Journal of Pharmaceutical and Biomedical Analysis 50 (2009) 210–215

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Use of multidimensional data analysis for prediction of lung malignity

Viera Mrázová[a], Ján Mocák[a,*], Elena Varmusová[b], Denisa Kavková[b], Adriana Bednárová[a]

[a] Department of Chemistry, University of SS. Cyril and Methodius, J. Herdu 2, Trnava SK-917 01, Slovak Republic
[b] Institute for Tuberculosis and Respiratory Diseases, Department of Clinical Chemistry, Kvetnica, Poprad, SK-058 87, Slovak Republic

ARTICLE INFO

Article history:
Received 20 February 2009
Received in revised form 27 March 2009
Accepted 16 April 2009
Available online 23 April 2009

Keywords:
Lung malignity
Tumor markers
Multidimensional analysis
ROC curves

ABSTRACT

Diagnosis of lung malignity can be predicted or confirmed not only according to the values of appropriate laboratory tests but also using multidimensional statistical analysis, which uses simultaneously all performed tests in the form of their optimal combination. The developed new way of diagnosis prediction is applied here to the results of laboratory analysis of lung tumor markers in serum as well as pleural effusion (exudate). Four laboratory tests were used and investigated in detail: carcinoembryonic antigen, CEA, in serum as well as in pleural exudate, and cytokeratin 19 fragment, CYFRA, in serum and exudate, as well. Each test represents one dimension in the investigated biomedical problem from the statistical point of view. Joint utilization of the performed laboratory tests is based on their optimized combination into a new statistical variable using a selected chemometric principle (principal component, discriminant function, or logit in logistic regression). This approach results in enhancement of diagnostic effectiveness applied for the specified purpose.

1. Introduction

Lung cancer is the leading cause of cancer deaths in Europe. The most effective and generally recognized positive test for lung cancer is based on histology of the appropriate tissue sample however this way is relatively invasive and, above all, takes a long time. On the contrary, the use of tumor markers is less invasive and takes much shorter time. It may, therefore, prevent the loss of time necessary for medical treatment in urgent cases. The most common tumor markers, which have been found to be of diagnostic significance for lung tumor diseases, are carcinoembryonic antigen (CEA), and cytokeratin 19 fragment (CYFRA). CEA was identified in 1965 and has been widely used for pursuing various tumors, i.e. colorectal cancers [1–3], breast cancer [4–6], and endometrial cancer [7]. Similarly to CEA, the CYFRA assay measures cytokeratin 19 fragment in the body fluids (mainly serum) and its concentration is increased with the extent of the malignant disease, e.g. oral cancer [8], bladder cancer [9,10], esophageal cancer [11], and gynecological cancer [12–14]. The content of the serum CYFRA differs significantly according to disease stage and performance status. Serum and pleural fluid CYFRA 21-1 are useful as the measures in differentiating malignant and benign disease [15]. CYFRA 21-1 assay may be a useful tumor marker for discriminating benign from malignant pleural effusion, especially in those of non-small cell lung cancer. The combined use of CEA and CYFRA 21-1 assay in the malignant effusion may increase the diagnostic yield compared with CEA or CYFRA 21-1 alone [16]. It is important for this study that the determined levels of CEA and CYFRA may help to establish efficiently the diagnosis of lung cancer [17–28].

Pleural effusion is common for several kinds of lung illnesses in clinical practice; malignancy is one of its main causes. Greater than 90% of malignant pleural effusions are due to metastatic disease, mainly from lung or primary breast malignancies. The initial diagnostic approach includes examinations: thoracocentesis, cytology, and biochemical laboratory tests. However, the sensitivity of several mentioned non-invasive techniques is considered to be only 50–70%. To improve upon these rates, a number of tumor markers (TM) in the pleural effusion have been intensively evaluated. It means that in addition to traditional determination in blood serum the same TM are monitored in pleural fluid. Malignant pleural effusions have higher levels of pleural fluid markers than did effusions induced at benign conditions [29].

In our chemometric paper [30], devoted to prediction of cardiovascular risk by means of nine cardiovascular markers, it was discovered that, in general, diagnosis of any disease can be predicted/confirmed not only inspecting a series of single laboratory tests but also using multivariate statistical analysis, in which all or the best performing tests are used simultaneously in the form of their optimal (usually linear) combination. This new approach is now applied also to the current lung malignity problem utilizing CEA and CYFRA tumor markers both in serum and pleural effusion.

* Corresponding author. Tel.: +421 3359214403; fax: +421 3359214403.
E-mail address: jan.mocak@ucm.sk (J. Mocák).

0731-7085/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2009.04.018

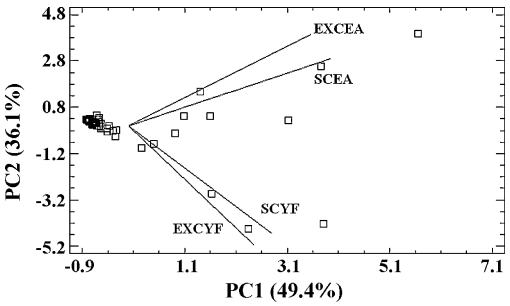

Fig. 1. PCA biplot showing 4 selected clinical variables (EXCYF, SCYF, EXCEA, SCEA) and 53 objects—patient samples. Software Statgraphics Plus 5.1.

2. Experimental

2.1. Investigated clinical data

Tumor markers CEA and CYFRA were determined at the Institute for Tuberculosis and Respiratory Diseases (ITRD) in Poprad—Kvetnica, Slovakia. Two sets of data were investigated:

(1) The first (smaller and older) data set of 53 patients comprising 27 men and 26 women, among them 25 patients with malignant tumor and 28 with benign tumor or other non-malignant disease (including tuberculosis), which was proved by histology. Four clinical variables, indicating the malignant/non-malignant status were composed: EXCEA (CEA in pleural effusion, i.e. exudate), SCEA (CEA in serum), EXCYF (CYFRA in exudate), and SCYF (CYFRA in serum).
(2) The second (larger and newer) data set of 182 probands containing 108 men and 74 women. Here, utilizing histology, 86 of the patient samples were ascertained malignant and 96 with benign tumors, tuberculosis or another non-malignant disease. Three clinical variables were utilized: EXCEA, SCEA, EXCYF, and, in addition, the patient's age (coded as AGE) and the gender of the corresponding individual (coded as SEXN). The reason of omitting SCYF was due to finding that it provides very similar information as EXCYF (their pair correlation coefficient was 0.606) so that it was practically redundant. It was proved also by the results presented further in Section 3.1 and exhibited by the smallest angle between the mentioned variables in Fig. 1.

When using classification of both data sets by the selected multivariate statistical techniques, the binary categorical variable Dg (diagnosis) was used for the patients' sample classification; it acquires two possible values: (1) indicating malignant disease, and (2) other, non-malignant disease. This categorization was performed on the basis of known histology results so that the corresponding categorized data established the training set when using chemometric terminology.

2.2. Statistical data analysis

Statistical calculations were performed using the following multidimensional techniques: principal components analysis (PCA) [31], cluster analysis (CA) [32], discriminant analysis (linear, LDA, and quadratic, QDA) [33], logistic regression (LR) [34], Kth nearest neighbor method (KNN) [33], and artificial neural network (ANN) [35]. In addition, analysis of variance (ANOVA), correlation analysis and ROC analysis was also implemented [36]. Several software commercial packages were used, particularly STAGRAPHICS Plus 5.1, SPSS 15.0, JMP 6.0.2, SAS 9.1.3, and Trajan 6.0 (for ANN calculations).

2.3. Analytical procedures

Tumor markers were analyzed in the clinical biochemical laboratory using automatic analyzers ELECSYS 1010 and ELECSYS 2010, based on immunoanalysis with electrochemically generated chemiluminiscent detection. Determination of tumor markers in pleural effusion was made according to the original procedure developed at the ITRD.

3. Results and discussion

3.1. Principal component analysis (PCA)

The smaller data set of the patients, containing four variables, namely SCEA, EXCEA, EXCYF and SCYF, was used for a preliminary study. It was found in the way, described above in Section 2.1 that the variable SCYF is of the least importance. Considering this as well as the economical aspects, the variable SCYF was used only in the preliminary studies and then it was omitted. Instead, the variable AGE, always accessible, was used in further PCA study.

PCA reveals a natural grouping of the studied objects (patient samples) as well as the selected variables in a reduced dimensional space. The first principal component (PC1) as well as further PCs perform a linear combination of four original variables, optimized with respect to preserving maximal variance of original data. The variables are demonstrated in the exhibited PCA biplot by the rays connecting the variable position in the PC2–PC1 plane with the origin. First two PCs contain the most part of the variance retired in original data so that PC3 and PC4 are practically unimportant. The inspection of the biplots depicted in Figs. 1 and 2 reveals that the PC1 axis represents malignancy. All tumor marker variables are positively correlated with the PC1, which means that all patient samples with a high PC1 value indicate malignancy. It is in accordance with the observation (different colors of the categories in original pictures) that the benign cases are located in a dense cluster located at most negative PC1 values and the malignant cases are located at high PC1 values. Similarity of the EXCEA and SCEA as well as EXCYF and SCYF couples, which is observable in Fig. 1, is expected. Large PC2 value (the sense of which was not clarified) and a perpendicular position of AGE with respect to PC1 (Fig. 2) indicate independence of AGE on malignancy as mirrored by three used tumor markers. However, AGE was not excluded from further investigations since it might provide another kind of information relevant to diagnosis improvement when using techniques of multivariate data analysis. Similar situation happened in our investigation of cardiovascular markers where CRP did not correlate with any further markers (cholesterols, lipoproteins, etc.) but its role in diagnosis prediction was significant [30]. Nevertheless, a risk when an unimportant variable is kept may be reflected by increased uncertainty of the achieved results, which can be proved by running data analysis with and without such a variable.

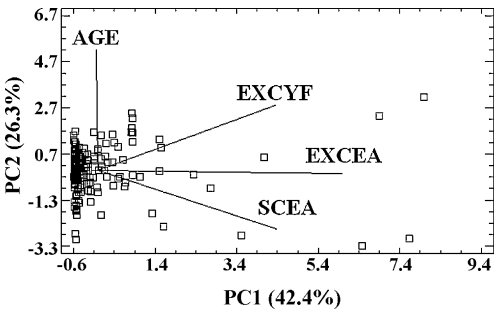

Fig. 2. PCA biplot showing 4 selected variables (EXCYF, EXCEA, SCEA, AGE) and 182 objects—patient samples. Software Statgraphics Plus 5.1.

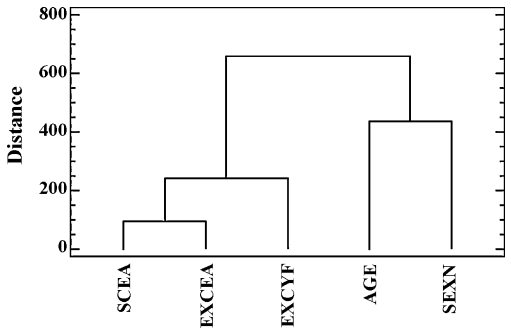

Fig. 3. Clustering of variables. Ward cluster analysis method and squared Euclidean distance metrics were applied. 182 patient samples with lung diseases. Software Statgraphics Plus 5.1.

3.2. Cluster analysis (CA)

Among several examined clustering techniques, Ward method [32] with squared Euclidean distance metrics were selected as the best for variable clustering. The obtained results achieved for the larger data set (Fig. 3) are in agreement with clinical expectations. EXCEA and SCEA are clustered with EXCYF and distances among them are the shortest so that all variables (tumor markers) indicating positive diagnosis result are most similar. Variable AGE is clustered with SEXN; their correlation may simply reflect the fact that the average age of women and men is different (higher at women compared to men).

3.3. Analysis of variance (ANOVA)

In this work a one-way analysis of variance was performed for each of three quantitative variables dependent on a single factor (diagnosis, Dg) used as independent variable. ANOVA was here used to test the hypothesis that several means are equal. The patient samples were assigned according to diagnosis and gender into four sample categories using categorical variable Dg: Dg = 0 for non-malignant men samples, Dg = 1 for malignant men samples, Dg = 2 for non-malignant women samples and Dg = 3 for malignant women samples. Then the connection between the diagnosis and gender on one part and the level of the analyzed tumor markers and age of the patients on another part were surveyed. Following this aim two types of statistical tests were used—LSD (least significance variance) test and Bonferroni test. From the obtained results it can be concluded that EXCEA and EXCYF are the best markers since they separate all four categories. The tumor marker SCEA is not as much ideally dependent on the created categories because it does not separate the non-malignant men samples from the malignant women samples. Variable AGE was proved insignificant in separating any of the created categories. The detailed outputs of ANOVA are summarized in Table 1.

Table 2
The correlations table showing pair correlation coefficients r for all continuous variables, software JMP 6.0.2, number of patients: 182, critical value $r_{crit} = 0.145$.

	SCEA	EXCEA	EXCYF	AGE
SCEA	1.000			
EXCEA	**0.478**	1.000		
EXCYF	0.043	**0.477**	1.000	
AGE	−0.035	−0.026	0.064	1.000

3.4. Correlation analysis

Table 2 is the correlation table summarizing Pearson sample correlation coefficients r, which express the strength of the linear relationships between each pair of the response variables. It is evident that the strongest dependence is between two pairs of tumor markers mutually (EXCEA vs. SCEA and EXCEA vs. EXCYF). They are significantly correlated since these r values are larger than the corresponding critical value however the values below 0.5 indicate not very strong mutual correlations. The critical values enabling to state significant correlation depend vastly on the number of observations n; for large n even a relatively small r value might be significant (e.g. for $n = 182$ any r larger than $r_{crit} = 0.145$ [37]). Variable AGE is not significantly correlated to any of the remaining variables, which indicates that the level of tumor markers does not depend on the patient's age.

Table 1
Selected ANOVA outputs of all significant differences between the specified categories[a] as obtained by software SPSS 15.0.

Multiple comparisons		Compared categories		p[b]	Multiple comparisons		Compared categories		p[b]
Dependent variable	Performed test	(I) Dg	(J) Dg		Dependent variable	Performed test	(I) Dg	(J) Dg	
SCEA	LSD	0	1	0.0023	EXCEA	Bonferr.	2	1	0.0022
	LSD	1	0	0.0023		Bonferr.	2	3	0.0163
	LSD	1	2	0.0079		Bonferr.	3	0	0.0039
	LSD	2	1	0.0079		Bonferr.	3	2	0.0163
	Bonferr.	0	1	0.0140	EXCYF	LSD	0	1	0.0001
	Bonferr.	1	0	0.0140		LSD	0	3	0.0008
	Bonferr.	1	2	0.0472		LSD	1	0	0.0001
	Bonferr.	2	1	0.0472		LSD	1	2	0.0003
EXCEA	LSD	0	1	0.0000		LSD	2	1	0.0003
	LSD	0	3	0.0006		LSD	2	3	0.0014
	LSD	1	0	0.0000		LSD	3	0	0.0008
	LSD	1	2	0.0004		LSD	3	2	0.0014
	LSD	2	1	0.0004		Bonferr.	0	1	0.0007
	LSD	2	3	0.0027		Bonferr.	0	3	0.0046
	LSD	3	0	0.0006		Bonferr.	1	0	0.0007
	LSD	3	2	0.0027		Bonferr.	1	2	0.0019
	Bonferr.	0	1	0.0003		Bonferr.	2	1	0.0019
	Bonferr.	0	3	0.0039		Bonferr.	2	3	0.0082
	Bonferr.	1	0	0.0003		Bonferr.	3	0	0.0046
	Bonferr.	1	2	0.0022		Bonferr.	3	2	0.0082

[a] Four specified categories belong to: non-malignant men samples (Dg = 0), malignant men samples (Dg = 1), non-malignant women samples (Dg = 2), and malignant women samples (Dg = 3).

[b] p - significance level; the difference of means of two categories (I) and (J) was qualified significant if $p \leq 0.0500$. Insignificant combinations were not included.

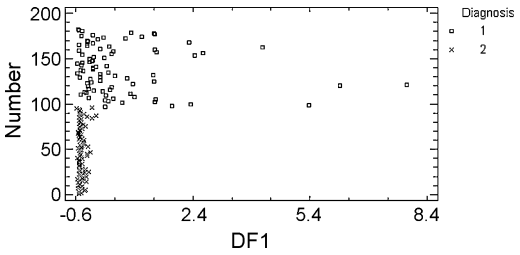

Fig. 4. Classification of the patient samples by linear discriminant analysis. The samples are specified by corresponding numbers on the vertical axis, DF1 denotes the only discriminant function. 86 examined patient samples confirmed by histology as malignant (Dg = 1) and 96 samples corresponding to benign tumors or tuberculosis (Dg = 2). Software Statgraphics Plus 5.1.

3.5. Classification of patient samples by diagnosis

Discriminant analysis, which is a set of multivariate classification techniques, can be used for classification of the investigated samples into known categories. Many different types of models can be employed for performing the discrimination including parametric (linear and quadratic) discriminators and nonparametric (e.g., kernel based or k-nearest neighbor) discriminators [33,38].

The overall procedure of the applied classification multivariate methods consists of three steps: (1) to create the training data set by means of diagnostic categories using the table entries of individual samples with known diagnosis, (2) to calculate a classification model using the categorized patient samples of the training set, and (3) to validate the categorization of the samples into the selected classes when using the samples not included into the training set. For this purpose the multidimensional techniques of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression (LR), and Kth nearest neighbor were used.

Fig. 4 represents the LDA graphical output, which shows that the non-malignant patient samples (numbers 1–96) are located in a narrow cluster at very negative values of the first discriminate function (DF1) whilst the malignant samples (97–182) form a wide and tailing cluster at higher DF1 values—such a behavior is typical for clinical studies performed by discriminant analysis.

The figures of classification performance of four applied classification methods are collected in Table 3. The exhibited classification results regard three types of the sample data: (1) the training set samples, from which the classification model is calculated, (2) the samples omitted from the training set in a step-by-step manner according to the leave-one-out validation procedure [39], and (3) the samples, which were selected to form a special validation set and not belonging to the training set.

The predictive ability of the used multivariate methods should be judged by the results shown in the last two columns, which are independent of the calculations performed with the training set. Surprisingly, in several cases the results achieved for the test set were better than those for the training set even though the opposite is generally expected. It should be stressed that the selection which of the samples are inserted into the test set was performed in a random way, that is generally accessible (e.g. in MS Excel); better results than those for the training set are therefore accidental. The best results pertaining to different groups of classification methods are marked in the table by bold typefaces. For leave-one-out validation the QDA results are better than the LDA ones. Several tabulated results for KNN correspond to various k-values (the used number of neighbors), with $k = 5$ as the best. However, the best results in total (90.0%) were achieved by logistic regression, where beside the variables used in other techniques, the patient's gender was additionally used in the form of the binary variable SEXN (expressing gender woman/man). It should be noted that among the utilized techniques only LR has an advantage of using also a categorical input variable so that the number of combined variables is here higher by one compared to other techniques.

Table 3
Classification results by diagnosis (success in %) for various multivariate methods and software SAS 9.1.3.

Classification method	Training set, success in %	Leave-one-out, success in %	Validation set, success in %
LDA	74.6	75.3	86.7
QDA	86.3	**85.2**	**86.7**
KNN, $k = 3$	89.0	78.6	83.3
KNN, $k = 5$	86.8	**82.4**	**86.7**
KNN, $k = 7$	85.2	79.7	83.3
KNN, $k = 9$	81.9	78.0	86.7
KNN, $k = 11$	80.2	76.4	86.7
LR	88.5	88.5	**90.0**
LR (+SEXN)	89.6	**89.6**	**90.0**

Note: Decision upon malignity is predicted using the measured values of four variables: EXCEA, SCEA, EXCYF, and AGE. The last row refers to LR where SEXN was also used as the fifth variable.

3.6. ROC curves

ROC is an old technique originally used for evaluation of radiolocation receivers (ROC-receiver operating characteristic), which is nowadays frequently used for qualitative evaluation of the performance of laboratory methods [30]. The ROC curve is a graphical plot of the sensitivity vs. (1-specificity) for a binary classifier system in which the discrimination threshold is varied. In order not to interfere with sensitivity and specificity defined in analytical chemistry these terms are sometimes denoted as the measure of sensitivity and measure of specificity [40]. When one positive and one negative case are randomly picked up, the area under ROC curve expresses the probability that the classifier (using the current threshold value) will assign a higher score to the positive example than to the negative.

The ROC curves pertaining to three tumor marker variables and further variables, combined in this work, are shown in Fig. 5. Compared to individual assessment of the tumor marker performance much more advantageous is a combination of all three marker variables together.

In three above mentioned multivariate techniques, i.e. PCA, LDA and LR, the most important new variables, combined from the originally used tumor marker variables, are the first principal component PC1, the first discriminant function DF1, and logit, respectively. These were used in the ROC analysis in the same way as the original tumor marker variables. The results of ROC analysis are depicted in Fig. 5. It is obvious that new combined variables provide a larger area under the ROC curve compared to any individual marker, which means they provide more information on malignity and can be considered as better diagnostic tools. The largest ROC area (0.908) was observed for logit_4 composed of four variables (EXCEA, SCEA, EXCYF and AGE). Only insignificantly smaller (0.906) is the area for logit_5 composed of the mentioned variables plus categorical variable SEXN (denoting gender of the patient).

3.7. Classification of patient samples by artificial neural networks

The artificial neural networks (ANN) were also used for prediction of the lung malignity. The neural network was defined by three layer perceptron; the back propagation algorithm was used for this classification problem. ANN has been used also for diagnostic purposes—e.g. Parekattil et al. [41] have developed a neural network to identify patients with bladder cancer more effectively than hematuria and cytology. Their algorithm, based on combined urine levels of nuclear matrix protein-22, monocyte chemoattractant protein-1 and urinary intercellular adhesion molecule-1, was

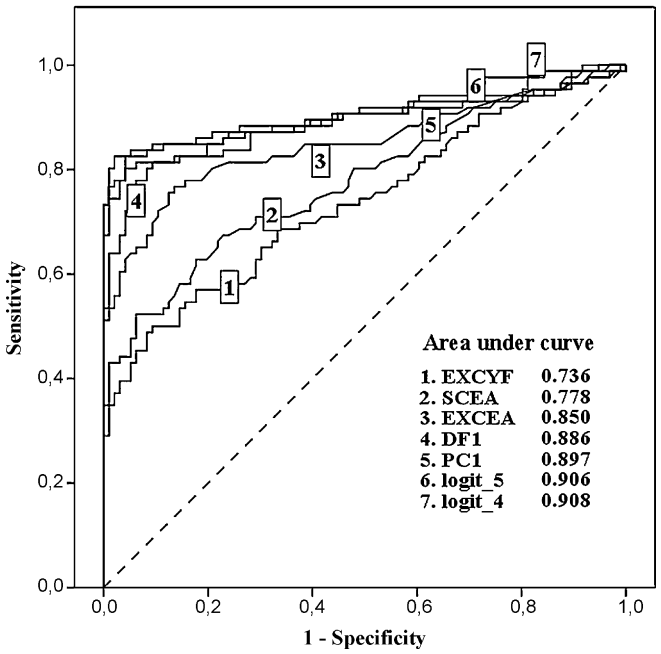

Fig. 5. ROC curves corresponding to three tumor markers and four combined variables obtained by principal component analysis, PC1 (first principal component), linear discriminant analysis, DF1 (first discriminant function) and logistic regression (logit_4, logit_5 composed of four and five variables, respectively). Software SPSS 15.0.

claimed to be superior to conventional screening tests for bladder cancer.

According to recommendations of the implemented Trajan software, our original data set (Section 2.1) was randomly separated into three parts: the training set (122 patient samples), the selection set (30 samples) and the test set (30 samples). The selection set was independently chosen and used by Intelligent Problem Solver of Trajan software for automatic evaluation of effectiveness of a large amount of possible networks and suggests the best possible neural network for further work. Test set, containing other independent samples, is then used for validation of the performed classification. The obtained results are collected in Table 4. The use of five variables in contrast to four variables mentioned in the table has not improved the classification results, which confirms the conclusions achieved by ROC curves. It should be noted that the sample selection performed by Trajan software for the test set could not be influenced by the software user so that the chosen samples differ from the way of selecting the validation set employed by other techniques (shown in Table 3).

3.8. Chemometrically aided prediction of lung malignity

Diagnosis prediction for a patient with unknown diagnosis is facilitated when using a combined variable calculated by PCA, LDA or LR. For instance, the PC1 is a multicomponent obtained in PCA by linear combination of original tumor marker variables using

Table 4
ANN classification results (success in %)—two classes formed by malignant/non-malignant diagnosis as achieved by software package Trajan 6.0.

Sets	Used variables	
	EXCEA, SCEA, EXCYF, AGE	EXCEA, SCEA, EXCYF, AGE, SEXN
Training set	88.5	86.9
Selection set	80.0	83.3
Test set	**96.7**	**96.7**

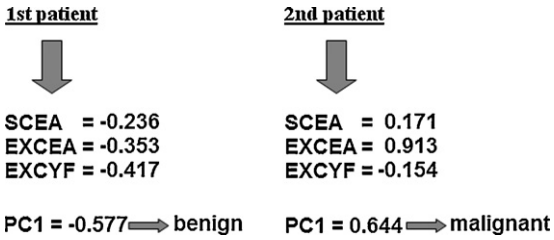

Fig. 6. Scheme for diagnosis prediction according to standardized values of tumor marker content using the cut-off value −0.34 established at the maximum of Efficiency.

standardized data. It means that the mean value of the respective variable (marker) is subtracted from every measured variable value and the result is divided by the variable standard deviation.

The PC1 value for any proband is easily calculated using the coefficients obtained as the eigenvector given in the PCA output:

$$\mathrm{PC1} = 0.6957\,\mathrm{EXCEA} + 0.5074\,\mathrm{EXCYF} + 0.5086\,\mathrm{SCEA} \quad (1)$$

(SCYF and AGE variables were omitted in this equation since it was found that they have only a minor prediction capability). This equation together with the found cut-off value −0.34 on the PC1 axis can be used for every individual either to confirm his/her diagnosis (if it is known) or for a fast diagnosis prediction—if histology has not been made before.

The cut-off value, separating the distributions of malignant and benign samples, respectively, can be found at the maximum of Efficiency, E, defined as

$$E = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{fn} + \mathrm{tn} + \mathrm{fp}} = \frac{\mathrm{tp} + \mathrm{tn}}{n_1 + n_2} \quad (2)$$

for any of the used variables. The symbols tp and tn denote true positive and true negative results achieved by the studied tumor marker, respectively, fp and fn denote false positive and false negative results, respectively. The correct diagnosis was obtained by histology.

The Efficiency values can be calculated either in MS Excel (where the fp and fn cases are filtered off) or by means of a suitable software, e.g. GraphROC for Windows [36,42]. In this way e.g. the cut-off value −0.34 was obtained for the dimensionless PC1 scale and a similar procedure was applied also for further multicomponent variables.

The described diagnosis prediction can be easily implemented in the hospital information system containing the patients' data when the needed PCA coefficients and the cut-off value are delivered, as it was the case in Department of Clinical Chemistry of the Institute for Tuberculosis and Respiratory Diseases in Poprad. Fig. 6 depicts an example of diagnosis prediction for two patients using the above mentioned approach. It should be noted that the standardized values of tumor markers may be also negative, which simply means that they are smaller than the corresponding mean value (as it happened also in four cases shown in Fig. 6).

4. Conclusions

Principal component analysis and cluster analysis allow display a natural grouping of the samples belonging to the individuals, which are under medical treatment, with regard to lung diseases. The exhibited results demonstrate a very good applicability of the used multidimensional statistical methods for graphical representation of the investigated samples in a reduced number of dimensions.

Four multivariate classification methods and artificial neural networks were successfully used for categorization of the patient samples by diagnosing malignant/non-malignant cases. All classification methods, which were used in this study, enable a very good

sample classification by diagnosis however the best results were obtained by artificial neural network and logistic regression. The mentioned types of diagnosis may be predicted or verified for the patients with a lung disease not only by evaluation of the results of the selected best individual laboratory test but also utilizing all performed laboratory tests jointly in the form of their optimal combination ensured by an appropriate multidimensional statistical technique. The use of tumor markers for diagnostic purposes is fast and sufficiently correct when applied in the way developed in this work. The attempts of embedding the mentioned newly created combined variables into the hospital information system were successful and have been used as new diagnostic tools.

Acknowledgements

The authors wish to acknowledge and thank the project VVCE-0004-07 and VEGA 1/1005/09 for support of this work.

References

[1] B. Mroczko, M. Groblewska, U. Wereszczynska-Siemiatkowska, B. Okulczyk, B. Kedra, W. Laszewicz, A. Dabrowski, M. Szmitkowski, Clin. Chim. Acta 380 (2007) 208–212.
[2] M.J. Duffy, A. van Dalen, C. Haglund, L. Hansson, E. Holinski-Feder, R. Klapdor, R. Lamerz, P. Peltomaki, C. Sturgeon, O. Topolcan, Eur. J. Cancer 43 (2007) 1348–1360.
[3] Y. Yamamoto, E. Hirakawa, S. Mori, Y. Hamada, N. Kawaguchi, N. Matsuura, Biochem. Biophys. Res. Commun. 333 (2005) 223–229.
[4] A. Nicolini, A. Carpi, P. Ferrari, G. Rossi, Cancer Lett. 263 (2008) 122–129.
[5] C.H.C.H. Chen, M.F. Hou, J.Y. Wang, T.W. Chang, D.Y. Lai, Y.F. Chen, S.Y. Hung, S.R. Lin, Cancer Lett. 240 (2006) 279–288.
[6] G. Sölétormos, D. Nielsen, V. Schioler, H. Mouridsen, P. Dombernowsky, Eur. J. Cancer 40 (2004) 481–486.
[7] Z. Yurkovetsky, S. Ta'asan, S. Skates, A. Rand, A. Lomakin, F. Linkov, A. Marrangoni, L. Velikokhatnaya, M. Winans, E. Gorelik, G.L. Maxwell, K. Lu, A. Lokshin, Gynecol. Oncol. 107 (2007) 58–65.
[8] S.S. Sawant, S.M. Zingde, M.M. Vaidya, Oral Oncol. 44 (2008) 722–732.
[9] Ch. Andreadis, S. Touloupidis, G. Galaktidou, A.H. Kortsaris, A. Boutis, D. Mouratidou, J. Urol. 174 (2005) 1771–1776.
[10] B. Nisman, V. Barak, A. Shapiro, D. Golijanin, T. Peretz, D. Pode, Cancer 94 (2002) 2914–2922.
[11] Y. Shimada, G. Watanabe, J. Kawamura, T. Soma, M. Okabe, T. Ito, H. Inoue, M. Kondo, Y. Mori, E. Tanaka, M. Imamura, Oncology 68 (2005) 285–292.
[12] E. Pras, P.H.B. Willemse, A.A. Canrinus, H.W.A. de Bruijn, W.J. Sluiter, K.A. ten Hoor, J.G. Aalders, B.G. Szabo, E.G.E. de Vries, Int. J. Radiat. Oncol. Biol. Phys. 52 (2002) 23–32.
[13] R. Molina, X. Filella, J.M. Auge, E. Bosch, A. Torne, J. Pahisa, J.A. Lejarcegui, A. Rovirosa, B. Mellado, J. Ordi, A. Biete, Anticancer Res. 25 (2005) 1765–1771.
[14] A. Gadducci, M. Ferdeghini, S. Cosio, A. Fanucchi, R. Cristofani, A.R. Genazzani, Int. J. Gynecol. Cancer 11 (2001) 277–282.
[15] W. Dejsomritrutai, S. Senawong, B. Promkiamon, Respirology 6 (2001) 213–216.
[16] R.S. Lai, C.C. Chen, P.C. Lee, J.Y. Lu, Jpn. J. Clin. Oncol. 29 (1999) 421–424.
[17] S. Holdenrieder, J. von Pawel, E. Dankelmann, T. Duell, B. Faderl, A. Markus, M. Siakavara, H. Wagner, K. Feldmann, H. Hoffmann, H. Raith, D. Nagel, P. Stieber, Lung Cancer 63 (2009) 128–135.
[18] T. Muley, T.H. Fetz, H. Dienemann, H. Hoffmann, F.J.F. Herth, M. Meister, W. Ebert, Lung Cancer 60 (2008) 408–415.
[19] M.M. van den Heuvel, C.M. Korse, J.M.G. Bonfrer, P. Baas, Lung Cancer 59 (2008) 350–354.
[20] I.C. Wagner, M.J. Guimarães, L.K. da Silva, F.M. de Melo, M.T. Muniz, J. Bras. Pneumol. 33 (2007) 185–191.
[21] G. Paşaoğlu, A. Zamani, G. Can, O. İmecik, Eur. J. Gen. Med. 4 (2007) 165–171.
[22] K. Matsuoka, S. Sumitomo, N. Nakashima, D. Nakajima, N. Misaki, Eur. J. Cardiothorac. Surg. 32 (2007) 435–439.
[23] S. Mizuguchi, N. Nishiyama, T. Iwata, T. Nishida, N. Izumi, T. Tsukioka, K. Inoue, T. Uenishi, K. Wakasa, S. Suehiro, Lung Cancer 58 (2007) 369–375.
[24] J. Schneider, Adv. Clin. Chem. 42 (2006) 1–41.
[25] D. Shitrit, B. Zingerman, A.B. Shitrit, D. Shlomi, M.K. Kramer, Oncologist 10 (2005) 501–507.
[26] T. Okamoto, T. Nakamura, J. Ikeda, R. Maruyama, F. Shoji, T. Miyake, H. Wataya, Y. Ichinose, Eur. J. Cancer 41 (2005) 1286–1290.
[27] F. Barlési, C. Gimenez, J.P. Torre, Ch. Doddoli, J. Mancini, L. Greillier, F. Roux, J.P. Kleisbauer, Respir. Med. 98 (2004) 357–362.
[28] C. Fuhrman, J.C. Duche, C. Chouaid, I. Abd Alsamad, K. Atassi, I. Monnet, J.P. Tillement, B. Housset, Clin. Biochem. 33 (2000) 405–410.
[29] J.M. Porcel, M. Vives, A. Esquerda, A. Salud, B. Pérez, F. Rodríguez-Panadero, Chest 126 (2004) 1757–1763.
[30] B. Balla, J. Mocak, H. Pivovarnikova, J. Balla, Chemom. Intell. Lab. Syst. 72 (2004) 259–267.
[31] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 2002.
[32] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, Wiley, New York, 2001.
[33] R. Khattree, D.N. Naik, Multivariate Data Reduction and Discrimination, SAS Institute, Cary, North Carolina, 2000.
[34] D.G. Kleinbaum, M. Klein, E.R. Pryor, Logistic Regression, Springer, Heidelberg, 2005.
[35] J. Zupan, J. Gasteiger, Neural Networks for Chemists: An Introduction, VCH, Weinheim, 1993.
[36] V. Kairisto, A. Poola, Scand. J. Clin. Lab. Invest. 222 (1995) 43–60.
[37] D.W. Stockburger, Introduction to Statistics: Concepts, Models, and Applications, Atomic Dog Publishing, WWW Version 1.0, 1996. Critical values for the correlation coefficient, Web interactive calculation in: http://www.psychstat.missouristate.edu/introbook/rdist.htm, 2009.
[38] C.D. Collins, S. Purohit, R.H. Podolsky, H.S. Zhao, D. Schatz, S.E. Eckenrode, P. Yanga, D. Hopkins, A. Muir, M. Hoffman, R.A. McIndoe, M. Rewers, J.X. She, Vasc. Pharmacol. 45 (2006) 258–267.
[39] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, in: B.G.M. Vandeginste, S.C. Rutan (Eds.), Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998, pp. 207–239.
[40] AOAC (Association of Official Analytical Chemists): definitions and calculations proposed for method performance parameters, 6–12 March 1995, The Referee.
[41] S.J. Parekattil, H.A. Fisher, B.A. Kogan, J. Urol. 169 (2003) 917–920.
[42] V. Kairisto, A. Poola, Software Manual to GraphROC for Windows 2.0, Turku–Tallinn 1994–1996, http://www.netti.fi/~maxiw, 2009.