Assessment and Modeling of Routinely Used Biochemical Laboratory Data of Healthy Individuals and End-Stage Renal Failure (ESRF) Patients by Three Different Chemometric Methods

Agelos Papaioannou,^{1*} George Rigas,² Panagiotis Plageras,¹ George A. Karikas,³ and George Karamanis⁴

¹Department of Medical Laboratories, Section of Clinical Chemistry—Biochemistry, Education & Technological Institute of Larissa, Greece

²Department of Animal Production, Education & Technological Institute of Larissa, Greece

³Department of Medical Laboratories, Education & Technological Institute of Athens, Greece

⁴Department of Biochemistry, Diagnostic Laboratories, General Hospital of Kavala, Greece

Background: In recent years, the use of biochemical markers has received increasing attention for purposes of risk assessment and clinical management in renal failure patients. Chemometric methods are often used in medical studies and there are already indications for their specific role as a tool of the medical statistics. **Methods:** Three chemometric methods, discriminant analysis (DA), binary logistic regression analysis (BLRA), and cluster analysis (CA), were used for assessment and modeling of routinely used biochemical laboratory data of 18 parameters that were determined from 185 healthy individuals (HIs) and 173 end-stage renal failure (ESRF) patients. **Results:** The above-mentioned chemometric methods were performed using the data

set of 14 parameters since the rest 4 parameters did not present significant difference between healthy and patients. DA created a model using only ALB (Albumin), K (Potassium), TG (Triglyceride), and ALP (Alkaline phosphatase); BLRA model also used the above four parameters; CA classified all the cases into two clusters using the same four parameters and one more parameter, AST (aspartate aminotransferase). **Conclusions:** This study provides models for assessment and modeling of routinely used biochemical laboratory data, finding groups of similarity among clinical tests usually determined on HIs and ESRF patients, contributing in data mining and reducing costs. J. Clin. Lab. Anal. 27:211–219, 2013. © 2013 Wiley Periodicals, Inc.

Key words: discriminant analysis; binary logistic regression analysis; cluster analysis

INTRODUCTION

Patients with chronic kidney disease represent a major healthcare problem. In spite of modern treatment, these patients are at high risk for subsequent other disease relapses. This population is however very heterogeneous and clinicians are therefore faced with the task of risk stratification to optimize individual patient care and avoid risk-filled therapies and procedures (1–8).

Improvements in the medical management of kidney disease have delayed the rate of progression to kidney failure (9). In the past 40 years, the most commonly used marker of overall renal function in clinical practice has been plasma creatinine (CREA) concentration. More pre-

cise renal function estimation can be obtained by using estimates of the glomerular filtration rate (GFR; (10–15). Mathematical models have been used since 1976 in an attempt to predict the progression of chronic renal failure. These models have used the serum CREA level as either a reciprocal or logarithmic plot against time. These studies

*Correspondence to: Agelos Papaioannou, Department of Medical Laboratories, Section of Clinical Chemistry – Biochemistry, TEI of Larissa, 41110 Larissa, Greece. E-mail: papaioannou@teilar.gr

Received 4 November 2012; Accepted 11 January 2013

DOI 10.1002/jcla.21586

Published online in Wiley Online Library (wileyonlinelibrary.com).

indicate that predictive models using serum CREA levels are of limited clinical use (16–22).

In recent years, the use of biochemical markers has received increasing attention for purposes of risk assessment and clinical management in renal failure patients. Assessment of these biomarkers in the diseased state however, requires knowledge of this marker in the healthy state taking into account variables that may affect levels such as age and gender. Hence, a matched control population is necessary when evaluating models with these markers (23).

Two major classes of multivariate methods can be distinguished (24, 25). The first comprises the methods concerning with the recognition of patterns. Well-known methods of this type are the DA and BLRA that have been applied to clinical chemical data on several occasions (26–31). The second class of pattern-recognition methods aims at the detection of patterns that remain unnoticed in clinical medicine. A single pattern may become apparent when distinct patterns in a group differ from other patterns of the same category—they form a “cluster”; CA was used for multivariate statistical modeling of the input data (32–37).

This study presents an example of routinely biochemical laboratory data intelligent analysis, which used linear discriminant analysis (DA) and binary logistic regression analysis (BLRA) on the one hand and hierarchical cluster analysis (CA) on the other, for assessment and modeling of the input data. With the help of chemometric methods of analysis, we can discern which biochemical parameters are important and which biochemical parameters are interrelated. This information cannot be acquired if non-multivariable methods of analysis are used. The target of the present study was to offer a new research strategy to classify, interpret, and modeling routinely biochemical results in order to reach better decision-making solutions concerning human health and disease prevention, and contribute to reducing costs.

METHODS AND MATERIALS

Experimental

We studied the distribution patterns of some analytes commonly assayed in clinical chemistry—biochemistry laboratories in healthy individuals (HIs) and end-stage renal failure (ESRF) patients. A total of 185 HIs and 75 ESRF patients from General Hospital of Drama (GHD) and 98 ESRF patients from General Hospital of Kavala (GHK) were among those tested. All ESRF patients were undergoing hemodialysis in the above hospitals (duration of hemodialysis: mean $\pm$ SD = 65.1 $\pm$ 55.7; median = 50.0 months). The selection of a normal subject is based on a historical review of past medical problems together with

TABLE 1. Methods Used for the Determination of the Different Quantities (37°C)

Quantity	Olympus method	Dimension RXL method
Alanine aminotransferase (ALT)	IFCC w/o p-5'-p	IFCC with (P-5-P)
Albumin (ALB)	Bromocresol green	(BCP) purple
Aspartate aminotransferase (AST)	IFCC w/o p-5'-p	IFCC with (P-5-P)
Alkaline phosphatase (ALP)	IFCC/AMP Buffer	AMP buffer
Calcium (Ca)	o-Cresolphthalein complex	o-Cresolphthalein complex
Cholesterol (CHOL)	CHOD/PAP	CHOD/PAP or CHOD/POD
Chloride (Cl)	ISE indirect	IMT Indirect
Creatinine (CREA)	Jaffe'	Jaffe'
Glucose (GLU)	Hexokinase	Hexokinase (HK/G-6-PDH)
High-density lipoprotein-cholesterol (HDL-C)	Direct enzymatic	Direct enzymatic
Iron (Fe)	TPTZ	Ferene
Phosphorus (P)	Phosphomolybdate	Phosphomolybdate U.V.
Potassium (K)	ISE indirect	IMT Indirect
Sodium (Na)	ISE indirect	IMT Indirect
Total proteins (TPs)	Biuret	Biuret
Triglycerides (TGs)	GRO/PAP	(CHOD/PAP or CHOD/POD)
Uric acid (UA)	Uricase/PAP	Uricase/PAP or Uricase/POD
Urea (UREA)	Urease U.V.	Urease/GLDH U.V.

a physical examination that is used prior to the donation of blood. Exclusion criteria were systemic diseases and pathophysiological disorders such as renal failure, congestive heart diseases, chronic respiratory diseases, liver diseases, malabsorption syndromes, and others (38).

The data used in this study were derived from the findings of the blood samples that were taken in the biochemical laboratories of GHD and GHK, monthly, for a period of 1 year (January to December). The preanalytical conditions and the analysis of the blood samples are described in details elsewhere (38–40). The analyses of 18 biochemical parameters were performed with an Olympus AU640 (Olympus, Tokyo, Japan) and a Dimension RXL analyzers (Dade-Behring, Deerfield, IL) at 37°C according to the methods listed in Table 1, immediately after centrifugation.

Before each determination, calibration and internal control of analyzers with calibrators and quality controls of the corresponding manufactures preceded, according

to manufacturer's instructions and international literature (41). The reagents provided in the commercial kits were used in the analyzers, and the methods were adapted according to the manufacturer's instructions. The water was free from metal ions and had a maximum receptivity of 18.2 M Ω cm at 25°C. Accuracy was checked (and achieved) by two external quality control programs: Radox (RIQAS) and Hospital of Evangelismos, Greece (ESEAP).

Statistical Methods

Descriptive statistics

Basic statistics and correlation calculations were carried out in order to give initial information about the biochemical data. Unless otherwise indicated, the characteristics of the subjects were described as mean value and standard deviation.

Tests for significance of observed mean differences were performed using the Student's *t*-test, and tests for significance of observed variance differences were performed using Levene's test (Levene's test is an inferential statistic used to assess the equality of variances in different samples for a given population). To evaluate the correlations between the levels of biomarkers of each studied group, the Pearson correlation coefficients were calculated. All these data analyses were performed with the Statistical Package for Social Sciences (SPSS 15.0 and Statistica 7.0) software for Windows.

Chemometric methods

In the present study, the classification of the individuals as patient or healthy according to their biochemical parameters was performed using a combination of DA, BLRA, and CA. DA is useful for building a predictive model of group membership based on observed characteristics of each case. The procedure generates a discriminant function (DF; or, for more than two groups, a set of DFs) based on linear combinations of the predictor variables that provide the best discrimination among the groups. The functions are generated from a sample of cases for which group membership is known; the functions can then be applied to new cases that have measurements for the predictor variables but have unknown group membership. The discrimination between two or more groups is obtained in terms of the means of discriminating variables of DFs. Moreover, an alternative way for building a predictive model is to use Fisher's linear DFs. In this case, DA constructs a DF for each group. We performed DA on the log-transformed data based on two different modes, that is, standard and stepwise, to construct the best DFs to confirm the groups determined by the patient and HIs.

The logistic regression model is a type of generalized linear model that extends the linear regression model by linking the range of real numbers to the 0–1 range. Start by considering the existence of an unobserved continuous variable, *Z*, which can be thought of as the “propensity toward” the event of interest. In the case of a doctor, *Z* represents an individual's propensity to be healthy, with larger values of *Z* corresponding to greater probabilities of being healthy. Logistic regression is applicable to a broader range of research situations than DA. BLRA is most useful when one wants to model the event probability for a categorical response variable with two outcomes. For example, a doctor wants to accurately diagnose possibly illness based on values of a set of biochemical parameters. A biochemist wants to be able to identify characteristics that are indicative of individuals who are likely to be patients and use those characteristics to identify patient and HIs. For example, using the BLRA, one could construct a model for predicting the probability of a given individual to be healthy or a patient. The BLRA is a useful tool for predicting the value of a categorical response variable with two possible outcomes and less assumption than DA. It must be noted that the collinearity of predictors could be a problem at the application of BLRA.

CA is another data reduction method that is used to classify entities with similar properties. The method divides a large number of objects into a smaller number of homogeneous groups on the basis of their correlation structure. The objective of CA is to identify the complex nature of multivariate relationships (by searching for natural groupings or types) among the data under investigation, so as to foster further hypothesis development about the phenomena being studied. CA imposes a characteristic structure on the data analysis for exploratory purposes. CA was conducted a group of biochemical data of HIs and ESRF patients, using Ward's method with Euclidean distance measure. We used CA to link variables in the configuration of a tree with different branches—branches that have linkages closer to each other indicate a stronger relationship among variables or cluster of variables. The dendrogram generated from tree clustering provides a useful graphical tool for determining the number of clusters that describe underlying processes that lead to spatial variation. We applied hierarchical CA on log-transformed standardized data using Ward's method with squared Euclidean distances.

RESULTS

Statistical Screening of Data

First, Levene's test for equality of variances and both pooled and separate variances *t*-tests for equality of means were conducted to see if there is a difference between the

TABLE 2. Descriptive Statistics for the Tested Clinical Parameters for Female, Male and All Individuals' Data Sets (Mean Value, Standard Deviation of the Mean)

Parameter	Cases	Female Mean $\pm$ SD	Male Mean $\pm$ SD	Total Mean $\pm$ SD
ALB (g/L)	Patient	3.6 $\pm$ 0.4	3.6 $\pm$ 0.4	3.6 $\pm$ 0.4
	Healthy	4.8 $\pm$ 0.4	5.1 $\pm$ 0.5	5.0 $\pm$ 0.4
ALP (U/L)	Patient	119.7 $\pm$ 86.5	96.6 $\pm$ 40.8	104.4 $\pm$ 60.7
	Healthy	54.4 $\pm$ 15.8	77.7 $\pm$ 25.7	63.7 $\pm$ 23.2
ALT (U/L)	Patient	36.3 $\pm$ 13.1	41.3 $\pm$ 27.1	39.6 $\pm$ 23.4
	Healthy	18.6 $\pm$ 14.6	27.2 $\pm$ 14.7	22.1 $\pm$ 15.2
AST (U/L)	Patient	14.7 $\pm$ 8.9	16.3 $\pm$ 13.3	15.8 $\pm$ 12.0
	Healthy	20.0 $\pm$ 6.3	24.8 $\pm$ 9.2	21.9 $\pm$ 7.9
Ca (mg/dL)	Patient	8.9 $\pm$ 1.1	9.0 $\pm$ 0.7	9.0 $\pm$ 0.8
	Healthy	9.6 $\pm$ 0.5	9.7 $\pm$ 0.5	9.6 $\pm$ 0.5
CREA (mg/dL)	Patient	8.8 $\pm$ 1.6	9.6 $\pm$ 2.4	9.3 $\pm$ 2.2
	Healthy	0.9 $\pm$ 0.1	1.0 $\pm$ 0.1	1.0 $\pm$ 0.1
HDL-C (mg/dL)	Patient	38.2 $\pm$ 12.1	31.3 $\pm$ 8.6	33.6 $\pm$ 10.4
	Healthy	51.4 $\pm$ 11.7	42.3 $\pm$ 10.0	47.4 $\pm$ 11.8
K (mmol/L)	Patient	5.6 $\pm$ 0.7	5.3 $\pm$ 0.9	5.4 $\pm$ 0.8
	Healthy	4.5 $\pm$ 0.4	4.3 $\pm$ 0.3	4.4 $\pm$ 0.4
Na (mmol/L)	Patient	137.5 $\pm$ 2.4	136.3 $\pm$ 3.4	136.7 $\pm$ 3.1
	Healthy	139.4 $\pm$ 2.3	141.1 $\pm$ 2.1	140.2 $\pm$ 2.4
P (mg/dL)	Patient	5.4 $\pm$ 1.9	5.5 $\pm$ 1.5	5.5 $\pm$ 1.6
	Healthy	3.9 $\pm$ 0.5	3.9 $\pm$ 0.7	3.9 $\pm$ 0.6
TG (mg/dL)	Patient	200.6 $\pm$ 142.6	159.1 $\pm$ 77.9	173.0 $\pm$ 105.4
	Healthy	75.9 $\pm$ 35.2	91.6 $\pm$ 44.5	82.2 $\pm$ 39.7
TP (g/L)	Patient	6.7 $\pm$ 0.4	6.8 $\pm$ 0.6	6.8 $\pm$ 0.5
	Healthy	7.6 $\pm$ 0.5	7.8 $\pm$ 0.5	7.6 $\pm$ 0.5
UA (mg/dL)	Patient	5.2 $\pm$ 1.0	5.8 $\pm$ 0.9	5.6 $\pm$ 1.0
	Healthy	3.8 $\pm$ 0.8	5.4 $\pm$ 1.0	4.5 $\pm$ 1.2
UREA (mg/dL)	Patient	164.9 $\pm$ 35.9	174.9 $\pm$ 36.0	171.6 $\pm$ 36.1
	Healthy	24.1 $\pm$ 5.0	28.4 $\pm$ 5.6	25.8 $\pm$ 5.6

studied groups' variances and mean values for each of the 18 biochemical parameters. For four of them (glucose, GLU; cholesterol, CHOL; iron, Fe; and chloride, Cl), there was no significant difference ($P < 0.01$) between the studied groups' mean values according to the two tests. Due to that, these four parameters were excluded from further chemometric evaluation. The application of DA, BLRA and CA was restricted to the rest of the 14 parameters (UREA; CREA; high-density lipoprotein-cholesterol, HDL-C; triglyceride, TG; uric acid, UA; potassium, K; sodium, Na; calcium, Ca; phosphorus, P; aspartate aminotransferase, AST; alanine aminotransferase, ALT; alkaline phosphatase, ALP; total protein, TP; albumin, ALB), which were significantly different ($P < 0.01$) between the two groups.

Descriptive statistics for the above 14 biochemical parameters for the total number of cases, the two distinct groups (HIs from the hospital of Drama and patients from the hospital of Kavala), and sex categories are presented in Table 2. The cross-correlation between the different biochemical test parameters of HI showed that although the overall significance of many of them was statistically sound, according to the Pearson test of the results for

r , a real logical interpretation ($r > 0.6$) for significance could be offered only for the couples of parameters such as AST/ALT (0.752, $P < 0.001$) and TP/ALB (0.790, $P < 0.001$). The correlated couples of parameters for the 98 ESRF patients from the GHK were also AST/ALT (0.718, $P < 0.001$) and TP/ALB (0.686, $P < 0.001$). These correlations were used to identify groups of highly correlated biochemical variables, and it is evident that the simple correlation analysis did not indicate specific links among the studied biochemical parameters.

Parameters Distribution Characteristics and Data Treatment

Most methods, such as DA and CA, require variables to conform to a normal distribution. Thus, the normality of the distribution of each variable was checked by Kolmogorov–Smirnov statistic, histograms, normality plots, and by analyzing kurtosis and skewness before multivariate statistical analyses.

The original data demonstrated that HDL-C and Na were more normally distributed, and UA, TP, ALB were less normally distributed; whereas the other parameters (except Ca) were positively skewed, with kurtosis coefficients significantly differed from zero and the most (except UREA and CREA) significantly greater than zero (95% confidence). After log-transformation of the parameters, all skewness and kurtosis values were significantly reduced (Fig. 1).

For CA, all parameters were also Z-scale standardized (mean = 0; variance = 1) to minimize the effects of differences in measurement units and variance and render the data dimensionless. Consequently, each column had zero mean and unit variance.

Analysis by Chemometric Methods

Three different statistical methods were applied for the analysis of clinical laboratory data. DA, BLRA, and CA interacted together harmoniously to model the 14-parameter data sets corresponding to (i) AI (HIs from Drama's hospital and patients from Kavala's hospital) group that consists of 185 HIs from GHD and 98 ESRF patients from GHK; (ii) HI group from Drama's hospital (DHI); (iii) KPI (patient individuals from Kavala's hospital) group that refers to the 98 ESRF patients from GHK; and (iv) DPI (patient individuals from Drama's hospital) group that contains the 75 ESRF patients from GHD, as follows.

Application of DA

DA was performed using the log-transformed data set of 14 parameters and their original classification into the

Fig. 1. Skewness and Kurtosis coefficients of original (●) and log-transformed (▼) data.

two major groups (healthy and patients) obtained from the clinical results of AI group. The groups were the dependent categorical variable, and the measured parameters were the independent variables.

First, DA was applied to AI group to create a DA model and the data set of DPI group was used to validate the model. Wilks' lambda and the Chi-square for the DFs obtained from the standard and stepwise modes of DA were 0.015, 0.015 and 1,157.001, 1,162.922, respectively, at $P < 0.001$, suggesting that the DA was credible and effective. The DFs or classification functions (CFs) obtained from the standard and stepwise modes of DA are shown in Table 3.

The discriminant model assigns the individual to the group whose CF obtained the highest score. The standard DA mode constructed DFs containing all 14 parameters, correctly assigning 100.00% of the AI and DPI. However, in the stepwise mode, the DA produced a classification matrix with 100.00% correct assignments using only six discriminant parameters. These results were similar to those obtained using the standard mode, but with significantly fewer parameters. Thus, the stepwise DA showed that CREA, ALB, UA, UREA, K, and ALP were significant discriminant parameters of individual's classification, with 100.00% of assignments correct. It was essential, therefore, to strengthen the determination accu-

TABLE 3. Classification Functions Coefficients for Discriminant Analysis

Parameters	Standard mode		Stepwise mode	
	Healthy	Patients	Healthy	Patients
UREA	95.811	156.398	146.427	201.174
CREA	-614.266	-396.422	-257.330	-41.552
HDL-D	335.695	348.939		
TG	168.415	171.597		
UA	-25.281	-91.530	69.663	5.976
K	905.552	827.535	292.368	211.220
Na	35,565.061	35,360.101		
Ca	2,262.777	2,272.784		
P	298.758	284.955		
AST	-286.555	-296.019		
ALT	153.761	164.490		
ALP	-197.036	-174.386	56.794	76.864
TP	4,010.099	4,009.387		
ALB	-3,134.015	-3,392.667	485.259	230.250
Constant	-40,575.949	-40,163.146	-442.634	-422.850

racy of CREA, ALB, UA, UREA, K, and ALP to clearly identify healthy and patient individuals in the future.

Since it is known that CREA and UREA levels are high among ESRF patients, from the original group of 14 parameters, CREA and UREA were excluded and performed DA (stepwise mode) to the group of the remaining 12 parameters and 8 of them were selected in the following order: ALB, K, TG, ALP, AST, ALT, P, and TP (Wilks' lambda and the Chi-square were 0.109 and 613.937, respectively, at $P < 0.001$). In this case, 100% of the HI, 98.98% of the KPI, and 78.67% of the DPI were correctly classified. Afterwards, we took the first six parameters (Wilks' lambda and the Chi-square were 0.124 and 580.399, respectively, at $P < 0.001$) and performed DA (standard mode), and the classification results were the same as above. Then, we left out ALT (as ALT correlated with AST, Pearson $r = 0.5$) and performed DA to the remaining five parameters with standard and stepwise mode (Wilks' lambda and the Chi-square were 0.148 and 532.336, respectively, at $P < 0.001$, the same for the two DA modes). The yielded model (based on the parameters ALB, K, TG, ALP, and AST) classified correctly 99.46% of the HI, 97.96% of the KPI, and 85.33% of the DPI. This last model gave the best results among produced DA models, while using only five biochemical parameters; it must be noticed that these parameters are not specific only for renal function, but they are used routinely in clinical laboratory practice.

Consequently, DA can be applied to new individuals who have measurements for the above predictor parameters but have unknown group membership. Each new individual is assigned to the group for which DF obtains the greater value.

Application of BLRA

The rationale for using linear logistic regression for analysis is that it is not required to assume the normal distribution for the independent variables. BLRA was applied to the original data set of the 14 biochemical parameters to create a logistic regression model to classify the individuals as patient or healthy, according to their biochemical characteristics.

At the beginning, we tried to create a model with a total of all 14 biochemical parameters that were measured, since it is known that levels of UREA and CREA are dramatically increased in ESRF patients. The only parameter selected by the forward LR stepwise method was CREA, and 100.00% of the AI and 98.70% of the DPI were correctly classified by the model, but the significance level of the Wald statistic was large (greater than 0.05) and consequently, CREA was not useful to the model. After that, CREA was left out of the model and the same procedure was repeated with the remaining 13 parameters. The resultant model included again one variable, that of UREA, and 100.00% of the AI and 100.00% of the DPI were correctly classified by the model; however, again the parameter UREA was not useful to the model (the level of the Wald statistic was greater than 0.05).

Thus, the parameters CREA and UREA were left out and a model was created taking into account only the first six parameters, ALB, K, TG, ALP, AST, and ALT, which were resulted by stepwise DA when CREA and UREA were excluded. In this case, the parameters selected by BLRA (forward LR stepwise method) were ALB, K, ALP, and TG, and all had significant changes in $-2\log$ -likelihood. Furthermore, the significance level of the Wald statistic was less than 0.05.

The equation of the model was as follows:

$$Z_i = -8.481 \cdot ALB_i + 4.599 \cdot K_i + 0.104 \cdot ALP_i + 0.048 \cdot TG_i, \quad (1)$$

where Z_i is the value of the unobserved continuous variable for the i th individual, and the probability function is

$$p_i = P(A_i) = 1 \cdot (1 + e^{-Z_i}),$$

where $P(A_i)$ is the probability the i th individual of being healthy or patient.

For the above model, the Hosmer–Lemeshow statistic χ^2_8 was calculated 0.082 (1.000); Cox and Snell's R^2 was 0.736; and Nagelkerke's R^2 was 0.981. The Hosmer–Lemeshow statistic indicates a poor fit if the significance value is less than 0.05; here the model adequately fits the data. Cox and Snell's R^2 , with categorical outcomes, has a theoretical maximum value of less than 1 even for a

“perfect” model. Nagelkerke's R^2 is an adjusted version of Cox and Snell's R^2 that adjusts the scale of the statistic to cover the full range from 0 to 1.

The above model resulted in better classification than the respective model of DA. In the case of BLRA, 98.92% of the HI, 98.97% of the KPI, and 90.67% of the DPI were correctly classified. It is important to notice that also in this case, the parameters used are routinely determined in clinical practice.

DA Application (after BLRA)

Based on the better results of BLRA when using only four parameters (ALB, K, ALP, and TG), we performed DA again with stepwise mode (Wilks' lambda and the Chi-square were 0.164 and 504.465, respectively, at $P < 0.001$), and with these four parameters. The results were about the same as those achieved by BLRA, as they correctly classified 98.92% of the HI, 97.96% of the KPI, and 89.33% of the DPI. This is explained by the fact that the removal of AST reduced the weak collinearity problem between AST and ALP (DA attempts to find linear combinations between independent variables).

Application of CA

Another exploratory approach involved the use of hierarchical CA on log-transformed standardized data of the above 14 biochemical parameters. Hierarchical CA produced a dendrogram with two groups: HIs and ESRF patients, respectively (Fig. 2).

The application of CA gave 100% correct prediction for the sample of AI and 98.67% correct prediction for the DPI. The CA was repeated using only five biochemical parameters (TG, K, AST, ALP, and ALB), which were extracted from DA and used in BLRA. The results are similar to those that were obtained by BLRA; 99.46% of the HI, 97.96% of the KPI, and 90.67% of the DPI were correctly classified. In the future, when a new individual would be classified as healthy or patient, this case could be added to the existing data set of individuals and the procedure repeated.

DISCUSSION

Two data sets (185 HIs and 173 ESRF patients), each one including 18 biochemical parameters, were analyzed. Only 14 of the 18 parameters were used for statistical analysis because they were significantly different between the groups of healthy and patients. More specific, the parameters GLU, CHOL, Fe, and Cl were excluded from the analysis as they did not differ significantly between the studied groups, according to Levene and t -tests.

Fig. 2. Hierarchical dendrogram for grouping of healthy individuals and ESRF patients.

Descriptive statistics and Pearson correlation test gave basic information about the biochemical laboratory data of the studied groups. As expected, ALB, AST, HDL-C, Ca, TP, and Na concentrations were higher; and ALP, ALT, CREA, K, P, TG, UA, and UREA were lower in HI group. Moreover, Pearson correlation test ($r > 0.6$) showed that there was strong relationship between the couples of parameters AST–ALT and TP–ALB, for both HIs and ESRF patients' data sets. It was evident that the simple correlation analysis did not indicate specific links among the studied biochemical parameters.

Three different environmetric methods were applied for the analysis of clinical laboratory data to model the above-mentioned 14-parameter data sets of AI, DHI, KPI, and DPI.

The application of DA (stepwise mode) on the above data sets finally yielded a model based only on five parameters (ALB, K, TG, ALP, and AST) and classified correctly 99.5% of the HI, 98.0% of the KPI, and 85.3% of the DPI. It is worthy of note that the above five biochemical parameters are not specific-only for renal function but instead, they are used routinely in clinical laboratory practice. Based on the results of DA, BLRA gave another model by using only four of the parameters used by DA model (ALB, K, ALP, and TG); the parameter AST was excluded, classified correctly 98.9% of the HI, 99.0% of the KPI, and 90.7% of the DPI. As it can be seen, BLRA model gave better results than DA model and used only four biochemical parameters. Consequently, when a new individual is to be examined, we need the values of the model parameters for this new person and depending on the value that would result from the model (0 or 1), this person can be classified as healthy or an ESRF patient. Then, DA (stepwise mode) was again applied to the data sets of the four biochemical parameters used by BLRA model. The yielded model of DA gave better result than the older DA model and they are similar as those of BLRA model, as they correctly classified 98.9% of the HI, 98.0%

of the KPI, and 89.3% of the DPI. Consequently, the interaction of DA and BLRA statistical methods produced models with similar results used only four biochemical parameters that are same for both models.

Another method, CA, was applied to the log-transformed standardized data of the 14 biochemical parameters and produced a dendrogram (of the cases [healthy and patients]) with two clusters, one of HIs and another of ESRF patients. The application of CA gave 100% correct prediction for the group of KAI and 98.7% correct prediction for the DPI group. The same method was applied to the data set of only five biochemical parameters that were extracted from DA and used in BLRA. The results were similar to those of BLRA; 99.5% of the HI, 98.0% of the KPI, and 90.7% of the DPI were correctly classified.

Therefore, the classification of all the studied cases by the use of CA revealed information about separation due to different levels of biochemical tests. It is indicated that the state of the various HIs and ESRF patients is dependent on different links among the above biochemical parameters. Thus, when a new individual would be classified as healthy or ESRF patient, this case could be added to the existing data set of individuals and the procedure repeated.

Finally, for a better overview of the results, Table 4 summarizes the produced models and their results. As it can be seen in Table 4, the model yielded by BLRA is better in terms of applicability in routine clinical practice because it uses the data as they are without log-transformation, includes only four parameters, and gives better results when compared with the other two models.

With the proposed models, we dispose of more tools to verify the health condition of an individual, meaning that we are able to classify an individual as healthy or ESRD patient by determining the biochemical parameters ALB, K, ALP, and TG, and for the CA model, one more parameter AST. The small number of biochemical

TABLE 4. Chemometric Methods, Produced Models, and Their Results

Method	Data matrix of 14 biochemical parameters	Parameters of the model	Correctly classified
DA	Log-transformed data	ALB, K, ALP, TG.	98.92% of the HI, 97.96% of the KPI, and 89.33% of the DPI
BLRA	Original data	ALB, K, ALP, TG.	98.92% of the HI, 98.97% of the KPI, and 90.67% of the DPI
CA	Log-transformed standardized data	ALB, K, ALP, TG and AST.	99.46% of the HI, 97.96% of the KPI, and 90.67% of the DPI

Note: DA, discriminant analysis; BLRA, binary logistic regression analysis; CA, cluster analysis; HI, healthy individual; KPI, Kavala's patient individual; DPI, Drama's patient individual.

parameters required in the proposed models does not add to the cost of hospitalization, but instead allows the clinician to realize a better diagnosis for the patient, thus achieving economies of scale and better health services to the citizens.

CONCLUSION

A combination of DA, BLRA, and CA is used to create models that could assess and model biochemical laboratory data of HIs and ESRF patients, based only on their routinely determined biochemical parameters. Consequently, our main findings were as follows:

- (i) DA creates a model that only used four parameters (ALB, K, ALP, and TG) and classified correctly 98.92% of the diagnosed HIs from Drama's hospital (DHI), 97.96% of the diagnosed ESRF patients from Kavalas hospital (KPI), and 89.33% of the diagnosed ESRF patients from Dramas hospital (DPI).
- (ii) BLRA, based on the results of DA, created another model by using the same parameters (ALB, K, ALP, and TG) of the DA model above, classified correctly 98.92% of the HI, 98.97% of the KPI, and 90.67% of the DPI.
- (iii) The application of CA classified correctly 99.46% of the HI, 97.96% of the KPI, and 90.67% of the DPI, based only on five parameters, the same four parameters used by DA or BLRA models above, and one more parameter AST.

It is important to note that the above environmetric methods interacted harmoniously to model the biochemical parameter data sets of AI, HI, KPI, and DPI.

REFERENCES

1. Ardissino G, Daccó V, Testa S, et al. Epidemiology of chronic renal failure in children: Data from the ItalKid project. *Pediatrics* 2003;111:e382–e387.
2. Gupta R, Birnbaum Y, Uretsky BF. The renal patient with coronary artery disease: Current concepts and dilemmas. *J Am Coll Cardiol* 2004;44:1343–1353.
3. Yerkey MW, Kernis SJ, Franklin BA, Sandberg KR, McCullough PA. Renal dysfunction and acceleration of coronary disease. *Heart* 2004;90:961–966.
4. Coresh J, Byrd-Holt D, Astor BC, et al. Chronic kidney disease awareness, prevalence, and trends among U.S. adults, 1999 to 2000. *J Am Soc Nephrol* 2005;16:180–188.
5. Asselbergs FW, Mozaffarian D, Katz R, et al. Association of renal function with cardiac calcifications in older adults: The cardiovascular health study. *Nephrol Dial Transplant* 2009;24:834–840.
6. McCullough PA, Soman SS, Shah SS, et al. Risks associated with renal dysfunction in patients in the coronary care unit. *J Am Coll Cardiol* 2000;36:679–684.
7. Curtis B, Barrett BJ, Levin A. Identifying and slowing progressive chronic renal failure. *Can Fam Physician* 2001;47:2512–2518.
8. Stigant C, Stevens L, Levin A. Nephrology: Strategies for the care of adults with chronic kidney disease. *JAMC* 2003;168:s1553–1560.
9. Curtis B, Barrett BJ, Levin A. Identifying and slowing progressive chronic renal failure. *Can Fam Physician* 2001;47:2512–2518.
10. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. *Nephron* 1976;16:31–41.
11. Larsson A, Malm J, Grubb A, Hansson LO. Calculation of glomerular filtration rate expressed in mL/min from plasma cystatin C values in mg/L. *Scand J Clin Lab Invest* 2004;64:25–30.
12. Poggio ED, Wang X, Greene T, Van Lente F, Hall PM. Performance of the modification of diet in renal disease and Cockcroft-Gault equations in the estimation of GFR in health and in chronic kidney disease. *J Am Soc Nephrol* 2005;16:459–466.
13. Selvin E, Köttgen A, Coresh J. Kidney function estimated from serum creatinine and cystatin C and peripheral arterial disease in NHANES 1999–2002. *Eur Heart J* 2009;30:1918–1925.
14. Tsinialis D, Thiel GT. An easy to calculate equation to estimate GFR based on inulin clearance. *Nephrol Dial Transplant* 2009;1–7.
15. Pöge U, Gerhardt T, Woitas RP. Estimation of glomerular rate by use of beta-trace protein. *Clin Chem* 2008;54:1403–1405.
16. Pizzarelli F, Lauretani F, Bandinelli S, et al. Predictivity of survival according to different equations for estimating renal function in community-dwelling elderly subjects. *Nephrol Dial Transplant* 2009;24:1197–1205.
17. Shlipak MG, Katz R, Sarnak MJ, et al. Cystatin C and prognosis for cardiovascular and kidney outcomes in elderly persons without chronic kidney disease. *Ann Intern Med* 2006;145:237–246.
18. Stevens LA, Fares G, Fleming J, et al. Low rates of testing and diagnostic codes usage in a commercial clinical laboratory: Evidence for lack of physician awareness of chronic kidney disease. *Am Soc Nephrol* 2005;16:2439–2448.
19. Soyulu A, Kasap B, Demir K, Turkmen M, Kavukcu S. Predictive value of clinical laboratory variables for vesicoureteral reflux in children. *Pediatr Nephrol* 2007;22:844–848.

20. Lee D, Levin A, Roger SD, McMahon P. Longitudinal analysis of performance of estimated glomerular filtration rate as renal function declines in chronic kidney disease. *Nephrol Dial Transplant* 2009;24:109–116.
21. Serra MA, Puchades MJ, Rodriguez F, et al. Clinical value of increased serum creatinine concentration as predictor of short-term outcome in decompensated cirrhosis. *Scand J Gastroenterol* 2004;11:1149–1153.
22. Kronborg J, Solbu M, Njølstad I, Toft I, Eriksen BO, Jenssen T. Predictors of change in estimated GFR: A population-based 7-year follow-up from the Tromsø study. *Nephrol Dial Transplant* 2008;23:2818–2826.
23. Tietz NW. Specimen collection and processing; sources of biological variation. *Textbook of Clinical Chemistry*, second edition, Philadelphia, PA: W.B. Saunders; 1994.
24. Deming S. Chemometrics: An overview. *Clin Chem* 1986;32:1702–1706.
25. Vanderginste B, Massart DL, Buydens L, De Jong S, Lewi P, Smeyers-Verbeke J. *Handbook of Chemometrics and Qualimetrics*. Amsterdam: Elsevier; 1998.
26. Moore JH, Parker JS, Olsen NJ, Aune TM. Symbolic discriminant analysis of microarray data in autoimmune disease. *Genet Epidemiol* 2002;23:57–69.
27. Knowles CH, Eccersley AJ, Scott SM, Walker SM, Reeves B, Lunniss PJ. Linear discriminant analysis of symptoms in patients with chronic constipation. *Dis Colon Rectum* 2000;43:1419–1426.
28. Januzzi JL, van Kimmenade R, Lainchbury J, et al. NT-proBNP testing for diagnosis and short-term prognosis in acute destabilized heart failure: An international pooled analysis of 1256 patients. *Eur Heart J* 2006;27:330–337.
29. Steyerberg EW, Eijkemans MJC, Harrel FE, Habbema JDF. Medical decision making. *Med Decis Making* 2001;21:45–56.
30. Chadwick D, Arch B, Wilder-Smith A, Paton N. Distinguishing dengue fever from other infections on the basis of simple clinical and laboratory features: Application of logistic regression analysis. *J Clin Virol* 2006;35:147–153.
31. Thwaites GE, Chau TT, Stepniewska K, et al. Diagnosis of adult tuberculous meningitis by use of clinical and laboratory features. *Lancet* 2002;360:1287–1292.
32. Vogt W, Nagel D. Cluster analysis in diagnosis. *Clin Chem* 1992;38:182–198.
33. Clinton D, Button E, Norring C, Palmer R. Cluster analysis of key diagnostic variables from two independent samples of eating-disorder patients: Evidence for a consistent pattern. *Psychol Med* 2004;34:1035–1045.
34. Ness R, Kip K, Hillier S, et al. A cluster analysis of bacterial vaginosis-associated microflora and pelvic inflammatory disease. *Am J Epidemiol* 2005;162:585–590.
35. Toraldo DM, Nicolardi G, De Nuccio F, Lorenzo R, Ambrosino N. Pattern of variables describing desaturator COPD patients, as revealed by cluster analysis. *Chest* 2005;128:3828–3837.
36. Nikolov M, Simeonova P, Simeonov V. Chemometrics as an option to assess clinical data from diabetes mellitus type 2 patients. *Central Eur J Med* 2009;4:433–443.
37. Bardarov V, Simeonova P, Neikova L, et al. Statistical interpretation of medical data of patients with alcohol abuse. *Central Eur J Med* 2012;7:465–474.
38. Papaioannou A, Simeonov V, Plageras P, Dovriki E, Spanos T. Multivariate statistical interpretation of laboratory clinical data. *Eur J Med* 2007;3:319–334.
39. IFFC. Approved recommendation on the theory of reference values. Part 4. Theory of reference values. Control of analytical variation in the production, transfer and application of reference values. *Clin Chim Acta* 1991;202:S5–S12.
40. Grossi E, Colombo R, Cavuto S, Franzini C. The REALAB project: A new method for the formulation of reference intervals based on current data. *Clin Chem* 2005;51:1232–1240.
41. Kafka MT. Internal quality control, proficiency testing and the clinical relevance of laboratory testing. *Arch Pathol Lab Med* 1988;112:449–453.