Clin. Lab. 2020;66:XXX-XXX
©Copyright

ORIGINAL ARTICLE

Optimization of the Preanalytical Phase by Estimating Serum Indices Using an Automated Classifier

O. Herráez, A. Velasco, A. Escobar, M. A. Asencio

Clinical Analysis Laboratory, La Mancha Centro General Hospital, Ciudad Real, Spain

SUMMARY

Background: **Most laboratory errors occur in the preanalytical phase. Among the most common preanalytical errors are interferences due to hemolysis, lipemia, and icterus. Our objective was to evaluate a serum interference estimation methodology by the RSD classifier, to identify other biochemical parameters affected by preanalytical interferences, and to determine the economic impact of its implementation.**

Methods: **The serum indices of 65,529 requests measured by the RSD system and by the analytical determination on the Cobas 711 or Cobas 8000 platforms were collected. We proceeded to the search for association patterns between the serum indices and laboratory analytical tests using data mining techniques. The influence of the preanalytical interferences was evaluated in 91 laboratory tests that include biochemistry, immunoassay, and coagulation. The savings estimation after the implementation of this methodology was made by time series models.**

Results: **The evaluation of the generated model showed compatibilities between the methods used (94.4% accuracy). The implementation of a protocol for serum indices estimation by the RSD would avoid the unnecessary analysis of the tests which are affected by interferences, achieving an estimated annual savings of €10,561. In addition, it allowed the estimation of bilirubin values which would add an annual savings of €4,900 in our laboratory. On the other hand, data mining techniques have allowed us to identify the following laboratory tests affected by hemolysis which are not usually considered in laboratories: iron, transferrin, fibrinogen, and alkaline phosphatase.**

Conclusions: **The RSD classifier is an efficient estimation method of serum indices and it allows the estimation of bilirubin values. The implementation of this methodology in our laboratory could lead to an estimated annual savings of more than €15,000 without increasing response times. Iron, alkaline phosphatase, transferrin, and fibrinogen should be included in the evaluated procedure.**

(Clin. Lab. 2020;66:xx-xx. DOI: 10.7754/Clin.Lab.2019.190541)

Correspondence:
María Ángeles Asencio
Clinical Analysis Laboratory
La Mancha Centro General Hospital
Avenida de la Constitución 3
13600 Alcázar de San Juan, Ciudad Real
Spain
Phone: +34 926 580574
Fax: +34 926 546882
Email: marian_asencio@yahoo.es

Manuscript accepted July 2, 2019

KEY WORDS

hemolysis, lipemia, icterus, preanalytical interferences, serum indices

INTRODUCTION

Most laboratory errors occur in the preanalytical phase [1]. The most common errors, especially those as a result of hemolysis, are icterus and lipemia interferences [2,3]. In past years, medical laboratories have introduced analytical platforms that allow the semiquantitative evaluation of the degree of these interferences by measuring the so-called serum indices [4].

In our laboratory, we have introduced in the RSD, a classification and aliquoted preanalytical system of sample tubes through the cap color comparison after photographing each tube, the QSI module (Quality System). This allows the sample quality evaluation by comparing the supernatant color, taking advantage of the photograph taken and comparing it with a set of stored patterns. The equipment, after communication with the laboratory information system, transmits a result that codifies the estimation of the preanalytical quality of the sample, so that the analysis is prevented of the tests affected by the interference.

Our objective was to validate the RSD preanalytical system for its application in the estimation of serum indices and in the management of interfered tests, as well as to evaluate the economic impact of its implementation. We also intended to identify other biochemical parameters affected by the preanalytical interferences.

MATERIALS AND METHODS

Data collection

In the laboratory, data is stored in a multidimensional model, allowing data mining using related queries in SQL. The data is adjusted with regard to its suitability for implementing the data mining techniques. For that, we perform projection, selection, and grouping queries in SQL.

RSD comparison with other analyzers

The serum indices of 65,529 samples measured by the RSD system and by the analytical determination in the Cobas 711 or Cobas 8000 platforms were collected (Roche Diagnostics, SL). Since both analyzers use the same measurement technique and reagent presentation, it is not necessary to segregate the serum indices measurements by the type of analyzer used. The comparison of both methods is carried out in the statistical platforms PASW Statistics 18 (IBM SPSS Statistics, Armonk, NY, USA) and WEKA (Waikato Environment for Knowledge Analysis, Australia), a software platform for data mining.

We selected the J48 algorithm, the RBF Network, and the Multilayer Perceptron for being fast algorithms with a precision higher than 88%; it is a result from classifying all the applications belonging to the most frequent type (ZeroR algorithm). The algorithm with the highest kappa (or concordance) index is based in decision trees (C4.5), defined with a 0.001 confidence factor, and uses Laplace smoothing. Moreover, a conventional statistical analysis is performed through mean comparison (Student's *t*-test).

Analysis of the influence of preanalytical interferences

For the analysis of the influence of preanalytical interferences in laboratory tests, we also use several data mining techniques in WEKA. First, algorithms which are based on association rules in order to identify the non-explicit relations among attributes (variables) are used. An association between two attributes happens when the frequency of occurrence in two specified values of each one together is relatively high.

In order to choose the most appropriate rules of all the possible rules that can be derived from a data set, restrictions can be used to apply support and confidence thresholds (using the *lift* metric). Broadly speaking, support is defined as the prevalence of all the items involved in the generated rule. Confidence is defined as the probability of finding the right part of a rule whitch is conditioned on finding also the left part. The *lift* indicator shows the observed support proportion of an item set regarding the theorical support of this set given the independence assumption. A *lift* value = 1 indicates that this set appears several times according to what is expected under independence conditions. A *lift* value > 1 indicates that this set appears a higher number of times than expected under independence conditions (so it can be inferred that there is a relationship that makes the products appear in the set more times than normal). A *lift* value < 1 indicates that this set appears several times less than expected under independence conditions.

The search of these association rules is performed on the discretized mineable table according to the definition of pathological change of laboratory results {Low, Normal, High}. Subsequently, variable selection algorithms are applied using the discretized value of serum index analysis as class variable, under the assumption that the selected diagnostic tests for the prediction of the serum index value will be correlated with this through reverse causality rules.

The influence of the preanalytical interferences was evaluated in 91 laboratory tests that include biochemistry, immunoassay, and coagulation. Depending on the model, the estimated index by the RSD (eliminating the lipemia and icterus results) or the discretized determination of the hemolysis index were used as the classification variable. In the two tables used for the model construction, the variables belonging to the laboratory tests with interference known as lipemia are removed, reducing the attribute number to 30. The data cleansing for the icterus evaluation is similar to the previous indices, except for the discretization of the index measurement, which is not considered necessary.

Costs evaluation by model implementation

The efficiency of the new detection methodology implementation was evaluated by the aid of time series models in the PASW Statistics 17 statistical platform (IBM SPSS Statistics, Armonk, NY, USA).

The cost estimate of the laboratory tests was performed through the automated costing system introduced in the laboratory which allocates to each determination, the direct reagents cost, the shared reagents cost, consumables, and the expenditure derived from the control analyses, gauges, and repetitions in the measuring period.

RESULTS

Serum indices comparison estimated by the RSD and analytical serum indices

After the application of the selected algorithms, we obtain the confusion matrix that is shown in Table 1. Using the C4.5 algorithm, 94.5% of correctly classified applications is achieved (kappa = 0.8).

According to the selected model, the samples identified by the RSD as hematic will have an analytical hematic index greater than 12, the samples identified as icteric by the RSD will have an analytical icteric index greater than 1, and those identified as lipemic will have an analytical lipemic index greater than 39. However, the samples with hemolysis (index > 27) and lipemia (index > 48) are only identified as lipemic samples by the RSD. In the three assessed cases through mean comparison ({no interferences, hemolysis), {no interferences, icterus}, {no interferences, lipemia}), some significant differences with a value $p < 0.001$ are observed.

Analysis of the influence of the preanalytical interferences

The variables of the model that are relevant according to the classification of the hemolysis degree are showed in Table 2. The following set of rules for samples in which the hematic index was positive are obtained:

Na=A I_HEM=A 109 ==> CL=A 50; lift: (81.87)
FBG=A CRP=A I_HEM=A 222 ==> CREA=A PT=B 48; lift: (61.33)
ALB=B I_HEM=A 198 ==> ALP=A 56; lift: (34.47)
CRP=A I_HEM=A 1191 ==> ALB=B 131; lift: (24.38)

Tests selected by the generation of models based on algorithms for the prediction of lipemia are shown in Table 3. The application of the M5P algorithm, based on the construction of a decision tree with linear regression function in the terminal nodes, allows the identification of the relevance of lipemia in triglycerides, total cholesterol, and LDL, discarding the potassium influence. The correlation coefficient obtained was 0.81 and the model error rate was 1.9%. The LDL determination is excluded from following the modelling because it is a calculated parameter which depends on the triglycerides value, among others.

The construction of a model based on the M5P algorithm, including cholesterol and triglycerides, allows the construction of a regression tree with a correlation coefficient of 0.80, where the LM regression lines are the following type:

LM = Coefficient x Triglycerides + value

In the regression tree, there is a theoretical possibility of predicting the triglycerides value from the measurement of the lipemic index. The first node indicates that if the triglycerides discretized value is less than 1.55, the value of the analytic lipemic index is between 0 and 50. In the tested models none are found with acceptable regression coefficients.

For the evaluation of icterus affected tests, the selection of relevant attributes is shown in Table 4. According to the tested models, the laboratory tests with relevant icterus interference were total bilirubin and serum iron. In this case, the algorithm with the most appropriate results is based on the logistic regression analysis of the relevant parameters with the index measurement. According to this algorithm, the value of the analytic icterus index can be calculated from the discretized value of the total bilirubin analytical result, with a correlation coefficient greater than 0.93. By implementing the same algorithm to the original results of the icteric index and total bilirubin (without discretizing), the following result was obtained:

$\text{TBil} = 0.807 \text{ x } I_{\text{icteric}} - 0.223$

This equation presents a correlation coefficient of 0.96. For TBil values < 1 mg/dL, the icteric index is lower than 1.47, which allows the reporting as TBil < 1 mg/dL of the total bilirubin in samples with a result of the icteric index less than 2.

After checking the total bilirubin values in 56,094 requests in which the RSD did not detect any preanalytical interference, it was observed that in 147 requests TBil > 1 mg/dL, of which in 75 TBil ≤ 1.5 mg/dL. We found that 40 of the 45 patients with TBil > 1.5 were newborns.

Costs evaluation by model implementation

Savings were estimated in the next 12 months after the implementation of the detection of serum interference by the RSD in October 2010. This estimation is performed by calculating the determination of cost reported as affected by hemolysis, icterus or lipemia, regardless of whether they have actually been analyzed or not. In order not to include in the model the possible variations derived from changes in the costs of the determinations, the current cost has been applied to all determinations. For the seasonally adjusted series, we selected the regression model with the highest correlation coefficient (0.687), which is the quadratic regression model. Based on this model, a predicted savings of €10,561 (Figure 1) was estimated for the next 12 months. In addition, the application of this screening method for bilirubin analysis in our laboratory represents an annual saving of €4,900.

DISCUSSION

In this study we evaluated the application of the RSD classifier for the detection of serum interferences using data mining techniques, so that the biochemical analysis of the diagnostic tests affected by interference could be avoided. By analyzing time series, it was estimated that the savings during the 12 months following the implementation of this system could save €10,561. The search for association patterns between serum indices and analytical tests showed the linear relationship between the value of the icteric index and total bilirubin, allowing the estimation of this by measuring the index, with an estimated annual saving of €4,900.

Table 1. Results of C4.5 model for prediction of interference estimation by RSD.

Serum index by RSD	Serum index predicted by the model from analytical measurements			
	0 No interferences	1 Hemolysis	2 Lipemia	3 Icterus
0 (No interferences)	57,272	702	29	188
1 (Hemolysis)	834	3,756	6	152
2 (Lipemia)	232	99	129	0
3 (Icterus)	1,150	224	1	724

Table 2. Tests selected by generation of models based on algorithms for the prediction of hemolysis.

Algorithm	Model	Variable Class	Relevant laboratory tests
CfSubset	Variables selection	RSD Index	Total bilirubin, albumin alkaline phosphatase, iron, chlorine, fibrinogen, insulin, and transferrin
Linear regression	Regression rules	Index analytical hemolysis	Total bilirubin, albumin, alkaline phosphatase, iron and transferrin

Table 3. Tests selected by generation of models based on algorithms for the prediction of lipemia.

Algorithm	Searching method	Selected tests
CfsSubsetEval	BestFirst -D 1 -N 5	Triglycerides, potassium
Infogain	Greedy Stepwise	Triglycerides, cholesterol, LDL
Wrapper MultilayerPerceptron	Greedy Stepwise	Triglycerides, cholesterol, LDL

Table 4. Tests selected by generation of models based on algorithms for the prediction of icterus.

Algorithm	Searching method	Selected tests
CfsSubsetEval	BestFirst -D 1 -N 5	Chlorine, iron, total bilirubin
Wrapper M5P	Greedy Stepwise	Iron, total bilirubin
Wrapper MultilayerPerceptron	Greedy Stepwise	Iron, total bilirubin

The construction of models by means of data mining techniques allows the detection of thresholds for serum interference information indicated by the calibration of the equipment. The threshold from which the RSD detects hemolysis is, depending on the model, the equivalent to an analytical serum index of 12. In the references, we found that the minimum interference threshold for hemolysis in other analyzers with similar characteristics to ours is 40 [5]. The use of this threshold would imply a diagnostic sensitivity of the RSD of 99.92% in the detection of hemolysis and is appropriate to be applied as a screening technique.

The sensitivity of the RSD for the detection of icterus and lipemia are also adequate, because its detection threshold is equivalent to an icteric and lipemic analytical index of 1 and 39, respectively. According to bibliographical sources, the level of interference on diagnostic tests begins with a value of the icteric index of 7 and a lipemic index higher than 50 [6].

The guide of the *Clinical and Laboratory Standards Institute, Interference Testing in Clinical Chemistry* (CLSI EP07-A2) recommends the evaluation of the in-

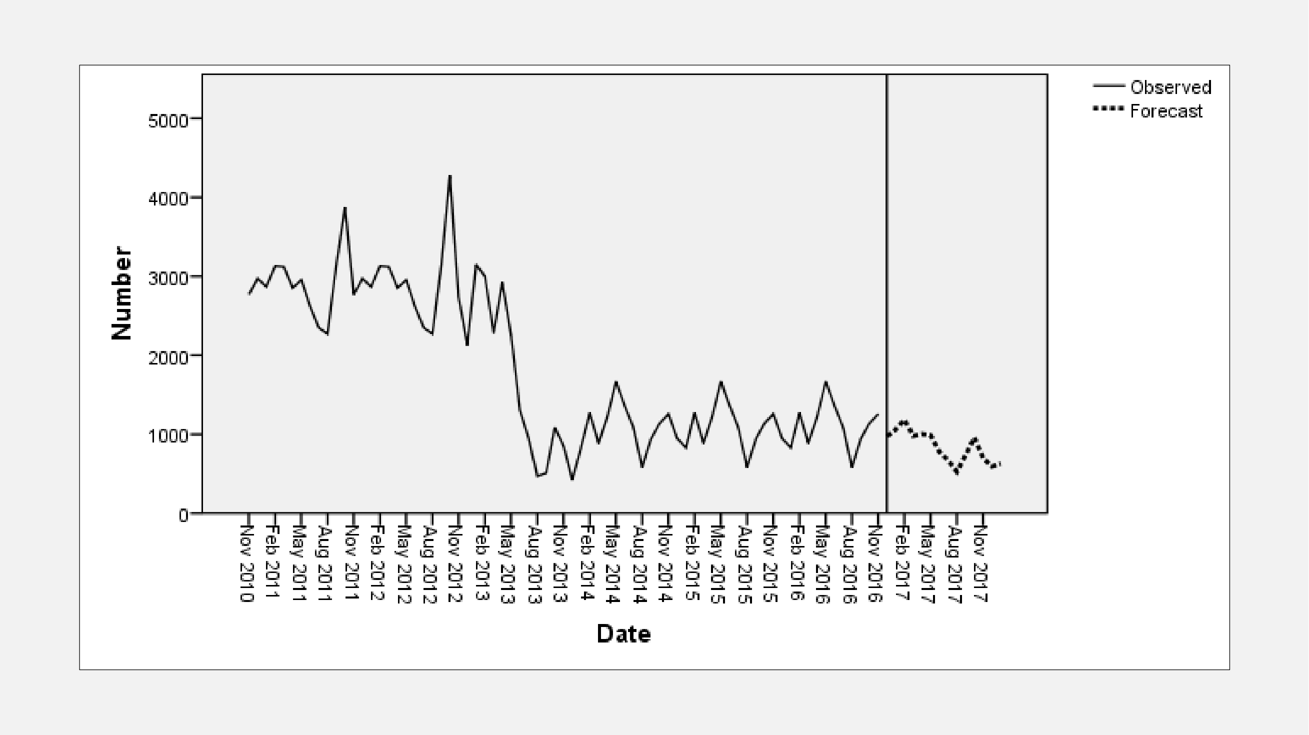

Figure 1. Evolution of the monthly cost of tests not informed by preanalytical interference.

terference's effects on the analytical determinations in each laboratory [7]. For the evaluation of interference by hemolysis, it recommends the creation of hemolysate by freeze-thawing, although some authors follow other procedures that try to imitate the phenomenon of hemolysis *in vivo* [8]. Our data suggest that hemolysis interferes with the analytical results of iron, transferrin, total bilirubin, and alkaline phosphatase. The interference with alkaline phosphatase values has already been detected by several authors using different techniques [8-11]. In our study, the concentration of total bilirubin, transferrin, and alkaline phosphatase are shown as variables with strong relevance for the prediction of hemolysis, both in the models that use as a variable to predict the analytical serum index and those that use index estimation by the RSD. The advantage of our method is that it does not require artificial modifications of the sample, and the evaluation is directly done from the laboratory data delivered to the patients. However, due to the processing to which the data must be submitted, it is difficult to define the interference degree. Other more powerful algorithms with different prior processing of data could expose this relationship and establish cutoff points according to the definition of the interference evaluation guides used by other authors.

Our data relate lipemia with variations in the analytical results of triglycerides, LDL, and cholesterol. Because of the LDL levels are calculated using the Friedewald formula in our laboratory, this parameter cannot be considered relevant *per se*, but indirectly through the other two parameters. In this case the model locates the cause of the lipemia, since according to several studies triglycerides and cholesterol measurement is not altered by the presence of turbidity [6,13-14].

In our study, the number of samples with a turbidity level analytically determined higher than 150, are less than 200 in each of the tables, so that the algorithms used to search relationships between the values of the laboratory tests and the turbidity value do not show tests such as IgA that have been related to it [15]. However, these algorithms have detected tests not influenced by turbidity determinations such as serum ions (sodium, potassium and chlorine), alkaline phosphatase, serum iron, total proteins, urea, and phosphorus, which appear in the references as tests with variations not significantly related to lipemia [6].

In our model, total bilirubin and serum iron are identified as laboratory tests related to icterus. Serum iron has been identified by other authors as a determination interfered by icterus [16]. The regression analysis between the value of bilirubin and the analytical value of the icteric index has allowed us to define a procedure for the estimation of bilirubin from it. Since linearity studies have not been carried out, we only propose the use this equation for results with an icteric index lower than 1, since then the value of the total bilirubin in all cases is below the upper limit of reference. The application of this screening method in our laboratory represents an annual saving of €4,900, although it slightly increases the cost for bilirubin determination and, therefore, increases the indirect cost related to the controls and calibrators analysis.

On the other hand, the implementation of the previous detection of the serum indices through the RSD classifier would allow the generation of an automatic comment in the laboratory tests possibly affected by the estimation of the preanalytical interferences, so that they would not be carried out by the laboratory analyzers. This work system would involve an economic saving in diagnostic tests estimated to be €10,561 for the next 12 months, without implying an increase in the response time or the need to reprocess the samples.

We consider significant the fact that the coagulation tests have not been found as affected by hemolysis, unlike other authors [17-19]. This, therefore, introduces a bias in the data for the three evaluated coagulation tests, which are not analyzed by the analyzer if the sample is hemolyzed. Another limitation of this study is due to the low prevalence of results measured by the RSD with lipemia detection, so that the algorithms used for modelling the influence of lipemia estimated with laboratory tests have not allowed the detection of relevance patterns. Other tests, such as CK and LDH, with interference by lipemia [13] could not be detected because the interference threshold is based on a value of the lipemic index of 1,200, a value that is not reached with the data collected for the study. Similarly, there were not many cases that would allow the algorithms used to find associations between some laboratory tests, such as CK, total proteins, and GGT, and the value of the icteric index, which have been detected by other methods [13]. In our study, the possibility of establishing a triglyceride value from the measurement of the analytical lipemic index is tested by regression analysis. However, none of the models analyzed was statistically acceptable to be able to report a triglyceride value from the values of the lipemic index. We consider that more specific studies could reveal relationships between diagnostic tests and serum indices that have not been analyzed in the present because they have a low determination number.

CONCLUSION

By using data mining techniques, we have defined an adequate method of estimating serum indices using the RSD classifier, avoiding the analysis of interfered laboratory tests by preanalytical errors without increasing response times, which makes it possible to issue quality analytical information. This methodology would involve a reduction in laboratory costs of approximately €15,000, both for the savings in analytical determinations and for allowing to directly estimate the value of the total bilirubin. Data mining techniques have helped our laboratory to identify laboratory tests such as iron, alkaline phosphatase, transferrin, and fibrinogen as tests affected by preanalytical interferences that must be included in the evaluated procedure. The implementation of measures that allow the optimization of the preanalytical phase would increase the overall efficiency of the laboratory work.

Financial Support:
This study was funded by J. L. Castaño Foundation, Spain.

Declaration of Interest:
The authors declare that they have no conflict of interest.

References:

1. Narayanan S. The preanalytic phase. An important component of laboratory medicine. Am J Clin Pathol 2000;113:429-52 (PMID: 10705825).
2. Lippi G, Blanckaert N, Bonini P, et al. Haemolysis: an overview of the leading cause of unsuitable specimens in clinical laboratories. Clin Chem Lab Med 2008;46:764-72 (PMID: 18601596).
3. Soderberg J, Jonsson PA, Wallin O, Grankvist K, Hultdin J. Haemolysis index-an estimate of preanalytical quality in primary health care. Clin Chem Lab Med 2009;47:940-4 (PMID: 195891 05).
4. Lippi G, Luca Salvagno G, Blanckaert N, et al. Multicenter evaluation of the hemolysis index in automated clinical chemistry systems. Clin Chem Lab Med 2009;47:934-9 (PMID: 19548845).
5. Green SF. The cost of poor blood specimen quality and errors in preanalytical processes. Clin Biochem 2013;46:1175-9 (PMID: 23769816).
6. Ji JZ, Meng QH. Evaluation of the interference of hemoglobin, bilirubin, and lipids on Roche Cobas 6000 assays. Clin Chim Acta 2011;412:1550-3 (PMID: 21575617).
7. Clinical and Laboratory Standards Institute (CLSI) (2005) Interference Testing in Clinical Chemistry; Approved Guideline-Second edition, CLSI document EP7-A2. Clinical and Laboratory Standards Institute, Wayne, Pennsylvania, USA.
8. Lippi G, Salvagno GL, Montagnana M, Brocco G, Guidi GC. Influence of hemolysis on routine clinical chemistry testing. Clin Chem Lab Med 2006;44:311-6 (PMID: 16519604).
9. Steen G, Vermeer HJ, Naus AJ, Goevaerts B, Agricola PT, Schoenmakers CH. Multicenter evaluation of the interference of hemoglobin, bilirubin and lipids on Synchron LX-20 assays. Clin Chem Lab Med 2006;44: 413-9 (PMID: 16599834).
10. Wang Z, Guo H, Wang Y, Kong F, Wang R. Interfering effect of bilirubin on the determination of alkaline phosphatase. Int J Clin Exp Med 2014;7:4244-8 (PMID: 25550938).
11. Cecco S, Rehak N. Rejection rules for hemolysis icterus and lipemia indices on Synchron LX20 Clinical System. Clin Chem 2004;50A:109.
12. Vermeer HJ, Steen G, Naus AJ, Goevaerts B, Agricola PT, Schoenmakers CH. Correction of patient results for Beckman Coulter LX-20 assays affected by interference due to hemoglobin, bilirubin or lipids: a practical approach. Clin Chem Lab Med 2007;45:114-9 (PMID: 17243928).

13. Calmarza P, Cordero J. Lipemia interferences in routine clinical biochemical tests. Biochem Med (Zagreb) 2011;21:160-6 (PMID: 22135856).

14. Rosenthal MA, Katz HB. An innovative method for determining lipemia interference in blood specimens. Clin Chim Acta 2011; 412:665-7 (PMID: 21108941).

15. Agarwal S, Vargas G, Nordstrom C, Tam E, Buffone GJ, Devaraj S. Effect of interference from hemolysis, icterus and lipemia on routine pediatric clinical chemistry assays. Clin Chim Acta 2015; 438:241-5 (PMID: 25128720).

16. Szoke D, Braga F, Valente C, Panteghini M. Hemoglobin, bilirubin, and lipid interference on Roche Cobas 6000 assays. Clin Chim Acta 2012;413:339-41, author reply 42-43 (PMID: 220015 18).

17. Lippi G, Montagnana M, Salvagno GL, Guidi GC. Interference of blood cell lysis on routine coagulation testing. Arch Pathol Lab Med 2006;130:181-4 (PMID: 16454558).

18. Laga AC, Cheves TA, Sweeney JC. The effect of specimen hemolysis on coagulation test results. Am J Clin Pathol 2006;126: 748-55 (PMID: 17050072).

19. Ahmad AL, Ismail S, Bhatia S. Optimization of coagulation-flocculation process for palm oil mill effluent using response surface methodology. Environ Sci Technol 2005;39:2828-34 (PMID: 158 84382).