CBC effectively stratifies the patients with different types of malignant tumors

M G Sadovsky¹,², A A Feller², E A Martynova², D V Chernyaev²,³, E V Semenov²,³, E V Slepov³ and R A Zukov²,³

¹ Institute of computational modelling of SB RAS, Krasnoyarsk, Russia
² V.F.Voino-Yasenetsky Krasnoyarsk state medical university, Krasnoyarsk, Russia
³ A.I. Kryzhanovski Krasnoyarsk Regional Clinical Oncology Hospital, Krasnoyarsk, Russia

E-mail: msad@icm.krasn.ru

Abstract. Early prediction of tumor process especially related to the detection of oncological diseases at early stage is a hot problem. We compared the complete blood count data for patients with various types of tumors; the patients with tumors of the hematopoietic system were excluded from the study. We used both linear (K-means) and non-linear (elastic maps) methods. No informative patterns have been found through the implementation of classic linear statistics methods. On the contrary, the elastic maps technique revealed the specific groups among the patients. Neither relation to the gender, nor age of patients has been found. However, the abundance of basophils, eosinophils, leukocytes and immunoglobulins exhibit specific pattern of the clusters occupancy. Thus, in our study we found that patients are stratified in accordance with quantitative changes in laboratory blood parameters, which is associated with the body's response to stress caused by a malignant tumor.

1. Introduction
Cancer is the second leading cause of death worldwide. Currently, a number of methods and techniques are implemented for early diagnostic and treatment of that latter [1–3]. Any human cell may become cancerous (malignant) one, resulting in a huge number of nosological forms and types of cancer occurrence, while them could be detected in the early stages of tumor formation.

Thus, the development and implementation of new methods for effective identification of cancer and further treatments is a major challenge in oncology. The central idea in the “elastic maps” method used in this study is to search for any inhomogeneities at the points distribution that might reflect relationships between variables or their combinations and some clinical effects not visible through the standard statistical analysis. The idea is as following: firstly, a researcher identifies the heterogeneity in the distribution of the data points and checks if these clusters are distinguishable. Secondly, an examination towards the connection between the composition of these clusters and the characteristics of the disease (or patient) should be carried out. This approach is very handy if large databases of medical records collected from hospitals or other medical institutions should be analyzed.

The relationship between CBC, biochemical and immunological data of breast cancer patients are discussed in [6]. The authors report the absence of significant intragroup scattering may reflect the stability of metabolic processes and the general stability of patients. Differential diagnostic of carcinoid...
tumors of the gastrointestinal tract according to the CBC discussed in [7]. CBC could be an effective tool for the differential diagnostic of some types of gastrointestinal cancer.

Complete blood count can be used as a prognostic marker of lymph node metastasis in endometrial cancer [8]. Patients who passed through the full course of treatment against the above disease could be reliably assigned to the risk group compared to the group with the lowest likelihood of metastasis. The classification was based on linear logistic regression; the latter revealed a hidden link between the growth of neutrophil cells and damage to the lymph nodes. Similarly, article [9] provides an assessment of the likelihood of endometrial cancer based on the results of a complete blood count.

The article [10] reports on the efficiency of CBC in examination of colon cancer for early diagnosis. The approach based on machine learning methods is reported in [11]. Finally, articles [12, 13] are fully consistent in the approach we used in our study. They present studies of simple screening techniques, including complete blood counts, for the diagnosis of myeloid leukemia. It is a simple and easy-to-use method for screening for myeloid leukemia. CBCs, namely absolute basophils, have been found to be effective in detecting chronic myeloid leukemia.

The key question of the article is can CBC be used to detect a tumor process. We have found that there are correlations between various indicators of a CBC that statistically reliably predict the development of cancer. Thus, the purpose of our study is to identify the distinctive features in the distribution of laboratory blood parameters in patients with various types of cancer, as well as to identify the relationships between the features of the course of the oncological process and the composition of the clusters identified in the multidimensional data space. The final goal of this work is to find the features of changes in CBC in cancer patients.

2. Materials and methods

Each patient record is a sequence of numbers derived from CBC results, thus including a dataset in a multidimensional metric space. Patients with tumors of the hematopoietic system were excluded from the study. We would like to emphasize that a detailed blood test is a mandatory laboratory test for all patients, including oncological patients, its cost, availability and ease of performance, along with the extensive use, make this analysis routine, and the results are comparable across the country. The principle of finding the relationship between blood test results and cancer is not new.

We carried out a software analysis of the database, which contains laboratory results of the CBC with differential of patients with various nosological forms of cancer. Blood samples were taken from patients upon the admission to A.I. Kryzhanovsky Krasnoyarsk Regional Oncological hospital. The database contains 867 records and includes the following parameters of a detailed blood test: BAS (absolute basophil content in peripheral blood), EOS (absolute eosinophil count in peripheral blood), HCT (hematocrit), HGB (hemoglobin concentration in whole blood), IG (immunoglobulin concentration in peripheral blood), LYM (absolute lymphocyte count in peripheral blood), MCH (average hemoglobin concentration per erythrocyte), MCHC (saturation of erythrocytes with hemoglobin), MCV (mean volume of erythrocytes), MON (absolute number of monocytes in peripheral blood), MPV (mean volume of platelets), NEUT (absolute number of neutrophils in peripheral blood), P-LCR (part of large platelets), PCT (percentage of platelet mass in blood volume), PDW (relative width of platelet distribution by volume), PLT (absolute number of platelets in peripheral blood), RBC (absolute number of red blood cells in peripheral blood), RDW-CV (change in the relative width of platelet distribution by volume), RDW-SD (standard deviation of the relative width of distribution of platelets by volume), WBC (absolute the number of leukocytes in peripheral blood) and the erythrocyte sedimentation rate (ESR).

Data preprocessing included standard statistical analysis, namely:

- mean and standard deviation were determined for each variable;
- each variable is checked for compliance with the normal distribution;
- for each variable, a histogram was developed to estimate the distribution model;
At the next stage, principal component analysis (PCA) [14] was implemented, as well as the correlation matrix. PCA was applied to determine the effective (linear) dimension of data space, and a correlation matrix was calculated to identify the couple of variables with high linear constraints. These latter were used to select variables from CBC records that have been excluded from further analysis.

The point is that numerous linear constraints can cause some bias in the disclosure of the internal structure. The following variable pairs are HCT-HGB (0.96), MCH-MCV (0.91), LYM-NEYT (0.96), MPV-PLCR (0.95), MPV-PDW (0.93), PLCR-PDW (0.94), PCT-PLT (0.96), RDWCV-RDWSD (0.85) usually show high and very high correlation coefficients. We used a threshold for excluding the variable from further analysis equal to ≥ 0.85; therefore, the variables MCV, NEUT, MPV, PDW, PLCR, PLT, RDWSD were excluded from further analysis. Thus, each patient is represented by a point in 21-dimensional metric space. Throughout the following, we have used Euclidean metrics to measure distance in space. Table 1 shows the number of neoplasms of various types in patients included in the database. So, we looked for heterogeneity in the distribution of patients included in the database, only in 21-dimensional Euclidean space according to CBC data. Then, if such clusters were found, then we carried out their subject analysis in terms of nosological forms, characteristics of patients or characteristics of the disease. We studied the structure and level of homogeneity of the distribution of patients in different clusters. In other words, it was very important to know which blood count plays a key role in the formation of the cluster.

### Table 1. Localization and type of neoplasm. Here I stands for malignant neoplasm, II stands for benign tumor, and III stands for unknown type.

| Localization                        | I | II | III |
|-------------------------------------|---|----|-----|
| head and neck cancer                | 33| 0  | 3   |
| digestion organs                    | 120| 15 | 30  |
| lung cancer                         | 63 | 2  | 15  |
| bones and conjuncture cartilage     | 3 | 0  | 0   |
| skin cancer                         | 9 | 5  | 7   |
| thyroid and other endocrinic glands | 3 | 8  | 0   |
| lymph, blood and related tissues    | 40 | 0  | 1   |
| mesothelial and soft tissues        | 11| 3  | 3   |
| mammary gland                       | 165| 1  | 11  |
| male and female reproductive system | 215| 2  | 17  |
| urine pathways                      | 29 | 5  | 17  |
| brain and other parts of CNS        | 6 | 1  | 0   |
| CUP syndrome                        | 23| 2  | 7   |

2.1. Clustering and visualization techniques
To select groups of similar indicators from the general data set, we used linear and nonlinear clustering methods. The former is K-means and the latter is the elastic map technique. K-means is a well-described and comprehensively researched clustering method (see [14]) and we shall not focus on it here.

The elastic map technique is a powerful and advanced technique based on approximation of multidimensional data with low-dimension manifold that looks like a custom geographical map. The elastic mapping technique provides non-linear dimension reduction: indeed, it reduces the dimension of the dataset to two. An implementation of elastic map starts from identification of the first and the second principal component [14]. Geometrically, the first principal component is the direction in the original Euclidean space yielding the highest scattering over the data. So, the first and the second principal components define the plane. Then, each data point must be projected onto a plane and the minimum square comprising all the projections must be determined. In the second step, each data point is connected to its projection with a (mathematical) spring; all springs are stipulated to have identical elasticity. Then the original rigid plane (or the square, to be exact) is changed for an elastic membrane. The membrane must be homogeneous: it has the same elastic properties at any point, in all directions.
So, at the third stage, the system is released to achieve a minimum of the total deformation energy. The latter is a combination of membrane elastic deformation (expansion, torsion and bending) and the mathematical expansion energy of the spring. Once an equilibrium configuration is reached, each data point must be redefined on the jammed surface. Indeed, it is necessary to find an orthogonal projection for each point; geometrically, this new image point is located at the jammed surface at the closest distance from the origin.

Figure 1. Distribution of patients on elastic map $16 \times 16$.

To reveal the cluster structure, it is necessary to return the jammed surface to a flat state. To do this, all the mathematical springs must be cut-off so that the jammed surface becomes flat. This is a non-linear transformation to so-called internal coordinates. Obviously, straightening of a jammed surface leads to a change in the position of the images of the points of the dataset (these were orthogonal projections). The projected images located on the convex part of the map will zoom in as the map straightens; conversely, projected images on the concave portion of the map become farther away. We used the freely distributed software VidaExpert [15].
In practice, due to computational constraints, no one implements an elastic map in software as a continuous object. On the contrary, the elastic map is presented in the form of a polyhedron, and the points are connected with the nodes of this latter, and not with the exact image of the projection. Depending on the number of nodes in a square, there can be three types of elastic maps: hard (usually $12 \times 12$), soft ($16 \times 16$), and detailed ($25 \times 25$); we used soft configuration in our research.

To visualize the distribution of the points of the newly obtained image on the map, one should introduce a certain clustering procedure by local density. There are several ways for this [16]; however, we implement, probably, the simplest of them, based on the local density function. To do this, it is necessary to provide each point of the image on the elasticity map with a Gaussian function,

$$f_j(r) = A \cdot \exp \left\{ \frac{(r - r_j)^2}{\mu^2} \right\}.$$  

(1)

$r_j$ is the vector of coordinates (in internal coordinates) of the $j$-th point, $r$ is the radius it, $A$ is a coefficient generally equal to 1, and $\mu$ (similar to the standard deviation in a normal distribution) is a contrasting parameter: it determines the width of the “field” covering the point.

It should be emphasized that for simplicity, function (1) is rotationally symmetric. Finally, the sum function, where $N$ is the total number of points in the dataset,

$$F(r) = \sum_{j=1}^{N} f_j(r)$$  

(2)

so that function (2) shows a local density function representing the cluster structure in the dataset, if any.

Figure 2. Variation of eigenvalues for the data set under consideration.

3. Results
First, let’s focus on principal component analysis and correlation analysis results of the original dataset. To determine the effective dimension of the dataset in a linear approximation, we carried out PCA; in particular, we started by calculating the eigenvalues of the covariance matrix obtained from the data. Figure 2 in the left shows the variation of eigenvalues. The curve has a clear break at the second eigenvalue; all other eigenvalues are almost linear in decrease. Thus, the effective spatial dimension in our database is one; however, the data can hardly be approximated by a one-dimensional linear manifold: the other eigenvalues are not small enough to be omitted. It means that the data are mostly dispersed along the line, but the distribution is rather complex.

Figure 2 in the right shows the above-mentioned data distribution in a special projection, where the first principal component is directed orthogonally to the plane of the drawing. Thus, the axes represent...
the contribution of each variable to the distribution pattern. We examined the classification of patients in the database using $K$-means. For $2 \leq K \leq 6$, no reasonable classification was observed; all classifications were very unstable, which means that most of the patients changed their class in a series of $K$-means runs. To get more information, we processed the database using the elastic maps technique.

Figure 1 shows a clustering sample obtained with a soft $(16 \times 16)$ elastic map, with default elasticity parameters. Obviously, there are three or four clusters on the map, depending on the contrast parameter. We also studied the distribution of different types of tumors on the map; the types are shown in Table 1. No specific predominance in the cluster structure was observed either in localization or in the type of tumor (malignant or benign).

At the next stage, the distribution of patients according to the elastic map was studied depending on the specific value of the parameters of the detailed blood test used. In fact, you can find the lowest and highest values for every 21 CBC parameters across the database. It should be emphasized that some patients fall outside the normal range of the indices, and some do not. So, the interval was divided into 10 equal intervals, and we identified patients with different values of CBC on the chart. Step by step, we labeled patients with a character level not exceeding $\lambda_j$; here $\lambda$ denotes a symbol, and $j$ denotes an interval number, $1 \leq j \leq 10$.

Surprisingly, we found that all 21 parameters of CBC split into two groups in terms of the distribution of patient patterns on the elastic map. The variant in which the data of patients, in the form of points, randomly placed on the map, in the form of the so-called “starry sky” observed for basophils, eosinophils, leukocytes and immunoglobulins. Of course, there are slight variations in the details of the “starry sky” appearance for some patients, depending on the blood counts. Figure 4 shows a picture of the “starry sky” type observed for the distribution of the content of eosinophils on the map.

The opposite type of distribution of points along the elastic map, called “wave”, observed in most of the patients. As the CBC score grows from the lowest figure to the highest one, the patients tend to occupy the map fairly regularly, layer by layer. Figure 3 really looks like a “moving wave”. Because the characteristics of patients have different starting positions on the map: this means there is (almost) no one in the database who has the minimum numbers of two characters at a time (see detailed discussion below). Figure 3 shows this wave pattern observed for the map distribution of monocyte counts.

4. Discussion

A malignant neoplasm development is a complex process involving a number of factors. Currently, there is no universal theory to classify these factors [17–20]. Thus, an implementation of logically apparent and informative predictors of the development of malignant neoplasms is still a hot problem, and significant progress has been reported here [1, 4, 5, 13, 20]. It is mostly based on advanced molecular biology techniques. There are two fundamental problems with the modern big data approach to medical data processing:

- simultaneous and combined analysis of a number of different characteristics that were not initially included in the general theory is the first problem;
- “modeling” data, that is, the approximation of multidimensional volumetric data by a set of low dimensions is the second problem;

In our work we have implemented both of these ideas. Neither the elastic map method (see figure 1), nor the $K$-means revealed a relationship between the cluster and the nosological form of malignant neoplasm; the same is true for benign growths. This failure may be the result of over-specifying the location and type of cancer.
Figure 3. Wave distribution of patients with different levels of monocytes on an elastic map $16 \times 16$; see text for explanation; monocytes.
Figure 4. Distribution by the type of starry sky of patients with different levels of monocyte content according to the elastic map 16 × 16; see text for explanation.

We examined all the characteristics of CBC, in terms of the map filling scheme. It shows that different characteristics start to fill the map from different places, and in their own way. This fact, apparently, is a manifestation of Liebig’s principle: any cancer tumor is a serious stress factor. Thus, stress causes the mobilization of the resources of the sick organism, concentrating them in a kind of “bottle-neck”. The most surprising thing here is that the database contains a number of patients with different pathologies in different organs and tissues. Regardless the specific type of tumor localization, clustering by elastic map technique shows that there is a kind of specialization of pathways and contours for resource mobilization.

Indeed, there is only one pair of variables that gives a fairly similar pattern of filling the map as the values increase: hematocrit and hemoglobin. This situation seems quite natural: the correlation coefficient between these two variables is very high. These two CBC options start to fill the map from the left edge and flow smoothly to the right, keeping a fairly straight line. In contrast, red blood cells
begin to fill the map from the upper left corner and do so as an arc wave centered in the upper left corner. Lymphocytes and average red blood cell volume give the most amazing filling patterns. Begin to fill out the map from two and three sections, respectively. Lymphocytes begin to fill the map from the lower left and upper right corners. Surprisingly, the patients with the highest lymphocyte counts are aggregated in the central part of the map, occupying an area along the “main” diagonal of the latter, extending from the upper left to the lower right. The distribution pattern of the average volume of erythrocytes is unique: it begins to fill the map from three areas. The latter are located in the upper left corner, in the upper right and lower left. The above-described patterns of filling in the card unambiguously confirm the distribution of patients over the space determined by the characteristics of the general blood test. There is no stratification of patients in terms of the correspondence of the clusters identified by the elastic map method to the age, gender or nosological form of the tumor. In contrast, patients are stratified according to the type of cancer response. There is an obvious correlation between the composition of the cluster and the nature of the filling of the map. The latter reflects the type and way of mobilizing the body’s resources caused by a malignant neoplasm, and the stratification strategy follows the Liebig’s principle. This is basically a new stratification seen in cancer patients, and further research is still needed into the detailed relationships between response type and relevant important clinical aspects.

Acknowledgments
This study supported be the grant from Voino-Yasenetsky Krasnoyarsk State Medical University, grant № 2.3 under the contract № 203.

References
[1] Arem H, Loftfield E 2018 American journal of lifestyle medicine 12(3) 200-10
[2] Kasting M L, Giuliani A R, Reich R R, Roetzheim R G, Nelson D R, Shenkman E et al. 2018 Cancer Epidemiology and Prevention Biomarkers 27(4) 503-13
[3] Albeshan S M, Mackey M G, Hossain S Z, Alfuraid A A and Brennan P C 2018 Clinical breast cancer 18(3) e381-e92
[4] Almugren N and Alshamlan H 2019 IEEE Access 7 78533-48
[5] Dubey A K, Gupta U and Jain S 2015 Asian Pac J Cancer Prev. 16(10) 4237-45
[6] Satparowa L, Kogina E, Satparov Y, Galimov S, Knyazeva O and Udut V 2018 Evaluation of some characters of biochemical, clinical and immunological analysis of blood of the patients with mammalian cancer. In: Contemporary aspects of control and regulation 246-53
[7] Arlekyia I and Pleten’ A 2017 Fundametal problems of science 13 131-4
[8] Taş E E, Özgen E and Yavuz AF 2018 Cyprus Journal of Medical Sciences 3(3) 168-72
[9] Yayla Abide C, Bostancı-Engen E, Cogendez E, Kiliciç C, Uzun F, Ozkaya E et al. 2018 Journal of clinical laboratory analysis 32(6) e22438
[10] Hornbrook M C, Goshen R, Choman E, O’Keefe-Rosetti M, Kinar Y, Liles EG et al. 2018 Digestive diseases and sciences 63(1) 270
[11] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM et al. 2017 Nature 542(7639) 115
[12] Ogasawara A, Matsushita H, Tanaka Y, Shirasugi Y, Ando K, Asai S et al. 2019 Clinica Chimica Acta 489 249-53
[13] Kourou K, Exarchos T P, Exarchos K P, Karamouzis M V and Fotiadis D I 2015 Computational and structural biotechnology journal 13 8-17
[14] Fukunaga K 1990 Introduction to statistical pattern recognition (London: Academic Press)
[15] Gorban A N and Zinovyev A Y 2015 Fast and user-friendly non-linear principal manifold learning by method of elastic maps In: 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015 Campus des Cordeliers Paris France 1-9
[16] Xu D and Tian Y 2015 Annals of Data Science 2(2)165-93
[17] Brown K F, Rumgay H, Dunlop C, Ryan M, Quartly F, Cox A et al. 2018 British journal of cancer 118(8) 1130
[18] Sakamoto Y, Kokudo N, Matsuyama Y, Sakamoto M, Izumi N, Kadoya M et al. 2016 Cancer
122(1) 61-70
[19] Islami F, Goding Sauer A, Miller KD, Siegel RL, Fedewa SA, Jacobs EJ et al. 2018 CA: a cancer journal for clinicians 68(1) 31-54
[20] Dubey A K, Gupta U and Jain S 2016 Chinese journal of cancer 35(1) 71