*Journal of Theoretical Medicine*, Vol. 2, pp. 317–327
Reprints available directly from the publisher
Photocopying permitted by license only

# The Prognosis of Survivance in Solid Tumor Patients Based on Optimal Partitions of Immunological Parameters Ranges

A. V. KUZNETSOVA[a,*], O. V. SEN'KO[b,†], G. N. MATCHAK[c], V. V. VAKHOTSKY[c], T. N. ZABOTINA[c] and O. V. KOROTKOVA[c]

[a]*Laboratory of Mathematical Immunobiophysics Institute of Biochemical Physics of Russian Academy of Sciences Kosygin str., 4, bld. 8, Moscow, 117977, Russia;* [b]*Computer Center of Russian Academy of Sciences Vavilova str., 40, Moscow, 117964, Russia;* [c]*Russian Cancer Research Center, Kashirskoye sh., 24, Moscow, 115478, Russia*

*(Received 22 November 1998; In final form 9 August 1999)*

New logical and statistical methods are used for the analysis of relationships between survivance and immunological variables. These methods are based on the search of the regularities (syndromes) in the multidimensional space. The syndromes are the elements of partitions of allowable areas of variables. To estimate the statistical validity of found regularities the new technique based on Monte-Carlo computer simulation was used.

We present some results from immunological research to illustrate the methods of logistical regularities search. Two tasks are described. The broad panel of monoclonal antibodies for differentiation lymphocytic antigens were used for lymphocytes subpopulations analysis. The purpose of the first task was the evaluation of significance of immunological parameters for prediction of 1-year metastasis-free survival in non-metastatic osteosarcoma of extremities. The second task was the construction of the predicting alghorithm for prognosis 2-years survival of patients with stomach cancer. The optimal sets of parameters for prediction of survivance was found for both tasks. We found out the high forecasting informativity of HLA-DR$^+$ cells percentage in the 1$^{st}$ task, and the percentage of adhesion cells (CD50$^+$-lymphocytes) in the 2$^{nd}$ task. Multivariate forecasting alghorithms are developed.

*Keywords*: Survival prognosis, solid tumor, immunological research, validity

## INTRODUCTION

Mathematical models are now widely used for predicting the outcome of treatment in cancer. The most popular are the proportional hazards Cox model, logistic regression and neural networks. In many cases these methods allow to achieve quite good results (Jefferson, 1997, Soong 1985). On the other hand complex biological systems are characterized by great number of various factors and nonlinear associations and dependencies. Standard statistical linear regression methods are not always

*Corresponding Author: Fax: +7-095-137 41 01; E-mail: bioimath@sky.chph.ras.ru
†E-mail: senko@ccas.ru

effective for the analysis of complex biological systems (for example the immune system). The analysis is more difficult in the small groups of cases which are typical for rare diseases.

Thus development of new mathematical approaches which are suitable for the biomedical data analysis and for the solution of prognostic tasks may be useful.

Algorithms named by voting recognition methods were developed by Dmitriev and Zhuravlev (1966), Bongard (1967), Karp and Kunin (1971). The Statistically Weighted Syndromes (SWS) technique is the further development of voting methods that is based on statistical estimates (Ryazanov, 1990, Senko, 1993; Kuznetsov, 1995, 1996 a,b; Kuznetsova, 1995). The SWS technique includes the search of so called "syndromes" in multidimensional space of predicting variables and the use of voting procedures by the sets of syndromes for prognosis.

The SWS technique was already used for the solution of several prognostic tasks. It has been successfully used for the prognosis for chemiotherapy of osteosarcoma (Ivshina, 1995; Kuznetsov, 1995). The problem of prognosis of treatment of bladder cancer patients was successfully solved by SWS technique while logistic regression analysis failed to discriminate the groups of bad and good responders (Jackson, 1998). The validity of results was established with the help of the permutation test. The task of prediction of long-term responses to antiretroviral treatment was also solved (Mueller, 1998). The SWS technique was compared with regression trees technique and the SWS predictions appeared to be more stable. The SWS method allowed to reveal the significant differencies between the group of patients with Wilson's disease and the group of healthy donors by immunological parameters (Zhirnova, 1998).

In this study we have investigated the possibility to use this approach in the particular clinical situations. We had to construct the algorithm for the prediction of short time disease-free survival in osteosarcoma and stomach cancer patients. We also used the new method of statistical validation of regularities described by syndromes.

## MATERIALS AND METHODS

Pretreatment immunological features were analyzed in two groups of patients. The set of samples for the first task consisted of 80 patients with non-metastatic osteosarcoma: 55 patients with bad outcome and 25 patients without progression during one-year period. Forty seven patients with gastric cancer were selected for the second task. At two-year follow-up period 27 of 47 patients were classified as disease-free survivors and 20 patients had progressive disease.

Cell phenotypes of the patients were studied using the broad panel of monoclonal antibodies for the differentiation antigens of peripheral blood lumphocytes: CD3, CD4, CD8, CD5, CD7, HLA-DR, CD24, CD16, CD38, CD25, CD71, HLA-1, CD45, CD50, CD45RA, CD95. Measurements were performed by the indirect immunofluorescence assay in a Becton Dickinson flow cytometer FACS-scan and presented in percentage and absolute (cells/ml) forms. In addition G, A and M immunoglobulins serum levels (by Manchini) were determined.

### Multivariate Pattern Recognition Analysis

Pattern recognition methods were used for the classification (recognition) of objects by the set of meanings of predicting variables. The recognition algorithm includes the analysis of empirical data table (so-called procedure of tutoring at the training set). The basic principle of SWS method is the voting procedure of syndromes. The syndrome is the subarea in multi-dimensional variable space that contains the projections of cases. The syndromes are constructed with the help of partitioning of parameters values ranges. The partitions with the maximal value of quality functional are searched. The quality functional characterizes how the objects from different groups are separated with the help of partition. The search for the sets of the most informative predicting variables is performed. The stepwise procedure is used, which implies the gradual escalating of parameters number in a set by adding the variables that give the best improvement of recognition by SWS method. The cross validation technique is used for statistically valid evaluation of effective recognition coefficient (ERC) that is used as the measure of an exactitude of recognition (see Appendix).

### The Method of Statistical Validation of Syndromes Constructed by Partitioning

The analyses of syndromes that was used to construct the predicting SWS algorithms may give additional information about informativity of different variables and the types of dependencies. The vizualization by 2-dimensional diagrams simplifies the interpretation by medical experts. The syndromes in SWS method are constructed by partitioning of intervals of single variables. However sometimes the unidimensional procedure does not allow to find the optimal solution. So the additional method of partitioning of 2-dimensional areas of pairs of variables was developed. The main problem of partition approach is the statistical validation of results. The standard Xi-square method is suitable only when data set is large enough to form the two subsets (Kendall, 1967). One of these subsets is used to construct the optimal partition and another is used to estimate the validity that the found dependence really exists. The statistical significance of results may be over estimated when the same set is used for partition construction and for validation. Usually the medical databases include only about hundred cases. So the procedure using two sets is ineffective and some another technique is necessary. To estimate the statistical validity of revealed regularity the Monte-Carlo technique was used (Ermakov, 1975). The large number of random tables was simulated in accordance with the supposition that compared groups have the equal distributions. The size of each random table and true table must be the same. The partitioning result at the true table is compared with the partitioning results at the randomized tables. The statistical significance level of regularity based on some partition is defined as ratio of random tables that allow achieving the better separation of two groups than the separation achieved at true table.

## RESULTS

### The Task N 1

*The 1-year metastasis-free survival prediction in osteosarcoma patients using of immunological parameters.*

Classic statistical analysis (Student's t-criterion) does not indicate significant differences between osteosarcoma patients groups with or without early progression. However slightly more sophisticated technique based on forming of subgroups of patients with the help of threshold meanings of immunological variables have shown the existence of valid relationship between survivance and some immunological parameters. Thresholds are found with the help of partition constructed by one parameter with one border (the 1st partitions model, see Appendix). The factors affecting survival are represented in the Table I. We found that the most prognostic power has the percentage of HLA-DR-positive lymphocytes.

The statistical validity was estimated by log-rank test. It must be noted that significant diversity of survivance curves exists at the initial period of time and it diminishes to the end of observation period. So we decided to estimate also the statistical difference between curves at initial period when all 1-year survivors are considered alive with the censoring time 12 month. The results are presented in 3rd column of Table I.

The survival curves calculated by Kaplan-Mayer method for 1st group of 23 patients with HLA-DR <

TABLE I The Log-rank Test Estimated Difference of Survivance Rate between Subgroups of Osteosarcoma Patients Formed with the Help of Optimal Partitions

| Parameters | Partition boundaries | Log-rank Significance | |
|---|---|---|---|
| | | 1-year censoring | Full observation period |
| Lymphocyte (%) | 27.5 | 0.00467 | 0.0514 |
| CD3 (%) | 71.65 | 0.00109 | 0.00025 |
| HLA-DR (%) | 7.6 | 0.00015 | 0.00247 |
| HLA-DR (cells/ml) | 95.0 | 0.00031 | 0.00615 |
| HLA-1 (cells/ml) | 944.0 | 0.0127 | 0.165 |
| CD7 (cells/ml) | 867.0 | 0.075 | 0.114 |
| IgG (IU/l) | 145.0 | 0.0668 | 0.0676 |

7.6% (15 cases [65.2%] belong to patients with good outcome and 8 cases [34.8%] belong to patients with bad one) and the 2nd group of 55 patients with HLA-DR > 7.6% (10 cases [18.2%] and 45 cases [81.8%] respectively) are shown at Figure 1.

The multivariate analysis was performed using Statistical Weighted Syndromes method. A stepwise procedure was used to select the factors that give the best prediction of 1-year survivance. The set of selected factors includes HLA-DR-positive lymphocytes (%), HLA-DR-positive lymphocytes (cells/ml), IgG (IU/l), CD3-positive lymphocytes (%) and percentage of lymphocytes. The exactness of forecasting was estimated by cross validation method. The number of correctly predicted cases was 59 (74%), the number of mistakes was 18 (22%), and the number of rejects was 3 (4%).

The optimal partition constructed by HLA-DR$^+$- and percentage of lymphocytes with one border at each parameter (the 3rd model, see Appendix) which separates investigated groups is presented at the scatter plot (Figure 2). One can see set of observations with a predominance of one of the groups inside of each subarea.

The 2000 tables were generated to estimate the validity of regularity demonstrated at Figure 2. And only in two cases the functional meaning exceeded the value received at true table. So the level of validity estimated in such a way is about 0.001.

## The Task N 2

*The survival evaluation in stomach cancer patients with using of immunological parameters.*

The comparison of means and distributions of immunological parameters values in patients groups with stomach cancer (see Table II) demonstrated significant increase of CD50- and CD16- positive lymphocytes (%) in the group with good outcome. Using the nonparametric U-criterion (Wilcoxon-Mann-Whitney, WMW) we found out the significant differences between groups by only one parameter — platelets (cells/ml).

The optimal thresholds are found with the help of partition constructed by one parameter with one border (the 1st partitions model, see Appendix). The most powerful factor affecting survival appeared to be CD50-positive lymphocytes percentage.

The survival curves calculated by Kaplan-Mayer method for the 1st group of 17 patients with CD50-positive lymphocytes percentage less then 83.6% (6 cases [35.3%] of good outcome and 11 cases [64.7%] of bad one) and the 2nd group of 29 patients with CD50-positive lymphocytes percentage greater then 83.6% (20 cases [69%] and 9 cases [31%] respectively) are shown at Figure 3.

The statistical validity for percentage of CD50-positive lymphocytes estimated by log-rank test is 0.002.

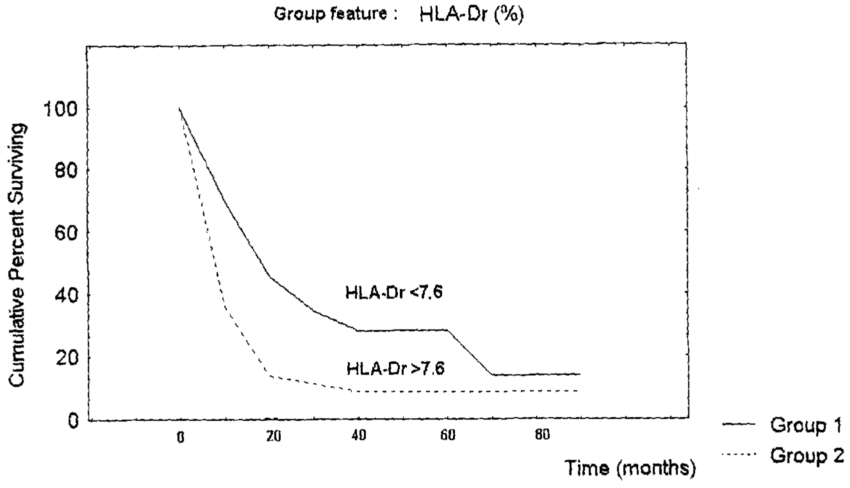

FIGURE 1 The survival curves calculated by Kaplan-Mayer method for the group of osteosarcoma patients with HLA-DR < 7.6% and the group of osteosarcoma patients with HLA-DR > 7.6%.

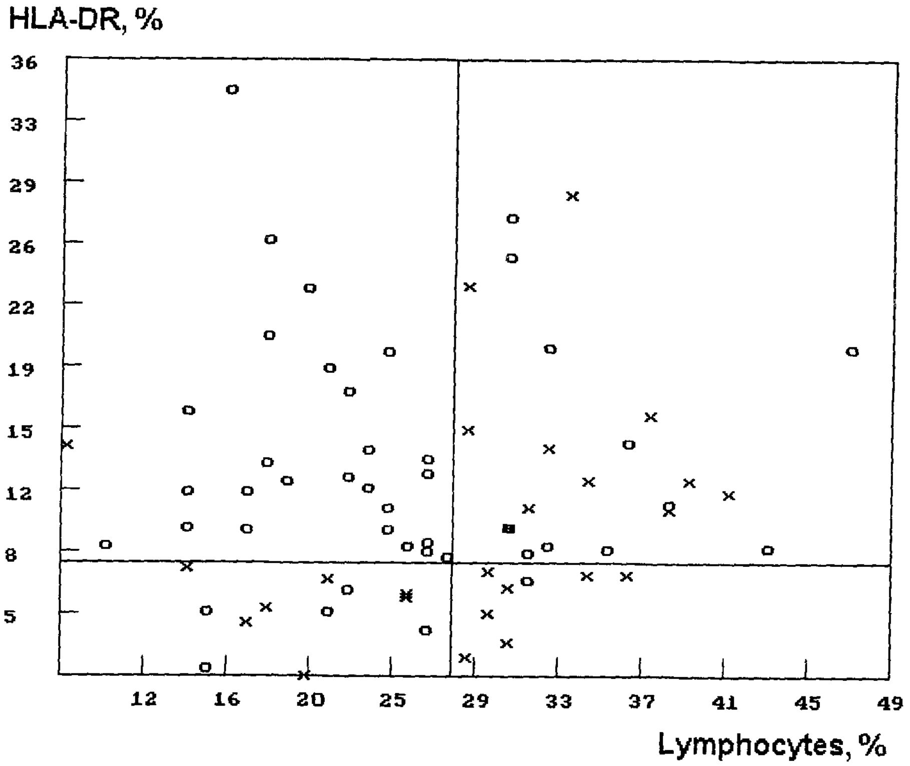

FIGURE 2 The optimal partition constructed by percentage of HLA-DR$^+$-lymphocytes and of common lymphocytes with one border at each parameter. The cases of the group with bad outcome are denoted "o", the cases of the group without earlier progression are denoted by "x".

TABLE II Statistical Estimation (M ± m) of Immunological Parameters of Patients with Different Stomach Cancer Outcome

| N | Parameter | Good outcome (n = 27) | Bad outcome (n = 20) | Criterion | |
|---|---|---|---|---|---|
| | | | | t- | WMW |
| 1 | Leukocytes | 7109 ± 721 | 5838 ± 434 | ns | ns ($p < 0.1$) |
| 2 | Monocyte,% | 8.2 ± 1.35 | 6.4 ± 1.09 | ns | ns |
| 3 | Lymphocyte,% | 21.8 ± 2.18 | 23.0 ± 2.16 | ns | ns |
| 4 | HLA-Dr,% | 12.2 ± 2.36 | 11.2 ± 1.98 | ns | ns |
| 5 | CD38,% | 24.0 ± 3.39 | 23.6 ± 2.65 | ns | ns |
| 6 | CD8,% | 21.2 ± 1.91 | 21.4 ± 2.1 | ns | ns |
| 7 | CD45,% | 87.5 ± 2.65 | 79.5 ± 6.02 | ns | ns |
| 8 | HLA-1,% | 89.3 ± 3.05 | 79.7 ± 5.94 | ns | ns ($p < 0.1$) |
| 9 | CD50,% | 85.4 ± 2.73 | 70.4 ± 6.1 | $p < 0.05$ | ns ($p < 0.1$) |
| 10 | CD45RA,% | 39.2 ± 4.94 | 46.3 ± 4.12 | ns | ns ($p < 0.1$) |
| 11 | CD5,% | 55.8 ± 3.06 | 55.1 ± 4.41 | ns | ns |
| 12 | CD4,% | 34.1 ± 2.23 | 29.6 ± 2.08 | ns | ns ($p < 0.1$) |
| 13 | CD7,% | 75.5 ± 2.8 | 78.2 ± 2.70 | ns | ns |
| 14 | CD71,% | 4.2 ± 1.59 | 7.2 ± 3.53 | ns | ns |
| 15 | CD3,% | 58.1 ± 3.07 | 61.3 ± 2.90 | ns | ns |
| 16 | CD22,% | 11.0 ± 4.84 | 9.1 ± 3.7 | ns | ns |
| 17 | CD25,% | 9.4 ± 3.38 | 8.3 ± 3.93 | ns | ns |
| 18 | CD16,% | 20.5 ± 3.68 | 10.4 ± 1.87 | $p < 0.05$ | ns ($p < 0.1$) |
| 19 | Platelets, cells/ml | 266 ± 17 | 316 ± 21 | ns ($p < 0.1$) | $p < 0.05$ |

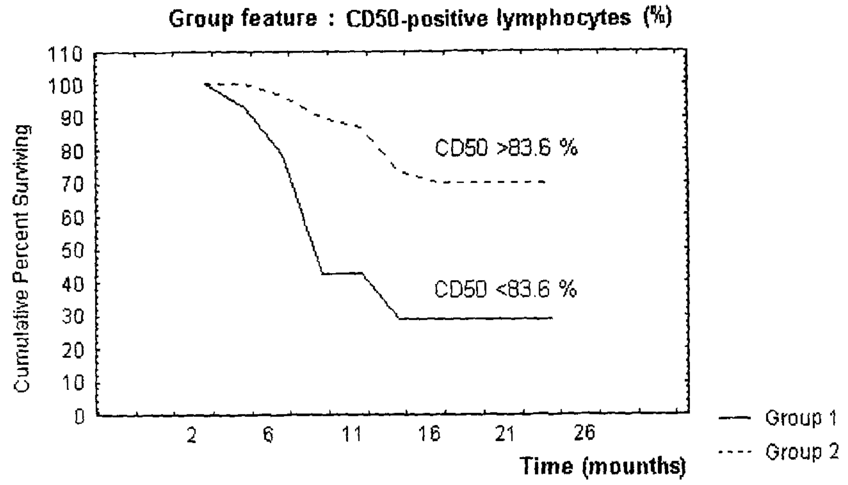

FIGURE 3 The survival curves calculated by Kaplan-Mayer method for the group of stomach cancer patients with CD50 > 83.6% and the group of stomach cancer patients with CD50 < 83.6%.

The optimal partitions constructed by pairs of parameters with one border at each parameter (the 3rd model, see Appendix) were searched. Rather good separation was achieved for the pairs presented at Table III. To validate the regularities based on found partitions the comparison with Monte Carlo generated random tables was implemented.

For example the optimal value of quality functional (see Appendix) at true table for pair of immunological parameters as percentage of CD45-positive and CD50-positive lymphocytes was 8.89. The value of quality functional was less than 8.89 in case of 1984 tables. So the statistical validity of regularity is estimated as 0.992 (significance level $p < 0.008$).

TABLE III The Monte Carlo Estimates of the Validity of the Regularities Based on Partitions Constructed by Pairs of Parameters

| 1st parameter and its boundary | | 2nd parameter and its boundary | | p |
|---|---|---|---|---|
| CD50 (%) | 51.8 | HLA-1 (%) | 88.6 | <0.05 |
| CD50 (%) | 83.6 | CD45 (%) | 96.2 | <0.008 |
| CD8 (%) | 24.5 | HLA-1 (%) | 95.4 | <0.02 |
| CD38 (cells/ml) | 406 | CD5 (cells/ml) | 1241 | <0.042 |
| CD8 (cells/ml) | 414 | CD4 (cells/ml) | 496 | <0.04 |
| CD5 (cells/ml) | 1247 | CD16 (cells/ml) | 266 | <0.03 |
| CD4 (cells/ml) | 476 | CD3 (cells/ml) | 474 | <0.035 |

In case of second pair (percentages of CD16 and CD5-positive cells) the optimal value of quality functional at true table was 7.96. The 2000 tables were generated using bootstrap technique. The value of quality functional received by stochastic tables was less than one at true table in case of 1946 tables. So we'll consider that the statistical validity of regularity is estimated as 0.97 (significance level $p < 0.03$).

The corresponding optimal partitions are shown on the scatter plots at Figures 4, 5. The patients from the group with bad outcome are denoted "o", the patients of the group without earlier progression are denoted by "x".

We have tried to solve the problem of discrimination of the two patients groups with the method of statistically weighed syndromes (SWS). We have used 19 parameters. Seven the most informative parameters were determined. They are (in an order of descending informativity) percentage of $CD50^+$-lymphocytes, plateletes (cells/ml), $CD16^+$-lymphocytes, leucocytes (cells/ml), percentage of $CD3^+$-, $CD45^+$-, HLA-$Dr^+$-lymphocytes. The effective recognition coefficient (ERC) (see Appendix) as coefficient of correlation between the group number prognosis and true numbers of group was 0.66. The correct prognosis is carried out in 79% of cases (22 valid prognoses in 27 patients with good outcome,

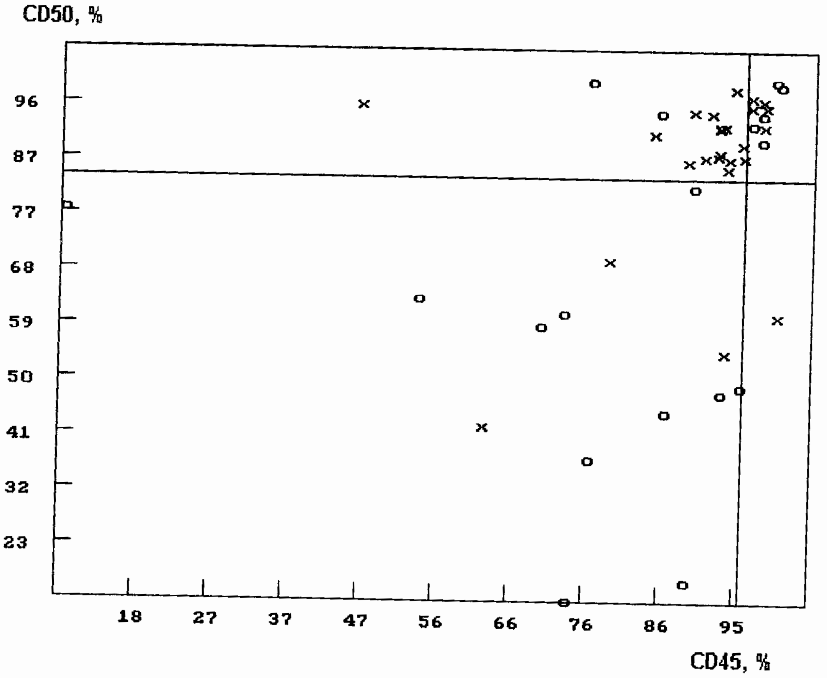

FIGURE 4 The optimal partition constructed by percentages of CD50$^+$- and CD45$^+$-cells with one border at each parameter. The patients from the group with bad outcome are denoted "o", the patients of the group without earlier progression are denoted by "x".

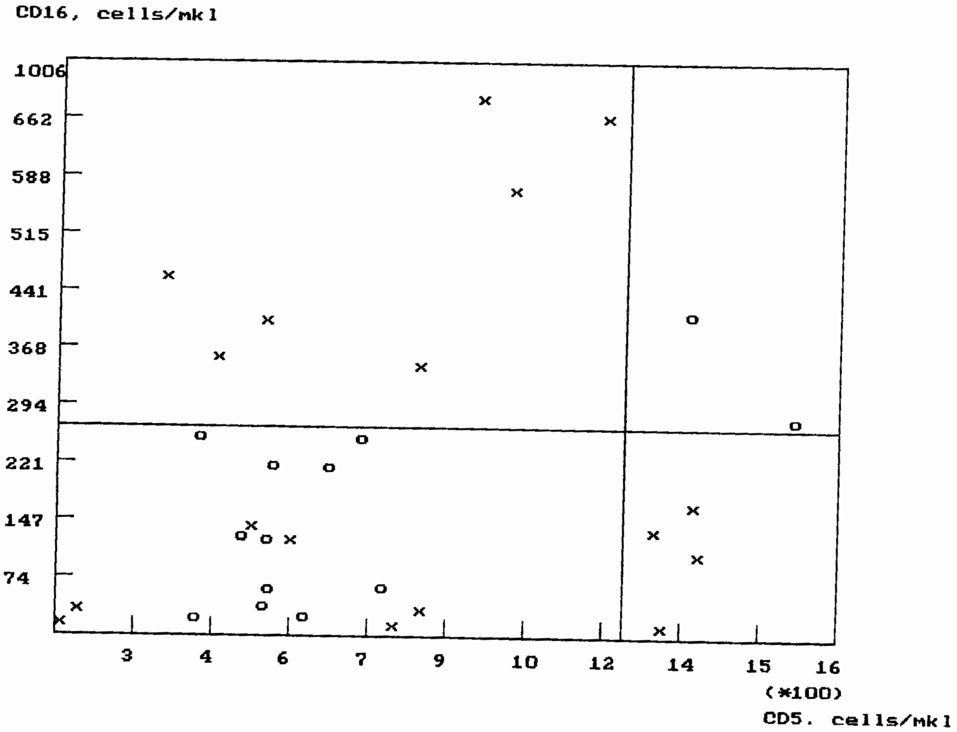

FIGURE 5 The optimal partition constructed by CD16$^+$- and CD5$^+$-cells (cells/ml) with one border at each parameter. The patients from the group with bad outcome are denoted "o", the patients of the group without earlier progression are denoted by "x".

15 valid prognoses in 20 patients with bad outcome). The number of mistakes was 9 (19%), the number of rejects 1 (2%).

## CONCLUSIONS

The discussed methods based on partitioning and voting procedures allow us to construct algorithms predicting 1-year survivance in osteosarcoma of extremities and 2 year survivance in stomach cancer by immunological prognostic factors. Our results allowed us to suppose that the count of HLA-DR$^+$-cells is the most informative parameter for prediction of 1-year survival in patients with osteosarcoma of extremities. The informativity of this parameter was also mentioned earlier in the works of Leskovar (1996), De Stefani (1996), Ishigami (1998). We ascertained the high forecasting informativity of adhesion cells (percentage of CD50-positive lymphocytes) in case for stomach cancer prediction. The log-rank test shows the strong difference between groups formed by percentage of HLA-DR$^+$-cells for the 1st task and percentage of CD50$^+$-cells for the 2nd task using thresholds. However it must be taken into account that thresholds are found by corresponding training sets. The same data set is used in each case for the calculation of log rank test statistics that have been used to calculate threshold. So the statistical significance of difference revealed by log rank test may be too high (Kendall, 1967).

Thus the new technique of the prognosis of survivance in solid tumor patients based on optimal partitions of immunological parameters ranges has been represented. We consider that our approach allow to reveal and describe dependencies in survivance evaluation. The main advantages of this approach are the possibility to ascertain the complex nonlinear regularities and simple visualized form of the analysis results representation. It can be used for analysis large variety of medical and biological investigation.

The developed program complex of algorithms can be recommended for the analysis of data in the cases of small groups, the large quantity of parameters and in presence of the missed data that is typical for medical and biological researches.

## APPENDIX

### The Method of Statistically Weighted Syndromes (SWS)

Suppose that variables $X_1,\ldots,X_n$ are used to describe patients that belong to 2 groups (classes) $K_1$ and $K_2$. Our goal is to construct by training information $\tilde{S}_0$ the algorithm recognizing if the patient belongs to class $K_1$ or $K_2$ not. The training information is the set of patients descriptions or $\tilde{S}_0 = \{(y_1,\bar{x}_1),\ldots,(y_m,\bar{x}_m)\}$. Pair $(y_i,\bar{x}_i)$ is the patient description with $y_i = 1$ ***if*** the patient belongs to class $K_1$ and $y_i = 0$ if the patient belongs to class otherwise, $\bar{x}_i$ is the vector of means of variables $X_1,\ldots,X_n$. In other words $y_1,\ldots,y_m$ may be considered as means of the indicator function $Y$ of class $K_1$.

1) At the first stage of training the set of "syndromes" is constructed in multidimensional space. The syndrome is defined as subarea in multidimensional space that contains descriptions presumably belonging to one of the classes. To construct the syndromes the partitioning of variables ranges is used. Suppose that the values of variable $X_i$ belong to interval $M_i$. The best partitions of $M_i$ from models **I** or **II** are searched using quality and stability functionals (*see below*). The partition with maximal value of quality functional $F_q$ and with the stability functional $F_s$ value not less than some fixed threshold is selected. The threshold is defined by the user. We usually used threshold from 0.7 to 0.95. In case when both models **I** and **II** not allow achieving the threshold we reject from further consideration of variable $X_i$. So the selection of variables takes place at the first stage of partitioning. Suppose that $r_{1i},\ldots,r_{qi}$ are elements of optimal partition of $M_i$. We shall say that syndrome $Q_{ji}$ is supported by element of partition $r_{ji}$ if it includes those and only those patients descriptions with $X_i$ belonging to $r_{ji}$.

The syndromes system $\tilde{Q}$ constructed at the first stage consists from all syndromes supported by elements of optimal partitions and all possible their intersections.

2) Suppose that we want to classify patient with the vector of predicting variables $\overline{x}$ belonging to syndromes $Q_1, \ldots, Q_l$ from $\tilde{Q}$. The voting procedure is used to calculate the estimate $\Gamma(\overline{x})$ of conditional probability $P(K_1|\overline{x}) : \Gamma(\overline{x}) = \frac{\sum_{i=1}^{l} wei_i v_i}{\sum_{i=1}^{l} wei_i}$, where $v_i = m_i^1/m_i$, $m_i$ is the full number of descriptions from $\tilde{S}_0$ in $Q_i$ and $m_i^1$ is the number of descriptions from $K_1$ among these descriptions. Parameter $wei_i$ is so called "weight" of syndrome $Q_i$ that is calculated as $wei_i = \frac{m_i}{m_i + 1} \frac{1}{D_i}$ where $D_i$ is dispersion of function $Y$ in $Q_i$ estimated as $D_i = v_i(1 - v_i)$. To classify patient you must compare the estimate $\Gamma(\overline{x})$ with two thresholds $d_1$ and $d_2$ calculated by training information $\tilde{S}_0$. If $\Gamma(\overline{x}) > d_1$ the patient is put to class $K_1$, if $\Gamma(\overline{x}) < d_2$ the patient is put to class $K_2$ and if $d_1 \geq \Gamma(\overline{x}) \geq d_2$ the reject from recognition takes place. To estimate the validity of recognition the cross validation mode can be used can be shortly described as follows. One of the object is removed from training table. The tutoring is carried out by the rest objects. Then the removed object is classified with the received solving rule. The procedure is repeated with every object from the training table. The effective recognition coefficient (ERC) is defined as the correlation coefficient between the true number of the class and estimates calculated by voting procedure (1). The more the ERC value the better are the recognition results.

3) The search for the sets of the optimal predicting variables is the important part of training procedure. The optimal set allows to achieve the possibly maximal exactness of recognition of the objects from compared groups. The stepwise procedure is used, which implies the gradual escalating of parameters number in a set by adding the parameters that give the best improvement of recognition by SWS method.

## The Models of Partitions

*The partition model is defined as the set of partitions with the number of elements less than some fixed number, which are built by the same apriori defined and fixed algorithm.* In this paper we used the models with the partitions formed by the boundaries parallel to coordinate axes are represented (Senko, 1998). The 1st model (Figure 6,a) includes all partitions with the number of elements less 3, which are constructed by single parameter. The 2nd model (Figure 6,b) includes all partitions with the number of elements less 4, which are made by single parameter. The 3rd model (Figure 6,c) includes all partitions with the number of elements less 5, which are constructed by the pair of parameters with the number of borders at each parameter less than 2.

## The Quality of Partitions

Suppose that we want to construct the optimal partition of allowable interval of predicting variable $X_i$. Suppose that some partition $R$ consists of elements $q_1, \ldots, q_p$, separated by boundary points $\alpha_1, \ldots, \alpha_{s-1}$. In case we use the model I s = 2 and in case of model II s = 3.

1) Let $m^1$ be the number of objects from class $K_1$ in $\tilde{S}_0$, $v_0$ is defined as ratio $m^1/m$. Let $m_j$ be the

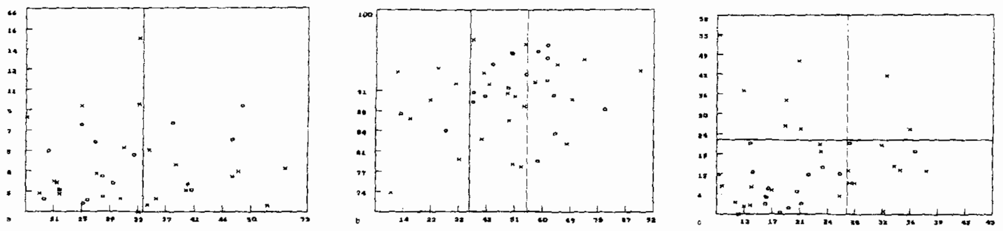

FIGURE 6 The example of partitions: a) model I, b) model II, c) model III.

number of objects in $\tilde{S}_0$ with variable $X_i$ meanings belonging to element $q_j$ and $m_j^1$ be the number of objects in $\tilde{S}_0$ with variable $X_i$ meanings belonging to element $q_j$ from class $K_1$. Then the $v_j$ is defined as ratio $m_j^1/m_j$.

The quality functional $F_q(R, \tilde{S}_0) = \sum_{j=1}^{s}[(v_j - v_0)^2 m_j]/D$ where $D = v_0(1 - v_0)$. The quality functional is defined not only for allowable intervals of single variables but also for multidimensional allowable areas. For example it can be used to estimate the quality of 2-dimensional diagrams partitions (models III).

2) The stability functional $F_s(R, \tilde{S}_0)$ is the measure of stability of boundaries $\alpha_1, \ldots, \alpha_{s-1}$ in case of small changes in training information. The $F_s(R, \tilde{S}_0)$ is calculated in cross validation mode. Let $\alpha_1^j, \ldots, \alpha_{s-1}^j$ are the boundaries calculated by $\tilde{S}_0$ without description of patient $s_j$. Let $D^i$ be the dispersion of variable $X_i$. The stability functional $F_s(R, \tilde{S}_0)$ is calculated as

$$F_s(R, \tilde{S}_0) = 1 - \frac{\sum_{i=1}^{s-1}\sum_{j=1}^{m}(a_i^j - \overline{a}_i)^2}{D^i m(s-1)},$$

where $\overline{a}_i$-is the mean value of all boundaries calculated by $\tilde{S}_0$ in cross validation mode.

## THE MONTE CARLO PROCEDURE OF PARTITION BASED REGULARITY VALIDATION

The coefficient $\alpha(\tilde{R}_k, \tilde{S}_0)$ we use as the measure of statistical validity of partition based regularity. Suppose that we have a process that generates random tables that coincide by the size with the true training table according with zero hypothesis $H_0$. The zero hypothesis suggests that analyzed variables have the same distributions for both classes $K_1$ and $K_2$. The coefficient $\alpha(\tilde{R}_k, \tilde{S}_0)$ is defined as the probability of casual arising of data set $\tilde{S}$ which does not allow to find the corresponding optimal partition $R(\tilde{S})$ from model $\tilde{R}_k$ with the meaning of quality functional $F_q(R(\tilde{S}), \tilde{S})$ exceeding the value $F_q(R(\tilde{S}_0), \tilde{S}_0)$.

To estimate the $\alpha(\tilde{R}_k, \tilde{S}_0)$ defined above the Monte-Carlo method is used (Ermakov, 1975). The "data sets" are constructed with the help of random numbers generator. Such tables will be further referred to as random tables. Then the procedure similar to bootstrap procedures is used (Efron, 1979). The two independent casual selections with replacement of the length $m$ are made from set $I_m = \{1, \ldots, m\}$. Suppose that selections $\{l_1, \ldots, l_m\}$ and $\{f_1, \ldots, f_m\}$ are generated. Then the random table corresponding to these selections is the set of pairs $\{(y_{l_1} \overline{x}_{f_1}), \ldots, (y_{l_m}, \overline{x}_{f_m})\}$.

Suppose that $2N$ independent selections with returns are made from $I_n$ and random tables set $\tilde{\mathrm{S}}_{\mathrm{R}} = \{\tilde{S}_1^r, \ldots, \tilde{S}_N^r\}$ is constructed. The $\alpha(\tilde{R}_k, \tilde{S}_0)$ is estimated as the ratio of random tables from $\tilde{\mathrm{S}}_{\mathrm{R}}$ with $F_q(R(\tilde{S}_*^r)) < F_q(R(\tilde{S}_0))$ and the full number of generated tables. Such approach of course gives the maximal possible similarity between the empirical distributions and measure P. It may be used in case when the number of parameters is little or we are interested in investigation only several pairs of parameters. In case when the parameters number is several tens as it is usually in medical research the approach demand great amount of calculations. Really, the random tables set $\tilde{\mathrm{S}}_{\mathrm{R}}$ must be a new constructed for the each pair of parameters. So we suppose to use the single set $\tilde{\mathrm{S}}_{\mathrm{R}}^{\mathrm{u}}$ common for all pairs of parameters. The set $\tilde{\mathrm{S}}_{\mathrm{R}}^{\mathrm{u}}$ is constructed using uniform and mutually independent distributions of ≪prognozing≫ and empirical distribution of prognozed value concentrated in points $\{y_1, \ldots, y_m\}$. Such approach allows to diminish very significantly the amount of calculations but it also can corrupt the estimates of regularities validity in case when real parameters distribution significantly differs from uniform. Some corrections must be made in these cases.

### *Acknowledgments*

We would like to thank prof. V. A. Kuznetsov for very useful discussions of concerning the application of "syndrome" approach in immunology, O. Yu. Rebrova (MD) and L. P. Semenova (PhD) for preparing of this paper for publication. The work is supported by grant of young scientists (N5, Russian Academy of Sciences presidium resolution from 13.07.1998) and grant RFFI N 99-07-90120.

## *References*

Bongard, M. M. (1967). The Recognition Problem, Moscow; Nauka (in russian).

Brieman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). Classification and Regression Trees, Wadswworth, Belmont, CA.

Cheredeev, A. N. and Kovalchuk, L. V. (1997). Pathogenetic principle of immune system evaluation in human: positive and negative activation. *Russian Journal of Immunology*, **2**(2), 85–90.

Chou, F. F., Sheen Chen, S. M., Chen, Y. S. *et al.* (1997). Surgical treatment of cholangiocarcinoma. *Hepatogastroenterology*, **44**(15), 760–765.

De Stefani, A., Valente, G., Forni, G. *et al.* (1996). Treatment of oral cavity and oropharynx squamous cell carcinoma with perilymphatic interleukin-2: clinical and pathologic correlations. *Journal of Immunother Emphasis Tumor Immunology*, **19**(2), 125–133.

Dmitriev, A. N., Zhuravlev Yu, I. and Krendelev, F. P. (1966). About mathematical principles of objects and phenomena classification. *// Discreet Analysis, Institute of Mathematics of RAS, Novosibirsk*, **7** 3–15 (in russian).

Efron, B. (1979). *Bootstrap Methods*, The Annals of Statistics, **7**(1), 1–26.

Ermakov, S. M. (1975). *Metod Monte-Carlo I Smezhnye Voprosy*, Moscow (in russian).

Ishigami, S., Aikou, T., Natsugoe, S. *et al.* (1998). Prognostic value of HLA-DR expression and dendritic cell infiltration in gastric cancer. *Oncology*, January, **55**(1), 65–69.

Ivshina, A. V., Kuznetsov, V. A., Kuznetsova, A. V. and Senko, O. V. (1995). Lymphocyte phenotype in the blood for estimation of tumor volume and their vascularisation in patients with osteosarcoma. *Immunologia*, **6**, 56- (in russian).

Jackson, A. M., Ivshina, A. V., Senko, O. V., Kuznetsova, A. V. *et al.* (1998). Prognosis of Intravesical Bacillus Calmette-Guerin Therapy for Superficial Bladder Cancer by Immunological Urinary Measurements: Statistically Weighted Syndromes Analysis. *Journal of Urology*, **159**(3), 1054–1063.

Jefferson, M. F., Pendleton, N., Lucas, S. B. and Horan, M. A. (1997). Comparison of a Genetic Algorithm Neural Network with Logistic Regression for Predicting Outcome after Surgery for Patients with Nonsmall Cell Lung Carcinoma. *Cancer*, **79**, April 1, 1338–1342.

Karp, B. P. and Kunin, P. E. (1971). Construction of decisive rule for solution of the problem of alternative diagnostics by the conjunction selection and directed training method. *Automation: Organization, Diagnostics*, Moscow: Nauka, 339 (in russian).

Kendall, M. G. and Stuart, A. (1967). *The Advanced Theory of Statistics*, **2**, London.

Kuznetsov, V. A., Ivshina, A. V., Kuznetsova, A. V. and Senko, O. V. (1995). Analysis of lymphocytes phenotype in the blood for prediction of metastases in patients with osteosarcoma. *Immunologia*, **5**, 52–58 (in russian).

Kuznetsov, V. A., Ivshina, A. V., Senko, O. V. and Kuznetsova, A. V. (1996 a). Syndrome approach for computer recognition of fuzzy systems and its application to immunological diagnostics and prognosis of human cancer. *Mathematical Computer Modelling*, **23**(6), 95–119.

Kuznetsov, V. A., Senko, O. V., Kuznetsova, A. V. *et al.* (1996 b). Recognition of the fuzzy systems by the statistical weighted syndromes method and its application to immune-hematology characteristics of a normal and of the chronic pathology. *Chemical Physics*, **15**(1), 81–100 (in russian).

Kuznetsova, A. V. (1995). Diagnostics and prediction of tumor growth on immunological datas with the help of the syndrome recognition methods. *Ph.D. Thesis*, Moscow, 23 (in russian).

Leskovar, P. and Bielmeier, J. (1996). Treatment of solid tumors should obligatorily be combined with the in vivo codepletion of tumor-protecting, CD8+/HLA-DR(+)-suppressor T cells by alloreactive donor T cells whose preprogrammed cell death allows a high GvL-effect before GvHD can be established. Results of animal experiments, including more than 6000 mice. *Pflugers Archiv*, **431**(6), Suppl 2, 229–230.

Mueller, B. U., Zeichner, S. L., Kuznetsov, V. A., Heath-Chiozzi, M., Pizzo, P. A. and Dimitrov, D. S. (1998). Individual prognoses of long-term responses to antiretroviral treatment based on virological, immunological and pharmacological parameters measured during the first week under therapy. *AIDS*, F191–F196.

Ryazanov, V. V. and Senko, O. V. (1990). On some voting models and methods of optimizing them, In *Recognition, Classification, Prediction (Mathematical Methods and their Application)*, N 3, 106–145, Nauka, Moscow (in russian).

Senko, O. V. (1993). The algorithm of prognosis, based on the procedure of voting by system of boxes on multidimensional space. *Pattern Recogn, Image Analysis*, 283–290.

Senko, O. V. and Kuznetsova, A. V. (1998). The use of partitions constructions for stochastic dependencies approximation, *Proceedings of the International conference on systems and signals in intelligent technologies*, 28–29 September, Minsk (Belarus), 291–297.

Soong, S. J. (1985) A computerized mathematical model and scoring system for predicting outcome in melanoma patients, In Cutaneous Melanoma, Clinical Management and Treatment Results Worldwide (Balch CM, Milton GW, eds), Philadelphia: Lippincott, 333–367.

Zadeh, L. A. (1984). Fuzzy sets and common sense knowledge. *Cognitive Science Report*, **21**, Berkeley: Univ, of California.

Zhirnova, I. G., Kuznetsova, A. V., Rebrova, O. Yu., Labunsky, D. A., Komelkova, L. A., Poleshchuk, V. V. and Senko, O. V. (1998). Logical and Statistical approach for the Analysis of Immunological Parameters in Patients with Wilson's Disease. *The Russian Journal of Immunology*, **3**(2), 173–184.

Zhuravlev, Yu, I., Gurevitch, I. B. and Ilyinsky, S. V. (1993). Development and investigation of the mathematical and computational basis for a system of information technologies of pattern recognition and image understanding. *Pattern Recognition and Image Analysis*, **3**, 266 (in russian).

Vapnik, V. P., Glazkova, T. G. and Kaschejev, V. A. (1984). Algorithms and dependence reconstruction programs, Moscow: Nauka.