proposed technique and implemented algorithms we investigated the possibilities of improving the accuracy of calculation methods focused on the assessment of cholesterol concentration in fractions. The multi-variant computational experiments have proved the efficiency of the studied approaches. It is confirmed that during performing calculations on the model based on the Shepherd’s method, the operation of "cleaning" of the initial sample can significantly improve the quality of the models (figure 2).

The hypothesis of the expediency of the sample division into parts and the construction of independent models on them is also partially confirmed by the computational experiment. As it can be seen from table 3 the allocation of women in a separate model allowed to reduce the variance of the error by an order of magnitude immediately. A separate model for men, on the other hand, has much worse performance than the model of women and the general model of all patients. However, during the procedure of "cleaning" its characteristics can also be brought into the acceptable range. Note that the procedure of sample division does not allow to improve the resulting models only, but it also significantly reduces the total calculation time, since the dependence of the calculation time on the sample size is quadratic.

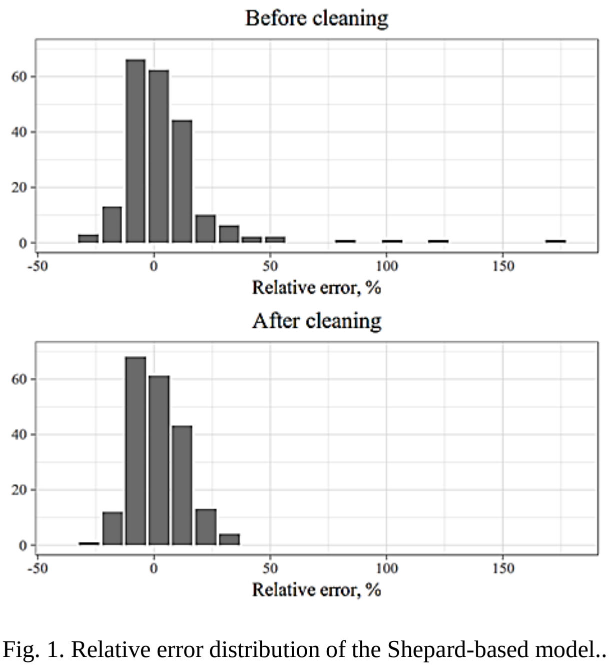

Fig. 1. Relative error distribution of the Shepard-based model..

Table 3.The parameters of relative errors in the calculation of LDL cholesterol, obtained using the Shepard method on a large sample (n=4384).

	Arithmetic mean of the relative error of the model (%)			The variance of the error		
	All	Men	Women	All	Men	Women
Before «cleaning»	5.01	8.73	1.87	31231.4	62430.9	200.36
After «cleaning» (1%)	1.45	2.01	1.40	114.72	142.73	105.80

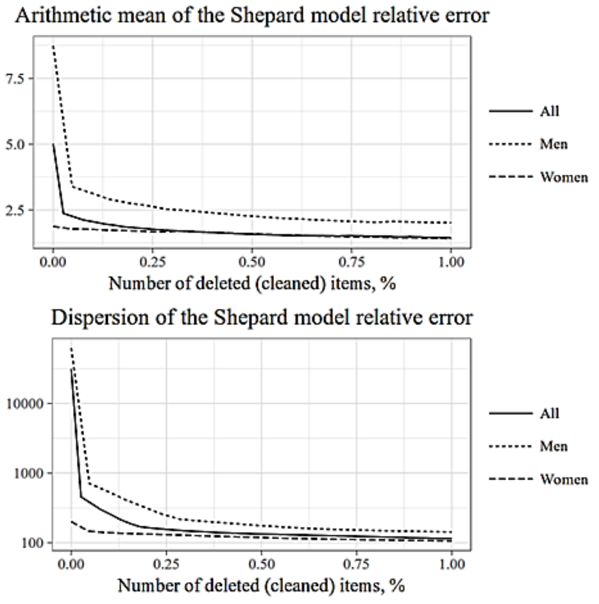

Fig. 2. Dynamics of decrease in the values of systematic and random errors after "cleaning" of the initial sample.

5. Conclusion

The calculation method in laboratory studies should be used in compliance with all the conditions and restrictions described by its developers. Calculation methods can be used during the screening stage, while the examination of patients with the revealed pathology should give preference to the direct measurement of the studied parameters. In the conditions of changing the significant elements of the original methods of application of calculation methods for the determination of cholesterol fractions, it is only possible when you replace a method (formula) of calculation of this indicator. In most cases the linear regression method is sufficient to refine the formula for LDL cholesterol calculations.

The use of mathematical models based on the Shepard's method makes it possible to obtain minimal errors. In this work the computational methods for calculating of the amount of LDL cholesterol are considered not to inferior practically in their characteristics to analytical methods. The models of estimation of the amount of this cholesterol fraction which were tested by us have allowed to reduce the errors of the calculation method significantly and can be used as a basis for practically used methods.

Acknowledgements

This work was partially supported by the research grant 18-07-00587 of Russian Foundation for Basic Research.

References

1. American Association of Clinical Endocrinologists and American college of endocrinology guidelines for management of dyslipidemia and prevention of cardiovascular disease, *J. Endocrine Practice* 23 (Suppl. 2) (2017) 87 p.
2. D. Shin, C. Bohra and K. Kongpakpaisarn, Novel method versus the Friedewald method for estimating low-density lipoprotein cholesterol in determination of the eligibility for statin treatment for primary prevention in the United States, *J. Medicine (Baltimore)* **97**(17) (2018) e0612.
3. W.T. Friedewald, R.I. Levy and D.S. Fredrickson, Estimation of the Concentration of Low-Density Lipoprotein Colysterol in Plasma, Without Use of the Preparative Ultracentrifuge, *J. Clinical Chemistry,* **18** (2) (1972) 499-502.
4. S. S. Martin, M. J. Blaha, et al., Friedewald estimated versus directly measured low density lipoprotein cholesterol and treatment implications, *J. Am. Coll. Cardiol*., **20** (2013). 732 –739.
5. D.S. Fredrickson, R.I. Levy and R.S. Lees, Fat transport in lipoproteins – an integrated approach to mechanisms and disorders, *The New England journal of medicine*, **276** (1) (1967) 34-42.
6. D.S. Fredrickson and R.I. Levy, Familial hyperlipoproteinemia, in *The Metabolic Basis of Inherited Disease*, 3rd edn. (McGraw-Hill, New York, 1972) 545-614.
7. D. Shepard, A two-dimensional interpolation function for irregularly-spaced data, in *Proc. of the 23 ACM National Conference* (ACM Press, New York, 1968), pp. 517-524.
8. R. Caira and F. Dell'Accio, Shepard-Bernoulli operators, *J. Mathematics of computation*, **257**(2) (2007) 299–321.
9. A. S. Anikin, CUDA-accelerated implementation of the multi-dimension functions approximation method, in *Proceedings of "Lyapunov readings" conference*, (ISDCT SB RAS, Irkutsk, 2009), pp. 3-4.
10. V. V. Kuz'menko, A. S. Anikin and A. Yu. Gornov, Algorithm for estimating the cholesterol concentration lipoproteins using the technique shepard static model, in *Proc. of the XV Baikal All-Russian Conference «Information and Mathematical Technologies in Science and Management»*, Tom 1st, (ISEM SB RAS, Irkutsk, 2010), pp. 145-151.