COMPUTER-AIDED DIAGNOSIS OF LUNG MALIGNITY USING MULTIDIMENSIONAL ANALYSIS OF TUMOUR MARKER DATA

VIERA MRÁZOVÁ[1], JÁN MOCÁK[1], ELENA VARMUSOVÁ[2], DENISA KAVKOVÁ[2]

[1]Department of Chemistry, University of SS. Cyril and Methodius, J. Herdu 2, Trnava, SK-917 01, Slovak Republic (viera.mrazova@ucm.sk)
[2]Institute for Tuberculosis and Respiratory Diseases, Department of Clinical Chemistry, Kvetnica, Poprad, SK-058 87 Slovak Republic

Abstract: The aim of this work is assessing diagnostic performance of lung tumour markers. Three clinical laboratory tests were used for indicating lung malignancy in order to verify or predict the patient's diagnosis. The data set of 182 patients was examined and two main groups of the patient samples were created – 86 with diagnosed malignancy (confirmed by histology) and 96 with diagnosed benign tumours or tuberculosis. The following tumour markers were analyzed: *carcinoembryonic antigen* and *cytokeratin 19 fragment*, which were sampled in the pleural exudates, and the same tumour markers in serum. In addition, the patient's age and the gender of the corresponding individual were used as further variables in the original data matrix. Three laboratory tests were used for indicating lung malignancy in order to verify or predict the patient's diagnosis not only by using the results of the chosen individual laboratory test but also applying multivariate statistical approach, which jointly utilizes all performed tests in the form of their optimal linear combination.

Key words: lung malignity, multidimensional analysis, tumour marker.

1. Introduction

Diagnosis of any disease can be confirmed or predicted not only using appropriate laboratory tests but also using multivariate statistical analysis, which uses simultaneously all performed tests in the form of their optimal (usually linear) combination. This new way of enhancing the diagnostic effectiveness, advocated by us, is applied here to the results of laboratory analysis of lung tumour markers in serum as well as pleural effusion (exudate).

Pleural effusion is common for several kinds of lung illnesses in clinical practice. Malignancy is one of the main causes of pleural effusion. 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 these non-invasive techniques is considered to be only 40 %–70 %. To improve upon these rates, a number of tumour markers (TM) in the pleural fluid have been intensively evaluated. The most common markers found to be of diagnostic significance were carcinoembryonic antigen (CEA), cancer antigen 15-3 (CA), cancer antigen 19-9, and cytokeratin 19 fragment (CYFRA 21-1).

CEA was identified in 1965 and has been widely used during the follow up of various tumours i.e.: colorectal cancers (MROCZKO *et al.*, 2007; DUFFY *et al.*, 2007;

YAMAMOTO *et al.*, 2005), breast cancer (NICOLINI *et al.*, 2008; CHEN *et al.*, 2006; SÖLÉTORMOS *et al.*, 2004). CYFRA 2l-1 assay measures cytokeratin 19 fragment and his concentration is increased with the extent of the malignant disease in non-small cell lung cancer. The serum CYFRA 21-1 distribution differs significantly according to histology, disease stage and performance status.

Lung cancer is the leading cause of cancer deaths in Europe. Levels of tumour markers, especially CEA and CYFRA 21-1, may help to establish the diagnosis of pleural malignancy (MATSUOKA *et al.*, 2007; SHITRIT *et al.*, 2005; OKAMOTO *et al.*, 2005; FUHRMAN *et al.*, 2000).

It should be added that the most effective positive test is histology of the appropriate tissue sample but this way is invasive and takes a long time. Therefore the use of TM may prevent the loss of time necessary for medical treatment in urgent cases.

2. Material and Methods

2.1 Description of the studied data

Tumour markers were determined at the Institute for Tuberculosis and Respiratory Diseases (ITRD) in Poprad - Kvetnica, Slovakia. The data set of 182 patients was examined; two main groups of the patient samples were created – 86 malignant (with malignancy confirmed by histology) and 96 benign tumours or tuberculosis. The following tumour markers were analyzed in *pleural effusion*: CEA (coded as EXCEA), CYFRA 21-1 (EXCYF) as well as in *serum:* (SCEA and SCYF, respectively). In addition, the patient's age (coded as AGE) and the gender of the corresponding individual (coded as SEXN) were used as the variables in the original data matrix. When using classification multivariate statistical techniques, two values of the categorical classification variable DG (diagnosis) were used: 1 - indicating malignant diseases and 2 - for others. This categorization was made on the basis of known histology results.

2.2 Multidimensional data analysis

Statistical calculations were performed using the principal components analysis (PCA), cluster analysis (CA), the linear discriminant analysis (LDA), and logistic regression (LR). Several software commercial packages were used: STAGRAPHICS Plus 5.1, SPSS 15 and JMP 6.0.2.

2.3 Analytical procedures

Tumour markers were analysed by automatic analysers ELECSYS 1010 and ELECSYS 2010, which use immunoanalysis with electrochemically generated chemiluminiscent detection. For determinations in pleural effusion an original procedure was used developed at the ITRD.

3. Results and Discussion

3.1 Principal Component Analysis (PCA)

The data set of the patients characterized by four variables, namely SCEA, EXCEA, EXCYF and SCYF was used for a preliminary study. In the way, described in detail in the part 3.1 it was found that the variable SCYF is of the least importance. Considering this as well as economical aspects the variable SCYF was exclude from further studies. Instead, the variable AGE, always accessible, was used in the detailed PCA study.

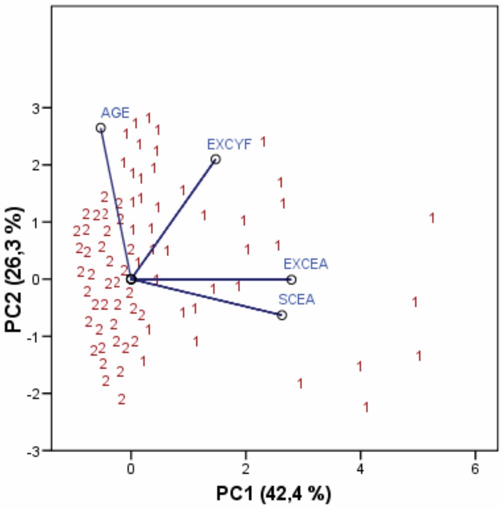

Fig. 1. PCA biplot showing 4 selected variables (EXCYF, EXCEA, SCEA, AGE) and 182 objects – patient's samples. Software SPSS 15.

PCA reveals a natural grouping of the studied objects as well as the used variables in a reduced dimensional space (Fig. 1). The first principal component (PC1) performs a linear combination of four original variables, optimized with respect to preserving maximal variance of the 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). The numbers 1 and 2 represent the category where the investigated sample belongs (1 – malignant, 2 – non-malignant). The inspection of the biplot depicted in Fig. 1 reveals that the PC1 axis represents *malignancy*. All tumour markers are positively correlated with the PC1, which is the proof that all patient samples with a high PC1 value are malignant and is in accordance with the observation that the malignant cases are located at high PC1 values.

3.2 Cluster Analysis (CA)

Among the clustering techniques, *Ward´s method* with Squared Euclidean distance metrics was selected for variable clustering. The obtained results are in agreement with clinical expectations: EXCEA and SCEA are clustered with EXCYF so that all tumour markers indicating positive diagnosis result are together. Variable AGE is clustered with SEXN since it simply reflects the fact that the average age of women is higher than that of men.

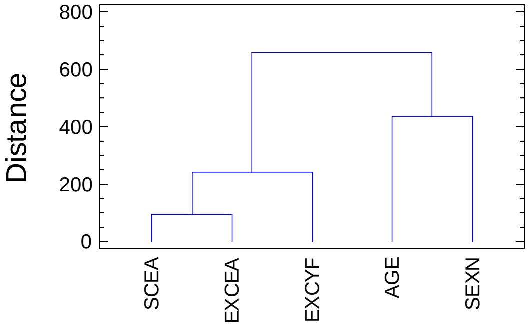

Fig. 2. Cluster analysis of variables using Ward`s method, Squared Euclidean. 182 patient samples with lung diseases. Software Statgraphics 5.1.

3.3 Discriminant Analysis

With regard to the solved problem the main goals of the applied classification multivariate methods, namely the linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression (LR) is: (1) to create diagnostic categories and the *training data set* using the entries of the individual samples with known diagnosis, (2) to elaborate a classification model using the categorized patient samples in the training set, (3) to perform the categorization of the not yet classified samples (belonging to the *test set* of data) into the selected classes.

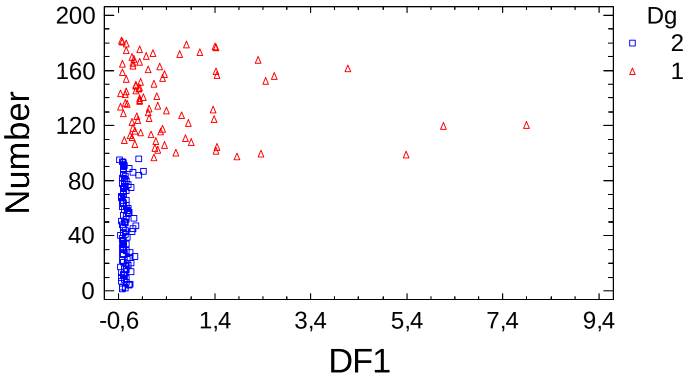

Fig. 3. Linear discriminant analysis of the samples denoted by the numbers on the vertical axis. DF1 denotes the only discriminant function. 86 samples corresponding to malignancy confirmed by histology (Dg=1) and 96 samples regarding benign tumours or tuberculosis (Dg=2). Software Statgraphics 5.1.

Figure 3 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 tailing cluster at higher DF1 values, which is a typical behaviour in many clinical studies.

The classification performance for different software packages providing the classification outputs is collected in Table 1. The exhibited results of the classification performance regard three types of samples: (1) the training set samples (used for calculating the classification model), (2) the samples omitted from the training set in a step-by-step manner according to the leave-one-out procedure (which was applicable only using the SPSS software), (3) the samples creating a special test set, which were not included into the training set.

Tab. 1. Classification results for various multivariate methods and software.

Classification method		**SPSS**			**JMP**		**Statgraphics**	
		Training set	Leave-1-out	Test set	Training set	Test set	Training set	Test set
LDA	true/all	139/182	137/182	26/30	139/182	26/30	139/182	26/30
	% true	76.4	75.3	86.7	76.4	86.7	76.4	86.7
QDA	true/all	150/182	N/A	25/30	N/A	N/A	N/A	N/A
	% true	82.4	–	83.3	–	–	–	–
LR	true/all	163/182	N/A	27/30	163/182	27/30	N/A	N/A
	% true	89.6	–	90.0	89.6	90.0	–	–

Note: The number of the patient samples is given by denominator in the "true/all" ratio. Decision upon malignity was predicted using four variables SCEA, EXCEA, EXCYF, AGE except LR where also SEXN was used. N/A means that the calculation was not possible when using the cited software.

The predictive ability of the used multivariate methods is expressed by the results referring to the last two types of the samples. It is better for the LDA (over 86 %) than the QDA. However, the best results (90 %) were achieved by logistic regression, where in addition to the variables used in other techniques, the patient's gender (woman/man) was used (in the form of the binary variable SEXN).

4. Conclusions

Principal component analysis and cluster analysis allow display a natural grouping of the samples belonging to the individuals treated for lung diseases. The obtained results demonstrate very good applicability of the used multivariate statistical methods for graphical representation and the samples classification in a reduced number of dimensions. Patient's diagnosis may be predicted or verified not only using the results of the selected 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.

References

CHEN, CH.CH., HOU, M.F., WANG, J.Y., CHANG, T.W., LAI, D.Y., CHEN, Y.F., HUNG, S.Y., LIN, S.R.: Simultaneous detection of multiple mRNA markers CK19, CEA, c-MeT, Her2/neu and hMAM with membrane array, an innovative technique with a great potential for breast cancer diagnosis. Cancer Lett., 240, 2006, 279-288.

DUFFY, M.J., VAN DALEN, A., HAGLUND, C., HANSSON, L., HOLINSKI-FEDER, E., KLAPDOR, R., LAMERZ, R., PELTOMAKI, P., STURGEON, C., TOPOLCAN, O.: Tumour markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines for clinical use. Eur. J. Cancer, 43, 2007, 1348-1360.

FUHRMAN, C., DUCHE, J.C., CHOUAID, C., ABD ALSAMAD, I., ATASSI, K., MONNET, I., TILLEMENT, J.P., HOUSSET, B.: Use of tumor markers for differential diagnosis of mesothelioma and secondary pleural malignancies. Clin. Biochem., 33, 2000, 405-410.

MATSUOKA, K., SUMITOMO, S., NAKASHIMA, N., NAKAJIMA, D., MISAKI, N.: Prognostic value of carcinoembryonic antigen and CYFRA 21-1 in patients with pathological stage I non-small cell lung cancer. Eur. J. Cardio-Thorac., 32, 2007, 435-439.

MROCZKO, B., GROBLEWSKA, M., WERESZCZYNSKA-SIEMIATKOWSKA, U., OKULCZYK, B., KEDRA, B., LASZEWICZ, W., DABROWSKI, A., SZMITKOWSKI, M.: Serum macrophage-colony stimulating factor levels in colorectal cancer patients correlate with lymph node metastasis and poor prognosis. Clin. Chim. Acta, 380, 2007, 208-212.

NICOLINI, A., CARPI, A., FERRARI, P., ROSSI, G.: Immunotherapy prolongs the serum CEA-TPA-CA 15.3 lead time at the metastatic progression in endocrine-dependent breast cancer patients: A retrospective longitudinal study. Cancer Lett., 263, 2008, 122-129.

OKAMOTO, T., NAKAMURA, T., IKEDA, J., MARUYAMA, R., SHOJI, F., MIYAKE, T., WATAYA, H., ICHINOSE, Y.: Serum carcinoembryonic antigen as a predictive marker for sensitivity to gefitinibhis in advanced non-small cell lung cancer. Eur. J. Cancer, 41, 2005, 1286-1290.

SHITRIT, D., ZINGERMAN, B., SHITRIT, A. B., SHLOMI, D., KRAMER, M. K.: Diagnosis value of CYFRA 21-1, CEA, CA 19-9, CA 15-3, and CA 125 assay in pleural effusions: Analysis of 116 cases and review of the literature. Oncologist, 10, 2005, 501-207.

SÖLÉTORMOS, G., NIELSEN, D., SCHIOLER, V., MOURIDSEN, H., DOMBERNOWSKY, P.: Monitoring different stages of breast cancer using tumour markers CA 15-3, CEA and TPA. Eur. J. Cancer, 40, 2004, 481-486.

YAMAMOTO, Y., HIRAKAWA, E., MORI, S., HAMADA, Y., KAWAGUCHI, N., MATSUURA, N.: Cleavage of carcinoembryonic antigen indeces metastatic potential in colorectal carcinoma. Biochem. Bioph. Res. Commun., 333, 2005, 223-229.