Methods for Discrete Analysis of Medical Data on the Basis of Recognition Theory and Some of Their Applications

Yu. I. Zhuravlev[a], G. I. Nazarenko[b], A. P. Vinogradov[a], A. A. Dokukin[a], N. N. Katerinochkina[a], E. B. Kleimenova[b], M. V. Konstantinova[b], V. V. Ryazanov[a]*, O. V. Sen'ko[a], and A. M. Cherkashov[b]

[a] *Dorodnicyn Computing Centre, Federal Research Center "Informatics and Control," Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119333 Russia*

[b] *Central Bank of Russia, Medical Center, Sevastopol'skii pr. 66, 117593 Russia*

*e-mail: *rvvccas@mail.ru, dalex@ccas.ru*

Abstract—Methods for the analysis of medical data and the results of their application to the treatment of a number of socially important diseases in important medical areas (cardiology, neurology, surgery, and oncology) are considered. The precedent approach is investigated. Practical methods of discrete analysis of training data, logical and statistical methods for searching logical regularities in data, combinatorial logic and logical statistical classification methods, and methods for estimating models and searching for "nonstandard" descriptions are presented. The results of experiments on real data are demonstrated.

Keywords: precedence, training, logical regularity, statistical significance, classification algorithm, feature, cardiology, neurology, oncology, surgery

DOI: 10.1134/S105466181603024X

INTRODUCTION

In this paper, we consider methods for the analysis of medical data and the results of their application to the treatment of a number of socially important diseases. In various medical institutions, large volumes of medical information on patients and their diseases have been accumulated. Quality registers are one of the sources of this information. Clinical registers are databases that systematically accumulate medical information on people exposed to some intervention, diagnosed with a certain disease or a state, and provided medical care with some resources. To assess the quality of treatment, information is used that can be presented in the form of clinical outcomes (for example, fatal outcomes or functional results of treatment), indicators of clinicodiagnostic processes, or indicators of quality (for example, compliance to clinical recommendations). The data of the registers on diseases or interventions can be represented as tables in which the rows correspond to various patients and columns correspond to various indicators of a patient related to a given disease (risk factors, clinical symptoms, the results of examination and treatment, and outcomes of treatment). The main problem consists in the discrete analysis of these data to detect regularities between different parameters of a given disease. After that, these regularities can be used as scientific hypotheses and checked empirically or with the use of various statistical techniques.

Received January 22, 2016

In Section 1, we give the main definitions related to the logical regularities of classes (LRCs), their search, and the solution of a classification problem. We describe an algorithm for searching elementary LRCs as linear functions of features and constructing LRCs of the second type on the basis of these functions. Then we describe a statistical approach to searching for LRCs for large training samples. For consistent training samples, we find the set of all minimal LRCs. In subsection 1.3, we consider the question of searching for LRCs for large training samples and statistical estimate of their reliability. In subsection 1.4, we consider some methods for discrete processing of the LRCs obtained. We solve the problem of searching for minimal LRCs that are equivalent to the original ones but have minimal complexity. As the characteristic function of each class of the training sample, one can use the disjunction of all minimal LRCs of this class. This expression may be very cumbersome. We give the definitions of the shortest and minimal logical descriptions of classes as those that are equivalent to the original logical descriptions but have minimal complexity. In subsection 1.5, we present the results of the prediction of complications of acute cardiovascular attack (ACVA), that are obtained by the method of optimal valid partitioning [1].

In Section 2, we present some methods that are used for the analysis of medical information. In subsection 2.1, we describe the principle of ROC analysis and its realization in the Recognition system. In subsection 2.2, we present an algorithm for smoothing an empirical distribution to remove random outliers in medical information.

Section 3 is devoted to the preparation of medical data for analysis and classification. We consider the general structure of the main clinical registers on neurological, cardiological, surgical, and oncological diseases.

Section 4 presents the results of application of models for searching LRCs, the RECOGNITION system, and the methods of classification and outcome prediction in some disciplines (cardiology, neurology, oncology, and surgery) based on the Medical Center of the Bank of Russia.

In Conclusions, we give a short discussion of the studies and the plans of future research.

1. Logical Regularities of Classes and Their Applications

Consider a standard supervised classification (recognition) problem by precedents with n features, l disjoint classes $K_1, K_2, \ldots, K_l$, and m standard objects $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ (the training sample). We will use the notations $\tilde{K}_i = X \cap K_i$, $i = 1, 2, \ldots, l$, and assume that $\tilde{K}_i \neq \varnothing$, $i = 1, 2, \ldots, l$. An arbitrary object $\mathbf{x} \in \bigcup_{i=1}^{l} K_i$ is identified with its feature description as a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. For simplicity, we assume that $x_i \in \mathbf{R}$ (binary and k-valued features are considered as particular cases of real-valued features). When analyzing a training sample, we will often write simply, without special stipulation, K_i, assuming that we always deal with $\tilde{K}_i$. At the Federal Research Center "Informatics and Control," Russian Academy of Sciences, various recognition models have been developed that are based on the discrete analysis of training data (a test algorithm [2], algorithms for calculating estimates [3], algorithms of voting by representative sets [4], recognition algorithms by valid partitioning [1], algorithms of voting by systems of logical regularities [5], and others). All these recognition models are based on the discrete analysis of training information and a search for and use of its informative fragments. One can use the common term "partial precedence models" for these models. Below we present the description of algorithms based on voting by systems of logical regularities. These algorithms are described in part in [6].

1.1. Main Definitions

Definition 1. A predicate

$$P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}) = \mathop{\&}_{j\in\Omega_1}(c_j^1 \le x_j) \mathop{\&}_{j\in\Omega_2}(x_j \le c_j^2) \qquad (1)$$

is called a *logical regularity of class* K_λ, $\lambda = 1, 2, \ldots, l$, if

1. $\exists \mathbf{x}_t \in \tilde{K}_\lambda : P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = 1$,

2. $\forall \mathbf{x}_t \notin \tilde{K}_\lambda : P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = 0$,

3. $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}) = \underset{\{P^{\Omega_1^*,\mathbf{c}^{*1},\Omega_2^*,\mathbf{c}^{*2}}(\mathbf{x})\}}{extr} \Phi(P^{\Omega_1^*,\mathbf{c}^{*1},\Omega_2^*,\mathbf{c}^{*2}}(\mathbf{x}))$,

where Φ is a quality criterion of a predicate, $\Omega_i \subseteq \{1, 2, \ldots, n\}$, $i = 1, 2$.

Definition 2. The following criterion is called a standard quality criterion of a predicate of class K_λ:

$$F(P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})) = |\{\mathbf{x}_i : \mathbf{x}_i \in \tilde{K}_\lambda, P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_i) = 1\}|.$$

Predicate (1) satisfying only the first two constraints is called an admissible predicate of the class.

Predicate (1) satisfying only the first and the third constraints is called a partial logical regularity of the class (LRC) K_λ.

The set $N(P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}) = \{\mathbf{x} \in R^n : c_j^1 \le x_j, j \in \Omega_1; x_j \le c_j^2, j \in \Omega_2\}$ is called the interval of the predicate $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})$ (an analog of the intervals of elementary conjunctions in logical algebra).

Two predicates $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})$ and $P^{\Omega_3,\mathbf{c}^3,\Omega_4,\mathbf{c}^4}(\mathbf{x})$ are said to be equivalent if $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = P^{\Omega_3,\mathbf{c}^3,\Omega_4,\mathbf{c}^4}(\mathbf{x}_t)$, $t = 1, 2, \ldots, m$.

Two intervals $N(P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2})$ and $N(P^{\Omega_3,\mathbf{c}^3,\Omega_4,\mathbf{c}^4})$ are said to be equivalent if $N(P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}) \cap X = N(P^{\Omega_3,\mathbf{c}^3,\Omega_4,\mathbf{c}^4}) \cap X$.

Definition 3. A logical regularity of class $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})$ is said to be minimal if there does not exist an equivalent LRC $P^{\Omega_3,\mathbf{c}^3,\Omega_4,\mathbf{c}^4}(\mathbf{x})$ such that $N(P^{\Omega_3,\mathbf{c}^3,\Omega_4,\mathbf{c}^4}) \subset N(P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2})$.

In [7], the authors described relaxation, combinatorial, and genetic algorithms for solving this problem. The search for all LRCs consists in searching for minimal logical regularities of class K_λ for every training object $\mathbf{x}_t, t = 1, 2, \ldots, m, \mathbf{x}_t \in K_\lambda, \lambda = 1, 2, \ldots, l$, provided that $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = 1$. In this case, finding a minimal LRC with a standard quality criterion reduces to searching for the maximum consistent subsystem of a system of linear inequalities under linear constraints and binary unknowns. All unknown coefficients of the system of linear inequalities are calculated by the training sample.

Suppose that, for every class K_λ, a set of linear regularities $\mathrm{P}_\lambda = \{P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})\}$ is found and that the set of intervals $\{N(P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}) : P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2} \in \mathrm{P}_\lambda\}$ covers $\tilde{K}_\lambda$. The functions $D_\lambda(\mathbf{x}) = \bigvee_{P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2} \in \mathrm{P}_\lambda} P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})$ take value 1 on the objects of $\tilde{K}_\lambda$ and 0 on other training objects. Suppose that the following estimate is calculated for an object $\mathbf{x}$ to be recognized on the basis of the sets P_λ:

$$\Gamma_\lambda(\mathbf{x}) = \frac{1}{|\mathrm{P}_\lambda|} \sum_{P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2} \in \mathrm{P}_\lambda} P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})$$

for a class $K_\lambda, \lambda = 1, 2, \ldots, l$ (the "degree of proximity" of $\mathbf{x}$ to K_λ). Then one applies, for example, the following general decision rule:

$$\alpha_j^A(\mathbf{x}) = \begin{cases} 1, & \sum_{i=1}^{l} \delta_i^j \Gamma_i(\mathbf{x}) \geq 1, \\ 0, & \text{otherwise}. \end{cases}$$

Here $\alpha^A_j(\mathbf{x}) = 1$ ($\alpha^A_j(\mathbf{x}) = 0$) denotes a solution $\mathbf{x} \in K_j$ ($\mathbf{x} \notin K_j$) of algorithm A. The parameters δ_i^j are determined during the optimization of the recognition model with the use of a test sample.

1.2. A Method for Searching for Logical Regularities of Classes by Constructing Polyhedral Sets

Definition 4. A predicate of the form

$$P_i(\mathbf{x}) = (\sum_{j=1}^{n} a_{ij} x_j - b_i \leq 0) \tag{2}$$

is called an *elementary predicate*.

We will also write it as

$$P_i(\mathbf{x}) = ((\mathbf{a}_i, \mathbf{x}) - b_i \leq 0). \tag{3}$$

Here $(\mathbf{a}_i, \mathbf{x})$ is the scalar product of vectors $\mathbf{a}_i, \mathbf{x} \in \mathbf{R}^n$ and $b_i \in \mathbf{R}$.

Definition 5. A logical regularity of the second type is a logical regularity represented in the form of a predicate of the form

$$P(\mathbf{x}) = \bigvee_{j=1}^{p} \&_{i=1}^{q} P_i^j(\mathbf{x}), \tag{4}$$

where $P_i^j(\mathbf{x})$ are some elementary predicates.

The method proposed for finding logical regularities for a given class consists in the construction of the convex hulls of subsets of standards of the training sample.

Introduce the following notations. Denote by $S(K_i)$ a subset of the training sample that consists of the standards of class K_i.

Denote by $V(K)$ the convex hull of an arbitrary finite set K, and by $V(K_i)$, the convex hull of the standards of class K_i. Thus, by definition, we have

$$V(K_i) \equiv V(S(K_i)).$$

Case $n = 2$.

A training sample $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ of objects from the classes $K_1, K_2, \ldots, K_l$ is defined on the plane. For an arbitrary class K_λ, we will construct logical regularities of the form (4) that separate the objects of a given class from all the other objects. Denote by CK_λ the set of all objects that do not belong to the class K_λ. Let $S(K_\lambda) = \{S_1, \ldots, S_r\}$ be the set of standards of class K_λ, where $S_i = (x_i, y_i)$.

First, construct the convex hull of the set of standards of class K_λ. To this end, we apply one of the known methods: the algorithms of Graham, Chan, Kirkpatrick, and other algorithms.

The convex hull $V(K_\lambda)$ obtained is determined by its extreme points (vertices of the boundary polygonal line). Denote them by $\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_t$.

After constructing the convex hull $V(K_\lambda)$ of the set of standards of class K_λ, one should check if "another's" objects (i.e., objects from the set CK_λ) fall into this class. If $V(K_\lambda) \bigcap CK_\lambda \neq \varnothing$, then we proceed to constructing a logical regularity of general form.

To this end, we first find the center of mass $\mathbf{p} = (\overline{x}, \overline{y})$ of the set of standards of class K_λ. We have

$$\overline{x} = \frac{1}{r}\sum_{i=1}^{r} x_i, \quad \overline{y} = \frac{1}{r}\sum_{i=1}^{r} y_i.$$

From the set of standards of class K_λ, we choose a point $\mathbf{q} = (x_q, y_q)$ closest to the point $\mathbf{p}$. Connect the point $\mathbf{q}$ with all the vertices of the boundary convex hull. Then the set $V(K_\lambda)$ is represented in the form of a union of triangles with vertices $\mathbf{q}, \mathbf{v}_i, \mathbf{v}_{i+1}$. We will successively analyze these triangles starting from $\Delta\mathbf{q}\mathbf{v}_0\mathbf{v}_1$ and ending with $\Delta\mathbf{q}\mathbf{v}_t\mathbf{v}_0$.

Consider $\Delta\mathbf{q}\mathbf{v}_i\mathbf{v}_{i+1}$. Let us check if "another's" objects fall into this triangle. To this end, it suffices to check the standards from CK_λ with coordinates x, y satisfying the following relations:

$$\min(x_q, x_i, x_{i+1}) \leq x \leq \max(x_q, x_i, x_{i+1}),$$

$$\min(y_q, y_i, y_{i+1}) \leq y \leq \max(y_q, y_i, y_{i+1}),$$

where $\mathbf{v}_i = (x_i, y_i)$, $\mathbf{v}_{i+1} = (x_{i+1}, y_{i+1})$.

If another's points fall into $\Delta\mathbf{q}\mathbf{v}_i\mathbf{v}_{i+1}$, then we should remove them from our set. Denote them by $\mathbf{w}_1^i, \ldots, \mathbf{w}_s^i$. Let us construct a set $V_i = V(\{\mathbf{w}_1^i, \ldots, \mathbf{w}_s^i\})$ – the convex hull of standards from the set CK_λ that fall into $\Delta\mathbf{q}\mathbf{v}_i\mathbf{v}_{i+1}$.

Having analyzed all triangles of the convex hull $V(K_\lambda)$, we obtain a collection of sets $V_0, V_1 \ldots, V_k$, $k \leq t+1$. Suppose that $S(K_\lambda) \bigcap \left(\bigcup_{j=0}^{k} V_j\right) = \varnothing$.

Then the condition under which an object $\mathbf{x}$ belongs to the class K_λ can be expressed as

$$(\mathbf{x} \in V(K_\lambda)) \,\&\, \left(\mathbf{x} \notin \bigcup_{j=0}^{k} V_j\right).$$

Here the condition $\mathbf{x} \notin \bigcup_{j=0}^{k} V_j$ is equivalent to the condition $\&_{j=0}^{k}(\mathbf{x} \notin V_j)$.

Suppose that $V(K_\lambda)$ is described by a predicate

$$\&_{i=1}^{q}((\mathbf{a}_i, \mathbf{x}) - b_i \le 0).$$

The condition $\mathbf{x} \notin V_j$ can be described by a predicate of the form

$$\vee_{t=1}^{u_j}((\mathbf{c}_t^j, \mathbf{x}) - d_t^j \le 0).$$

Then a logical regularity for the class K_λ has the form

$$\&_{i=1}^{q}((\mathbf{a}_i, \mathbf{x}) - b_i \le 0)\ \&_{j=0}^{k}\ \vee_{t=1}^{u_j}((\mathbf{c}_t^j, \mathbf{x}) - d_t^j \le 0).$$

It is obvious that it can be reduced to the form (3).

The logical regularity obtained accurately describes the class K_λ when the sets V_j do not contain standards from the given class. Otherwise we assume that this logical regularity is an approximate description of the class. Anyway, the condition of this predicate is sufficient (but not always necessary) for it to belong to the class.

The main idea of the approach proposed consists in obtaining a representation of each class in the form of a difference between a convex set and a polyhedral set. A polyhedral set is a union of a finite number of convex sets.

The complexity of the construction of a logical regularity of the form (3) for one class is estimated as $O(m \cdot \log m + mt)$. Here m is the number of precedents in the training sample and t is the number of extreme points in the convex hull of the set of precedents of the given class.

The algorithm described consists of several stages:

—construction of the convex hull of standards of this class;

—finding the center of mass for the set of standards of the class;

—decomposition of the convex hull constructed into triangles;

—construction of the convex hull of "another's" elements in each triangle;

—finding a logical regularity for the class–a predicate of the form (3).

Denote the algorithm constructed by $\mathbf{A_2}$.

Case $n > 2$.

Next, we propose an approach that allows one to use the algorithm constructed above for $n > 2$.

We consider all possible pairs of features (x_i, x_j), $i \ne j$. For each such pair, we analyze the projections of objects of the training sample onto the plane of axes i, j. For a given class K_λ and a pair (x_i, x_j), we should try to separate the projections of standards of this class from the projections of standards from the set CK_λ.

Introduce the following notations.

Let $\Pi(K_\lambda, i, j)$ be the set of projections of standards of class K_λ onto the plane of axes i, j.

Analogously, $\Pi(CK_\lambda, i, j)$ is the set of projections of standards from CK_λ onto the plane of axes i, j.

By algorithm $\mathbf{A_2}$, we construct a predicate of the form (3) for the set $\Pi(K_\lambda, i, j)$. However, here, in contrast to the case of $n = 2$, there may be domains where the projections of "own" and "another's" objects are mixed. Then, the predicate constructed will separate many own standards together with another's standards. Denote by ρ_{ij} the number of separated own standards for a pair of features (x_i, x_j). Take a pair for which this quantity is minimal. For the standards that fall into the domain separated by the predicate, we can add the third feature, or consider other pairs of features. Finally, we choose a family of pairs or triples of features for which the union of predicates makes the minimum number of errors (leaves the minimum number of unrecognized standards of the class). Each of these predicates depends only on the values of two (or three) features. The disjunction of these predicates is considered as a logical regularity for the given class.

1.3. Finding LRCs for Large Samples and Statistical Estimation of Their Reliability

An algorithm for searching all LRCs suggests the solution of special discrete optimization problems for each training object**.** The complexity of the main problem significantly increases with the length of the training sample. The LRCs obtained when outlier objects are used as support elements will have a low value of the quality criterion. The effective solution for samples of large length is possible for the following modification of the search problem.

Let us set up the problem of searching for LRCs $P^{\Omega_1, \mathbf{c}^1, \Omega_2, \mathbf{c}^2}(\mathbf{x})$ such that $F(P) \ge h$, where $h > 0$ is a parameter. Henceforth, we assume that $F(P)$ is normalized (i.e., $0 \le F(P) \le 1$). It is assumed a priori that logical regularities of given quality exist. Let g be a parameter ($0 < g < 1$).

Let us describe an algorithm for searching LRCs whose result is formulated as follows: In the calculated set of logical regularities $\mathbf{P_j} = \{P(\mathbf{x})\}$ of class K_j, with probability of at least g, there is a logical regularity $P(\mathbf{x})$ such that $F(P) \ge h$.

Suppose that the problem of searching for logical regularities of a class is solved for k randomly chosen "support" standards of class K_j and that all the found logical regularities are combined into a single set $\mathbf{P_j}$. Then the probability that the sought predicate is not

found is estimated from above as $(1-h)^k$. Then the value of the parameter k is determined from the relation

$$(1-h)^k \le (1-g). \tag{5}$$

The value of the parameter k is an important factor of the efficiency of the algorithm. For example, for $g = 0.9$ and $h = 0.1$, equality (2) implies $k \ge 22$. The value of $k = 22$ is quite acceptable for problems of large dimension.

In conclusion, consider the problem of choosing the value of the parameter h. How to estimate it for a given specific problem and what its value is valid from some viewpoint?

The logical regularities of class K_j may and may not be "statistically significant." To this end, one applies an approach called a "permutation test." A series of the following calculations of the same type is performed (t is the "number of random permutations"). A random permutation of the rows of the training table is performed; in this case, as before, the first m_1 rows of the new table $\tilde{T}^i_{nml}$, $i = 1,2,\ldots,t$, are assumed to be the standards of the first class, the next m_2 rows, the standards of the second class, and so on (i.e., the numbers of classes of standard objects are randomly changed while preserving the total number of standards of the class). For the class K_j of the table $\tilde{T}^i_{nml}$, we find a regularity P_{ji} with quality $F(P_{ji})$. Then a logical regularity P_q from the set $\mathbf{P_j} = \{P(\mathbf{x})\}$ is assumed to be statistically significant if at least $100q$, %, $0 < q < 1$ (q is the significance level), of inequalities $F(P_q) \ge F(P_{ji})$, $i = 1, 2, \ldots, t$, are satisfied. Thus, the values of $F(P_j)$ obtained as a result of permutation tests are a natural variant for the choice of the parameter h.

1.4. Logical Descriptions of Classes

Earlier, algorithms for searching logical regularities of class K_λ, i.e., predicates of the form $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}) = \mathop{\&}_{j\in\Omega_1}(c^1_j \le x_j)\mathop{\&}_{j\in\Omega_2}(x_j \le c^2_j)$ have been considered. Consider the problems of searching for logical regularities that are equivalent to the original ones but have minimal complexity. In other words, the problem consists in searching for an LRC that is equivalent to $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})$ and is such that $|\Omega_1| + |\Omega_2|$ is minimal.

Consider the following integer linear programming problem.

$$\sum_{j=1}^{n}(y^1_j + y^2_j) \to \min,$$

$$\begin{gathered}\sum_{j=1}^{n}(1-(c^1_j \le x_{ij}))y^1_j \\ +\sum_{j=1}^{n}(1-(x_{ij} \le c^2_j))y^2_j \ge 1, \forall x_i \notin K_\lambda,\end{gathered} \tag{6}$$

$$y^1_j, y^2_j \in \{0,1\}.$$

The set of all unit components of the solution $(y^1_1, y^1_2,\ldots, y^1_n, y^2_1, y^2_2,\ldots, y^2_n)$ biuniquely defines the corresponding subsets of features Ω_1, Ω_2. Indeed, the predicate $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}) = \mathop{\&}_{j\in\Omega_1}(c^1_j \le x_j)\mathop{\&}_{j\in\Omega_2}(x_j \le c^2_j)$ obtained satisfies the definition of a logical regularity, and the fulfillment of the linear constraints in (6) corresponds to the fulfillment of the second condition of the main definition. The set of all units of the predicate constructed is an analog of a maximal interval in the Boolean case.

Suppose that the set of LRCs $P_t(\mathbf{x}), t \in T$, is calculated for a class K_λ. We assume that the training sample is consistent (there are no equal objects from different classes).

Definition 6. A logical description of a class K_λ is the logical sum $D_\lambda(\mathbf{x}) = \bigvee_{t\in T} P_t(\mathbf{x})$.

It is clear that $D_\lambda(\mathbf{x}_t) = 1$ for all training objects from the class K_λ and $D_\lambda(\mathbf{x}_t) = 0$ for all standard objects that do not belong to the class K_λ. Thus, $D_\lambda(\mathbf{x})$ coincides with the characteristic function of the class K_λ on the set of descriptions of standard objects.

Obviously, $D_\lambda(\mathbf{x})$ is a direct analog of reduced DNFs for Boolean functions. Then it is natural to consider analogs of minimal and shortest DNFs.

Definition 7. The shortest logical description of class K_λ is a logical sum $D^s_\lambda(\mathbf{x}) = \bigvee_{t\in T'\subseteq T} P_t(\mathbf{x})$ in which $|T'| \to \min$ and the function $D^s_\lambda(\mathbf{x})$ coincides with $D_\lambda(\mathbf{x})$ on the training sample.

Denote also by $P_t(\mathbf{x})$, $t \in T$, a set of LRCs of minimal complexity found by the original set of logical regularities. Suppose that the complexity of each minimal LRC of minimal complexity is $|\Omega_{t1}| + |\Omega_{t2}|$. Then we adopt the following definition.

Definition 8. A minimal *logical description of a class* K_λ is a logical sum $D^m_\lambda(\mathbf{x}) = \bigvee_{t\in T''\subseteq T} P_t(\mathbf{x})$ in which $\sum_{t\in T''\subseteq T}(|\Omega_{t1}| + |\Omega_{t2}|) \to \min\limits_{T''}$ and the function $D^m_\lambda(\mathbf{x})$ coincides with $D_\lambda(\mathbf{x})$ on the training sample.

The definition of a minimal logical description of a class is introduced by analogy with the Boolean case.

Logical (shortest, minimal) descriptions of classes are analogs of representations of partial Boolean functions in the form of abridged (shortest, minimal) disjunctive normal forms, and the geometric images of logical regularities of classes are analogs of maximal intervals.

The problem of searching for the shortest logical descriptions is formulated as follows:

$$\sum_{t\in T} y_t \to \cdot\min,$$

$$\sum_{t\in T} P_t(\mathbf{x}_i)y_t \geq 1, \cdots \forall \mathbf{x}_i \in K_\lambda, \cdot y_t \in \{0,1\}.$$

The problem of searching for minimal logical descriptions is formulated as a covering problem:

$$\sum_{t\in T} (|\Omega_{t1}| + |\Omega_{t2}|)y_t \to \cdot\min,$$

$$\sum_{t\in T} P_t(\mathbf{x}_i)y_t \geq 1, \cdots \forall \mathbf{x}_i \in K_\lambda, \cdot y_t \in \{0,1\}.$$

The unit components of the vector $\mathbf{y} = (y_1, y_2, \ldots, y_{|T|})$ biuniquely define the sets T' and T''.

Note that the original sets $P_t(\mathbf{x})$, $t \in T$, may contain equal or close elements and "degenerate" solutions corresponding to maxima with small absolute value; the cardinality of these sets may be very large (which, however, is favorable in recognition procedures). These "properties" essentially depend on the length of the training sample and the search algorithm itself. At the same time, the shortest and minimal logical descriptions of classes form subsets that are no longer redundant and reflect both the main properties of these sets and the properties of the classes themselves. Therefore, the calculation of $D_j^s(\mathbf{x})$ and $D_j^m(\mathbf{x})$ can be considered as one of approaches to the processing of the sets of logical regularities of classes. The LRCs appearing in $D_j^s(\mathbf{x})$ and $D_j^m(\mathbf{x})$ can be considered as the most compact representations of classes, which include both the most representative knowledge (predicates covering a large number of standards) and unique or rare (predicates covering a small number of standards or certain individual ones).

Definition 9. The logical complexity (compactness) of classes is given by:

1. $\psi_1(K_j) =<$ the number of conjunctions in $D_j^s(\mathbf{x})>$;

2. $\psi_2(K_j) =<$ the number of variables in $D_j^m(\mathbf{x})>$.

The quantity $\Phi(X) = \sum_{i=1}^{l} \psi(K_i)$, where ψ is a logical complexity criterion of a class, is called the logical complexity of the problem.

It is natural to expect that if a class is a "compact," simple set of objects that is well logically separable from other classes, then it has a small number of variables in a minimal logical description of a class and/or a small number of conjunctions in the shortest description.

1.5. Prediction of ACVA by the Method of Optimal Valid Partitioning

The method of optimal valid partitioning (OVP) is based on the search for the partitions of one-dimensional intervals of individual features or two-dimensional ranges of indicator pairs that allow one to separate observations with different values of objective variable over the training sample in the best way. In the present case, the objective variable is a binary indicator function that points to the presence or absence of complications when treating ACVA.

The training sample has the form $\tilde{S}_0 = \{(\alpha_1, \mathbf{x}_1), \ldots, (\alpha_m, \mathbf{x}_m)\}$, where $\alpha_j = 1$ in the presence of complications and $\alpha_j = 0$ in their absence Partitions are constructed so that to separate domains that are characterized by maximally different rates of complications. Suppose that R is a partition of the training sample $\tilde{S}_0$ into subsamples $\tilde{S}_1, \ldots, \tilde{S}_q$. Then the maximum difference corresponds to the maximum of the quality functional, which is defined as the sum

$$F_I(\tilde{S}_0, R) = \frac{1}{\nu(\tilde{S}_0)[1-\nu(\tilde{S}_0)]} \sum_{i=1}^{q} [\nu(\tilde{S}_i) - \nu(\tilde{S}_0)]^2 m_i, \quad (7)$$

where $\nu(\tilde{S}_*)$ denotes the fraction of objects with complications in the sample $\tilde{S}_*$, $m_i = |\tilde{S}_*|$.

The search for regularities reduces to the construction of boundaries, parallel to the coordinate axes, for which $F_I(\tilde{S}_0, R)$ attains its maximum value. When two-dimensional models are used, an optimal solution is sought by enumeration of all possible combinations of boundaries.

The verification of regularities is performed by the permutation test, which allows one to estimate the significance of differences with respect to each variable in the case of two-dimensional models.

Estimation of the reliability of simplest one-dimensional models. For the simplest one-dimensional model, a permutation test consists in a multiple repetition of the search for optimal boundaries on the samples obtained from the original sample by random permutations of the goal variable, in our case of the indicator of presence of complications, the variables used for calculating a prediction. To this end, by a random number generator, we generate a set of permutations $\{\tilde{f}^1, \ldots, \tilde{f}^N\}$ of natural numbers from the set $\{1, \ldots, m\}$. By a permutation $\tilde{f}^i = (\tilde{f}_1^i, \ldots, \tilde{f}_m^i)$, we construct a sample $\tilde{S}_i^r = \{(\alpha_{f_1^i}, \mathbf{x}_1), \ldots, (\alpha_{f_m^i}, \mathbf{x}_m)\}$. As a result, we obtain a set of random samples $\{\tilde{S}_1^r, \ldots, \tilde{S}_N^r\}$. As the significance measure of a simple one-dimensional regu-

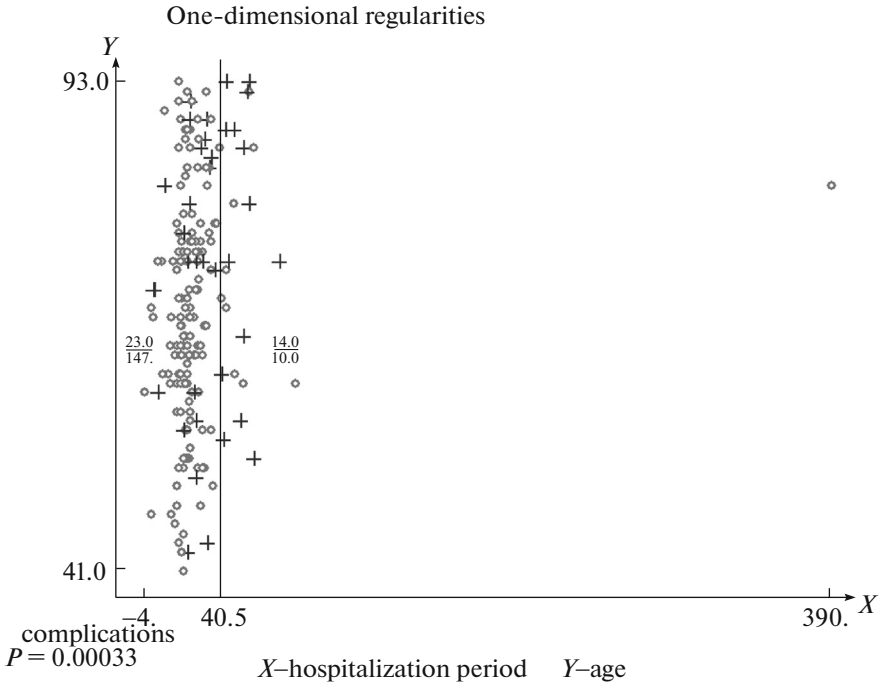

Fig. 1. Regularity relating the emergence of complications to the duration of the hospitalization period.

larity, we use a fraction of samples from $\{\tilde{S}_1^r,\ldots,\tilde{S}_N^r\}$ for which $F_I(\tilde{S}_*^r,R)$ is not less than $F_I(\tilde{S}_0,R)$. In [1], it is shown that the procedure described estimates the probability of exceeding $F_I(\tilde{S}_0,R)$ provided that the null hypothesis on the independence of the goal variable of the features of X is valid and the condition is satisfied that the marginal distributions of the goal variable and the features of X correspond to empirical distributions for the original sample $\tilde{S}_0$.

Verification of two-dimensional regularities. To verify a two-dimensional regularity formed by variables $X_{i'}$ and $X_{i''}$, we checked the null hypothesis on the exhausting description of data by simple one-dimensional regularities with a single threshold for the same variables. In this case, the threshold values for simple regularities are taken to be equal to the threshold values of the two-dimensional regularity. To verify if the variable $X_{i'}$ enters the two-dimensional regularity, we generate a set of permutations $\{\tilde{f}^1,\ldots,\tilde{f}^N\}$ of natural numbers from the set $\{1,\ldots,m\}$ by a random number generator. However, mutual permutations between the numbers of objects belonging to different segments of optimal partitioning with respect to the variable $X_{i''}$ were completely eliminated from the collection of permutations. As the significance measure of the appearance of the variable $X_{i'}$ in a two-dimensional regularity, we used a fraction of samples from $\{\tilde{S}_1^r,\ldots,\tilde{S}_N^r\}$ for which $F_I(\tilde{S}_*^r,R)$ is not less than $F_I(\tilde{S}_0,R)$. The appearance of the variable $X_{i'}$ in a two-dimensional regularity was verified similarly. This procedure was justified in [8, 9]. Note that permutation tests do not require assumptions on the types of probability distributions of data and can be applied for arbitrary sample sizes.

One of advantages of the one-dimensional regularity method is that it allows one to detect nonlinear phenomena. Another advantage is a convenient visual form of representation of the results of analysis in the form of two-dimensional diagrams.

Figure 1 graphically demonstrates one-dimensional and two-dimensional regularities revealed at the level of significance. The cases with complications are indicated by + and the cases without complications, by **o**. Each quadrant of partition contains a fraction that indicates the number of cases of each type: (number of +) / (number of **o**).

One-dimensional regularities. Quadrant I (to the left of the boundary) contains 23 cases with complications and 147 cases without complications. That is, the fraction of cases with complications is less than 15%. Quadrant I (to the right of the boundary) contains 14 cases with complications and only 10 cases without complications. That is, the fraction of cases with complications is greater than 58%. When estimating the significance of this regularity, the value of the partition quality functional (7) is exceeded only for one case of 3000 permutations. That is, the estimate for the p-value was equal to 0.00033. When the value of the partition quality functional (7) is not exceeded for any of the permutations, we will assume that $p < 0.00033$.

Table 1. All one-dimensional regularities with a single boundary

No.	Indicators	Quality functional	Boundaries	Significance	Distribution in quadrants	
1	Time between the symptoms onset and hospitalization	40.2	2.500	$p < 0.00033$	23/151	14/5
2	Length of stay	27.21	40.5	0.00033	23/147	14/10
3	Age	10.96	79.5	0.032	21/129	16/28
4	Atrial fibrillation	6.5	–	0.01	18/111	19/46
5	Severe cardiac insufficiency	10.14	–	0.005	32/154	5/3
6	Assessment by the Rankin scale on admission	36.73	4.5	$p < 0.00033$	12/121	20/18
7	Assessment by the Glasgow scale on admission	12.5	14.5	0.003	16/14	2/23

The complete list of one-dimensional regularities is presented in Table 1. The column "Indicator" shows an indicator for which a regularity is revealed. The column "Quality functional" shows the value of the functional $F_I(\tilde{S}_0, R)$ obtained for the optimal boundary shown in the column "Boundary." For two categorical features indicating the presence or absence of severe cardiac insufficiency and auricular fibrillation, the boundary is not set. The column "significance" presents the p-value calculated by a variant of the permutation test for the simplest one-dimensional models. The cells denoted "Distribution in quadrants" in two columns present the fraction (number of **+**)/(number of **o**). Here the arrangement of cells fully corresponds to the arrangement of quadrants in Fig. 1.

Table 1 shows that the relation between the frequency of emergence complications to the period of time between the onset of symptoms and hospitalization is the most reliable and informative parameter. For late hospitalization time (encoded "3"), complications arose in 14 cases, while complications did not arise only in 5 cases. Note that the assessment of the state of patients by the Rankin scale on admission also shows high reliability and informativeness. Figure 2 represents the regularity between the emergence of complications and the length of stay and with the assessment of the state of patients by the Rankin scale on admission.

Two-dimensional regularities. Altogether, we revealed four two-dimensional regularities between the frequency of emergence of complications and paired combinations of indicators. Figure 2 demonstrates the regularity relating the emergence of complications and the duration of hospitalization period to the estimate of the state of patients by the Rankin scale on admission. The symbols "+" denote the cases with complications, and "o" denote the cases without complications.

The maximum separation of groups with and without complications is reached when a boundary of 44.5 days is used for the indicator of the period of hospitalization and a boundary of 4.5, for the estimate of states of patients by the Rankin scale on admission.

The diagram shows that there is a clear relation between the emergence of complications and the estimate of the state by the Rankin scale for a hospitalization period of at most 44 days:

—in quadrant I, when estimating by the Rankin scale equal to 5 and a hospitalization period of less than 45 days, 16 cases with complications, and 15 cases without complications are observed (i.e., the fraction of cases with complications is 52%);

—in quadrant II, when estimating by the Rankin scale equal to 5 and a hospitalization period of less than 45 days, four cases with complications and three cases without complications are observed (i.e., the fraction of cases with complications is 57%);

—in quadrant III, when estimating by the Rankin scale less than 5 and a hospitalization period of less than 45 days, five cases with complications and four cases without complications are observed (i.e., the fraction of cases with complications is 55.5%);

—in quadrant IV, when estimating by the Rankin scale less than 5, seven cases with complications and 117 cases without complications are observed (i.e., the fraction of cases with complications is 6%).

Note also that there is a relationship between the frequency of emergence complications and the duration of the hospitalization period estimated by the Rankin scale less than 5: in quadrant III, the fraction of cases with complications is much greater than the fraction of cases with complications in quadrant IV.

All two-dimensional regularities relating the emergence complications to different biological and clinical indicators are shown in Table 2, which has the following structure. The first column numbers regularities. The column "Indicators" presents two factors related to a regularity. The factor in the upper cell corresponds to the "X" axis, while the factor in the lower cell, to the "Y" axis. The column "Quality functional" shows the value of the functional $F_I(\tilde{S}_0, R)(7)$ for optimal partitioning. The column "Boundaries" indicates the optimal boundaries of partitions. The column "Significance" shows p-values that estimate the reliability of partitioning on indicators. The upper cell presents the reliability with respect to the indicator

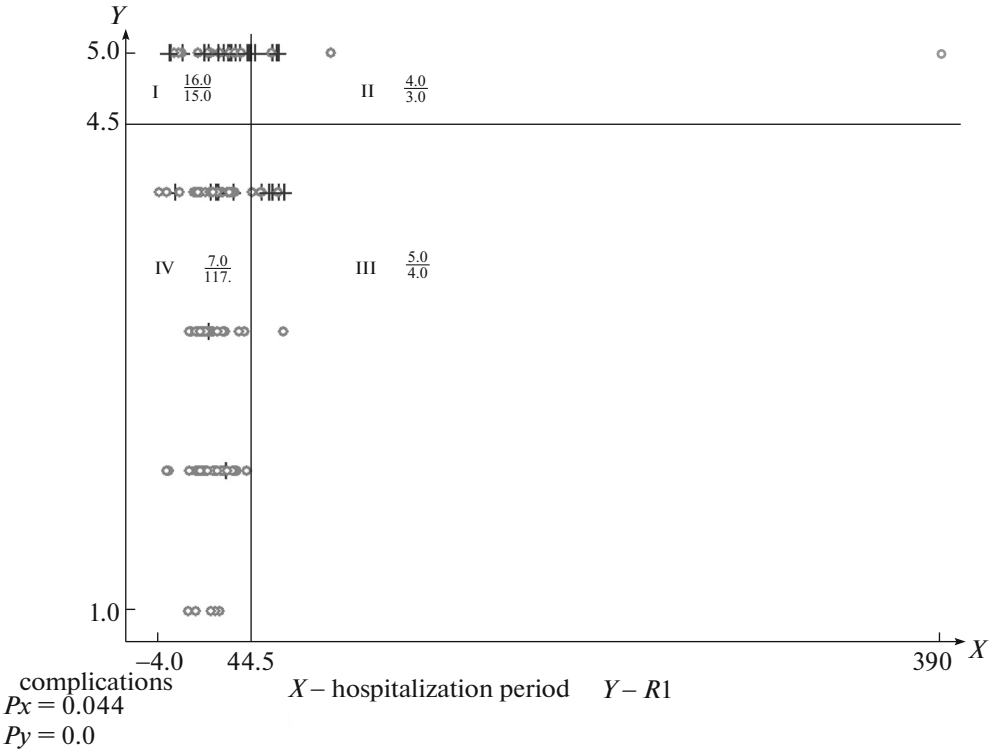

Fig. 2. Regularity relating the emergence of complications to the duration of the hospitalization period and the estimate of the state of a patient by the Rankin scale on admission.

corresponding to the factor plotted on the X axis. The lower cell presents the reliability corresponding to the factor plotted on the Y axis. The cells contained in two columns and titled "Distribution in quadrants" presents the fractions (number of +)/(number of **o**). Here the arrangement of cells fully corresponds to the arrangement of quadrants in Fig. 2.

Table 2 shows that the most informative regularities are regularities 1 and 2. Regularity 1 relates the frequency of emergence of complications to the time between the symptoms onset and hospitalization and assessment by the Rankin scale. An unfavorable combination of these indicators was observed in 14 patients. Nine of these patients had complications. A favorable combination of indicators was observed in 129 patients. Complications appeared also in nine patients. The reliability of this regularity was estimated by the above-described variant of the permutation test for verifying two-dimensional regularities at a level of 0.001 by the indicator "Time between the symptoms onset and hospitalization" and at a level of $p < 0.00033$ by the indicator "Assessment by the Rankin scale on admission."

Multifactor prediction of complications by a set of indicators. Multifactor prediction of complications was performed with the use of seven indicators shown

Table 2. All two-dimensional regularities that are reliable simultaneously with respect to both indicators

No.	Indicators	Quality functional	Boundaries	Significance	Distribution in quadrants	
1	Time between the symptoms onset and hospitalization	54.9	2.500	0.001	11/13	9/5
	Rankin score on admission		4.500	$p < 0.00033$	9/120	3/0
2	Time between the symptoms onset and hospitalization	56.7	2.500	$p < 0.00033$	8/10	6/0
	Length of stay		40.500	0.006	15/141	8/5
3	Length of stay	50.5	44.5	0.044	16/15	4/3
	Rankin score on admission		4.5	$p < 0.0003$	7/117	5/4
4	Age	57.972	77.500	0.044	4/3	1/0
	Severe heart failure	–	–	0.02	7/117	5/4

Table 3

	Group without complications	Group with complications	AUC
MSWS (5, 0.1)	129 (82.2%)	28 (75.7%)	0.847
SVM (Gauss 10)	151 (96.2%)	14 (37.8%)	0.826

in Table 1. The results for the method of multimodel statistically weighted syndromes (MSWS) [10] and the known method of support vectors (SVM) are presented in Table 3. The accuracy of recognition was estimated by the sliding control method. The cells show the numbers and percentages of correctly recognized cases in groups with and without complications.

2. METHODS OF ANALYSIS OF MEDICAL INFORMATION

In this section, we describe mathematical methods for the analysis of medical information that have been applied to the solution of some problems of analysis of medical data.

2.1. ROC Analysis of Precedent Data

ROC analysis is used in medicine when considering the applicability of a certain recognition model for classifying precedent data. Consider a classification problem for two classes and keep to medical terminology. The first class by training sample is p "positive" patients $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ (a disease is confirmed) and r "negative" patients $\mathbf{x}_{p+1}, \mathbf{x}_{p+2}, \ldots, \mathbf{x}_{p+r}$ (there is no disease). Suppose given a one-parameter family of recognition algorithms $A(t)$, $-\infty < t < \infty$. Classification by an algorithm $A(t)$ of the objects of the training sample may give four variants of solutions (truepositive, falsepositive, falsenegative, and truenegative) and, accordingly, four groups of objects. The following parameters are used to estimate the results of classification: "fraction of correctly classified positive objects" (true positive rate, TPR) and "fraction of falsely classified positive objects with respect to negative objects" (false positive rate, FPR).

$$TPR = \frac{|TP|}{p}, \tag{8}$$

$$FPR = \frac{|FP|}{r}. \tag{9}$$

In practical recognition with two classes, it is important to find an algorithm that makes a small number of errors on objects of both the first and second classes. Assume that the rates TPR and FPR monotonically decrease with t. If we consider a graph with axes TPR and FPR, assign the point (1,1) to the point $t \to -\infty$, and the point (0,0) to the point $t \to +\infty$, then we obtain a parametric curve, called a ROC curve, in the axes $y = TPR$ and $x = FPR$. The efficiency of the classification model is estimated by the area under the curve (the AUC characteristic); the greater the area under a curve, the better. The AUC characteristic has important statistical properties.

Suppose that, in the case of two classes, estimates for the classes $\Gamma_1(\mathbf{x}_i), \Gamma_2(\mathbf{x}_i)$ are calculated in the sliding control mode for an arbitrary recognition model and a decision rule of recognition by the maximum estimate is used. Take a one-parameter set of decision rules: if $\Gamma_1(\mathbf{x}_i) - \Gamma_2(\mathbf{x}_i) > t$, then $\mathbf{x}_i \in K_1$, where $-\infty < t < +\infty$. In this case, we fall "within the requirements of ROC analysis." Let us arrange all possible different differences of estimates $\Delta_j = \Gamma_1(\mathbf{x}_i) - \Gamma_2(\mathbf{x}_i)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, \tau$, $j = j(i)$ in decreasing order $\Delta_{j+1} < \Delta_j$. Each possible pair $\{\Gamma_1(\mathbf{x}_i), \Gamma_2(\mathbf{x}_i)\}$ corresponds to a point $\{y = TPR, x = FPR\}$. Connecting neighboring points, we obtain a ROC curve for a given family of algorithms and can estimate the quality of the model.

2.2. Algorithm for Smoothing an Empirical Distribution to Remove Random Outliers

Medical data often contain records that are essentially different in the set of parameters from average values for a given type of patients. This may be due to either hardware malfunction, incorrect use of measurement methods, errors in recording the results, etc. An important problem in the mathematical processing of data is screening out such outliers in order to construct and train correct algorithms for analysis and prediction. For a small size of the training sample $X \subset R^N$, each object $x \in X$ makes a noticeable contribution to the formation of essential data clusters, including logical regularities of classes. It is usually assumed that an object x has its own attraction domain; today, a large number of approaches are known in which the geometric shape of the attraction domain is modeled by one or other method: balls, hyperparallelepipeds, Gaussian caps, etc. Such heuristic models allow one to compensate for the deficit of training information under the assumption that the classes are compact. In the opposite situation, when a sample has a large size, the application of an appropriate model allows one to optimize a solution, in particular, allows one to reduce the effects of "overfitting."

Below we consider the problem of screening out outlying objects ("outliers") from the training sample, which may arise in case of both small and large samples. Here the instrument is the restriction of the attraction domain of outliers. The approach involves a simple model of the attraction domain in the form of a uniformly filled hyperparallelepiped with center at x, volume $\prod_{n=1}^{N}(2a_n + 1)$, and density $1/\prod_{n=1}^{N}(2a_n + 1)$, where a_n is the half-size of the smoothing interval

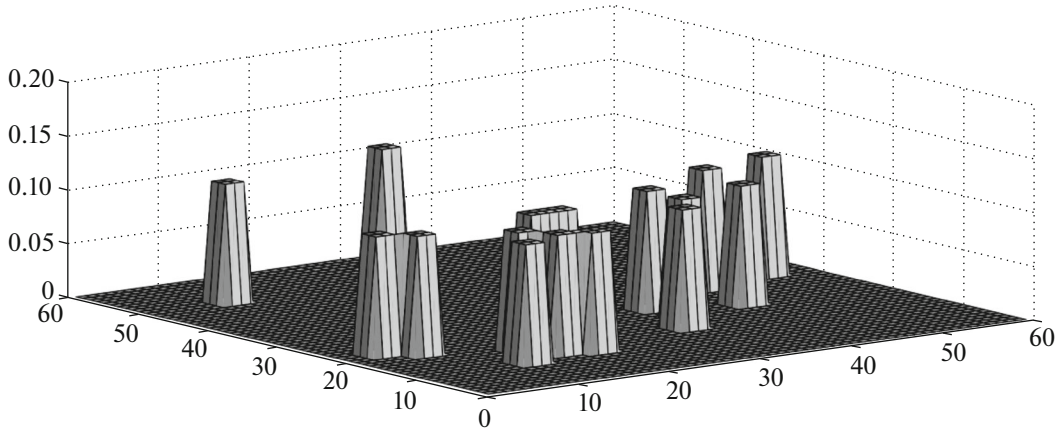

Fig. 3. Model example of a sample with two outliers after two stages of smoothing.

along the axis n, $n = 1,2,...,N$. As a result of smoothing, the object x is uniformly represented at all points of the hyperparallelepiped. The fact that a new object falls into a given domain indicates that it belongs to an appropriate class. It is suggested that one should take into account this effect only if the total generated density is higher than a given threshold.

We will iterate the smoothing process; in this case, each descendant of the central object (a point with nonzero generated density) obtained at the preceding stage is considered as a new attraction center. Under s smoothing operations, the attraction domain becomes a hyperparallelepiped with volume $\prod_{n=1}^{N}(2sa_n + 1)$, which is now uniformly filled with the total generated density. It is easily seen that the distribution inside the hyperparallelepiped is rapidly normalized as the parameter s increases, and, already for $s > 3$, its approximation by a Gaussian may sometimes be acceptable for constructing numerical estimates for classes. As the domain $\prod_{n=1}^{N}(2sa_n + 1)$ expands, close objects start to join their attraction domains, which is manifested in the summation of estimates for classes. For isolated objects, the generated density rapidly decreases, including the maximum at the central point, and, for an appropriate choice of the threshold of estimates, all isolated outliers can be ruled out. The process of normalization of the generated density for a single point is well studied, and the necessary threshold can be calculated in advance with high accuracy.

There are two variants of the algorithm. The first, simpler variant allows one to construct sets of LRCs in the explicit form. The second variant has a structure analogous to the standard scheme of the algorithm of k nearest neighbors. It has enhanced capabilities for the reconstruction of the generated density in smoothed clusters, and, hence, for constructing estimates for classes. Below, we present only the second variant in detail, because the first variant is easily implemented on its basis.

Figure 3 demonstrates a model example of a sample with two outliers. At the initial stages of smoothing, close objects yet do not fall into the attraction domain of each other. In the attraction domain of each object, the generated density is appreciably lower than the initial density 1. Figure 4 illustrates the effect of neighbors on the value of the total density after several iterations.

Let x_n^m, $n = 1, \ldots, N$, $m = 1, \ldots, M$, be a training table and K^l, $l = 1, \ldots, L$, be its layout with respect to classes. The characteristic function $k(m) = \{l, \text{if } m \in K^l\}$ yields the number of a class by the number m of an object in the table. For the object x^0 to be recognized, we will seek a set $list(p)$ of close points of the sample (i.e., nearest neighbors) of the vector x^0, that are located within the hyperparallelepiped $[x_n^0 - sn, x_n^0 + sn]$, $n = 1,\ldots,N$, $m = 1,\ldots,M$, with volume $\prod_{n=1}^{N}(2sa_n + 1)$, which arises as a result of application of s smoothing operations, and then, using these points, construct estimates for the density $r(l)$ and votes for classes for the object x^0. In contrasts to the ordinary method of k nearest neighbors, here one should not calculate the distance immediately during the choice. One should just find all the points of the sample that fall within the hyperparallelepiped; the fact whether the points fall within the hyperparallelepiped can be checked independently with respect to each of the axes n, $n = 1,2,...,N$, for all the points of the sample x_n^m, $n = 1, \ldots, N$, $m = 1, \ldots, M$. Actually, the distances and contributions of neighbors to the total generated density will be calculated at the final stage.

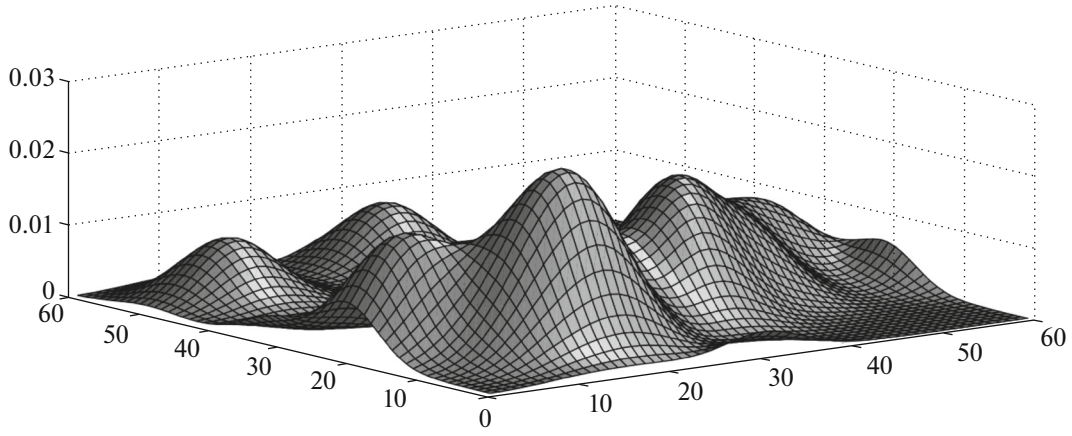

Fig. 4. Result of several iterations of the smoothing procedure.

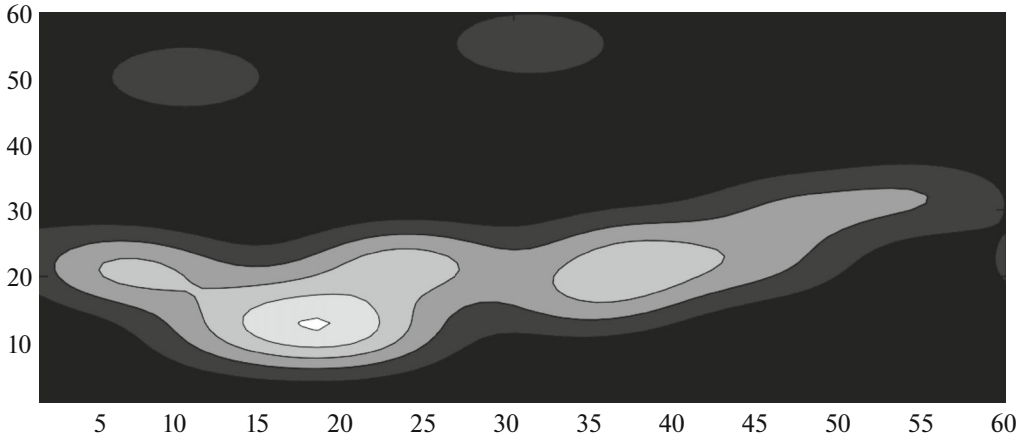

Fig. 5. Total density levels for Fig. 4. Leading levels can be used when constructing decision rule that does not take into account the effect of outliers.

For these reasons, one can start the test from any axis, say, from the first, $n = 1$, and, on each subsequent axis, check only those points that withstood the proximity test on the previous axes. Having constructed the set $list(p)$, we find all the points of the sample that extend at least minimal attraction to the object x^0. For the points of some class l, $l = 1,\ldots, L$, an estimate for the total probability density $r(l)$ can be obtained without binding to the parameter s, for example, when one uses a library Gaussian function in the computations. In this case, one should appropriately correct the screening threshold for outliers. One can further deal with the estimate vector $r(l)$ constructed to make a decision on the object x^0, for example, with the use of the maximum likelihood criterion, etc.

Figure 5 demonstrates the levels of total generated density from Fig. 4. The first level serves as a screening threshold for two outliers located in the upper part of Fig. 5.

3. PREPARATION OF MEDICAL DATA FOR THE ANALYSIS AND CLASSIFICATION

In cardiology, neurology, oncology, surgery, including neurosurgery, unified standard forms of reports have been developed on the basis of information from registers. Quality registers are designed for a systematic data acquisition and the application of instruments for improving the quality of healthcare; they can be classified into two categories: disease registers and intervention registers. Quality registers differ from other clinical registers in the existence of special tools that are used in combination with the systematic data acquisition and are aimed at improving the healthcare quality. The tools of support of decision-making analyze the structured data on a patient introduced into the register and form treatment recommendations on the basis of clinical instructions. During 2013–2014, registers of four directions were introduced with the formation of report templates at the Medical Center. The register for percutaneous cor-

onary interventions (PCIs, balloon angioplasty and/or stenting of coronary arteries) contains data on 288 patients, of which 137 patients were subjected to planned PCIs and 150 patients were subjected to emergency PCIs. The register contains 230 indicators, the report on coronary interventions contains 7 sections: implementation of clinical protocols, demographic indicators, characteristics of patients with PCIs, preprocedural state for planned PCIs, preprocedural state for PCIs in case of acute coronary syndrome (ACS) without ST elevation, specific features of the procedure, and postoperative indicators. The register of the acute cerebrovascular accident (ACVA) is represented by three types: ischemic stroke, hemorrhagic stroke, and transitory ischemic attack (261 patients had ACVA during 2012–2014, of which 197 patients had an ischemic stroke, 25 a hemorrhagic stroke, and 39 a transitory ischemic attack). The register of ACVA is formed of 240 indicators, and a report on each type of ACVA consists of six sections. By an example of an ischemic stroke, we can represent the contents of the sections: demographic indicators, indicators at prehospital stage, main risk factors, and the estimate of clinical data and the results of examinations (during the first day) during hospitalization and at discharge from the hospital. The register of general surgery includes 3 nosological forms by which a decision is made on surgery: cholecystitis, appendicitis, and inguinal hernia. The register contains data on 403 patients operated during 2013–2014: cholecystectomy (214), appendectomy (67), and herniotomy (122). The register contains 260–240 indicators for each patient, taking into account a specific character of pathology. Reports are formed automatically, separately for each nosology according to the following sections: demographic features, estimate of the condition of a patient before operation, the hospital stage, and the audit of the healthcare quality. Oncological registers include prostate cancer (104 patients), gastric cancer (94), renal cancer (64), and pancreas cancer (11).

The scope of indicators ranges within 316–150. The description of sections is made by an example of CPG: demographic data, regular medical check-up, diagnostics, initial treatment, local recurrence, remote metastases, hormone-resistant CPG, and outcomes. The patient's condition and the general and recurrence-free survival rate are evaluated, and the causes of death and the presence of bone fractures are characterized. Electronic forms are developed for all registers: electronic registration form, protocol of observance of clinical recommendations depending on the stage of a disease and individual risk factors, an electronic form for the audit of the results of treatment and clinical outcomes, and an outpatient form of regular medical check-up for assessing remote results.

The forms are filled for all patients. The stages or the severity of diseases and the risk of complications are calculated by selected calculators and scales. Key indicators influencing the outcomes of the diseases are selected for further analysis and the development of a model for assessing the results of treatment. Unified standard forms of reports are developed in all directions for all the above-listed sections on coronary interventions, oncology, surgery, and neurology.

As a result of a three-year registry maintenance on a stroke, coronary interventions, five oncological diseases, and three surgical interventions, data were collected and analyzed that are necessary for predicting clinically significant outcomes (lethal outcomes and complications and functional outcomes of diseases). One of important applied aspects of the construction of prediction models has been taking into account the indicators studied for estimating the results of treatment and carrying out a comparative analysis of the activity indicators both in the same hospital in dynamics and in different hospitals (benchmarking). A necessary condition for solving this problem is the specification of a model that can most accurately predict an outcome using the whole array of statistically and clinically significant risk factors of patients. Traditionally, the solution of such problems involves logistic regression. However, the accuracy of traditional statistical methods is quite limited for small samples. For example, the analysis of the dynamics of lethal outcomes at the Medical Center for four years during 2012–2015 yielded the following values of stationary lethality according to the register data: ischemic stroke: 6–5–7–14% (3/49; 3/58; 6/86; 7/51), hemorrhagic stroke: 17–0–10–17% (2/12; 0/10; 1/10; 1/6), myocardial infarction: 8–0–12–8% (6/78; 0/68; 8/67; 8/98), prostate cancer: 0–0–9–3% (0/46; 0/32; 3/34; 1/37), respectively, which highly complicates the interpretation of the data. Since the populations of patients in different hospitals may differ significantly (the dispersion may be greater than 50%), carrying out a comparative analysis and benchmarking without taking account the features of treated patients and correcting for these factors become senseless. The processing of the data obtained in the registers of the MC during 2012–2015 by traditional statistical methods has not allowed one to carry out an objective comparative analysis of the results of treatment due to small samplings. The analysis of the data with the use of recognition methods by precedents has allowed one to predict the development of lethal outcomes and complications at stroke, coronary interventions, and gastric cancer by the risk factors of a specific patient with high accuracy. The accuracy of the models was estimated by statistical (chi-square, S-statistics) methods. The standardized risk coefficient was calculated by the following formula: "event frequency with correction for risk = (actual frequency / predicted frequency) x normalized frequency according to international registers."

The use of the prediction models for quality indicators analysis helps to estimate the impact of individual risk factors on clinical outcomes and to calculate annual mortality in a general hospital adjusted to case mix.

4. RESULTS OF APPLICATION OF CLASSIFICATION MODELS TO RECOGNITION AND ANALYSIS IN CARDIOLOGY, NEUROLOGY, ONCOLOGY, AND NEUROSURGERY

We have analyzed the models of classification theory for processing medical information. In medicine, two classes of applied special problems are distinguished. The first class of medical problems is related to the estimate of a condition of a patient on admission in relation to some disease or the estimate of the result of treatment. Here an additional problem of introducing objective scales arises, which requires, first of all, large volumes of data; therefore, we considered the simplest binary case, when one class is interpreted as a "norm" or a "favorable state," and the second class, as an "outlier" or an "unfavorable state." It is a priori assumed that the class "norm" should be a "compact" set. Here a two-level scheme is created for solving the main problem. First, expert doctors form a training sample for two classes, then the sample is analyzed by recognition and clustering methods of the RECOGNITION system (first of all, by the methods of hierarchical grouping). From the first class, objects are removed on which many recognition algorithms commit errors and which are "outliers" of the class from the viewpoint of clustering the objects of the first class. As a result, a compact first class of favorable outcomes (estimates) and a second class of "outliers" are constructed. The second class of medical problems is related to the choice of an optimal treating method. We developed the following two-level scheme. For each method of treatment of a disease for the known number of methods, a table of feature descriptions of patients is given in which the patients are divided into two classes: patients for which the application of this method of treatment has been successful, and patients for which the application of this method has been unsuccessful. For a given training table, a prediction algorithm is constructed that gives a positive or negative outcome of a given type of treating for some new patient. Thus, for each table (i.e., method), a prediction algorithm is constructed for an appropriate method of treating. When choosing a treatment method for a patient, the doctor has the results of prediction for each method. If the results of prediction by the method chosen by the doctor in charge are negative, this is a reason for a more detailed examination of the patient. The present problem was solved by an example of the choice of a method of surgical treatment of degenerative diseases of the lumbar region of the backbone. We considered four types of surgery: percutaneous laser diskectomy, microdiskectomy, radio-frequency destruction of facet nerves, and spinal fixation. We considered the descriptions of 390 patients subjected to one of the above-mentioned methods. The choice of a method was made by the doctor. The number of correctly chosen methods (positive outcomes) was estimated over a year and amounted to 359 patients (92%). As a result of application of this man-machine method, 46 descriptions (11.8%) out of the initial 390 patients were eliminated. The accuracy of the algorithm on the remaining 344 patients (a part of patients for which doctor's opinion concerning the choice of the method did not contradict the corresponding algorithm) was 96.8%.

We have developed a numerical algorithm for assessing the quality of recognition models of the RECOGNITION system in the case of two classes, which is based on the use of the ROC analysis. We consider a one-parameter family of decision rules with respect to a single parameter: the difference between estimates is greater than or equal to the value of some quantity. The estimates for classes are calculated within an arbitrary model of the system. By the training data, a matrix of estimates is formed in the sliding-control mode for an arbitrary recognition model. A table is compiled whose rows correspond to different values of the difference of estimates. The rows of the table are arranged in decreasing order of the difference of estimates. As the source information, we used the matrices of estimates formed in the RECOGNITION system. For every row of the table, the parameters "fraction of correctly recognized objects of the first class" and "fraction of errors of the second class" are calculated. Thus, a sample of points is formed starting from the point (0,0) and ending with the point (1.1), which correspond to the ROC analysis curve. We calculate the area under the given curve, which is assumed to be the quality of the corresponding model. The algorithm was implemented in the form of a program in the C# language with graphic interface and has been successfully approved.

It is known that the method of sliding control calculates unbiased estimates for the accuracy of classification. The application of the sliding-control method in combination with the Chi-square method allows one to estimate the reliability of the predictive ability of a recognition method. When searching for empirical regularities, we applied the method of statistical verification, which is based on the use of permutation tests. A permutation test (PT) checks the null hypotheses on the independence of the goal variable Y of the prediction variables X by comparing the quality of the regularity obtained on the initial training sample with the results of search on random samples obtained from the initial one by random permutations of the positions of Y with respect to the fixed positions of the vectors of the variables X. In this case, by the quality of regularity is understood a functional that characterizes the accuracy of approximation of data by the mathematical model describing the regularity. Notice that a PT does not require any a priori assumptions about the type of distributions and can be applied for arbitrary sample sizes. PTs can be used not only for the verification of regularities, but also for statistically reliable estimate of the necessity to use a complex model instead of a simpler one. For example, with the use of

PTs, we proved the necessity of using piecewise linear models for describing the dependence of the content of parathyroid hormone in the organism on the content of vitamin D, which proves the existence of a threshold value of the content of vitamin D below which a compensatory production of parathyroid hormone starts [9].

We carried out an analysis of the problem of prediction of the quality of treatment of cardiological diseases depending on the qualification of doctors. As the main basic (classifying) feature, we took "the frequency of hospitalization of a patient" (the first class corresponds to one hospitalization a year or more, and the second class corresponds to the lack of hospitalization). The distribution on classes is as follows: 120 patients in the first and 210 patients in the second class. We used 34 features: age, sex of a patient, duration of a disease, annual frequency of crises, coronary heart disease, diabetes, heart failure, renal failure, height, weight, body mass index, systolic and diastolic blood pressure, blood glucose and cholesterol, smoking, ACE inhibitors, beta-blockers, diuretics, Ca antagonists, antithrombolic and hypolipidemic drugs, angiotensin II receptor blockers, alpha-blockers, imidazoline receptor agonists, complications, doctor's recommendation compliance, visit of an AH school by the patient, risk factor control, drugs acquisition, psychological support, as well as three features that are directly related to a doctor: qualification of the doctor (intern, the first or the highest qualification), visit of AH school by the doctor, doctor's practice experience. We solved optimization problems for various recognition models and constructed optimal algorithms. It turned out that the best methods are "voting by logical regularities of classes" (88.5% of correct answers on the training sample in the sliding-control mode), "neural network algorithm" (87.5%), and "method of support vectors" (91.5%). We found that the list of highly informative features includes the "length of service of the doctor," while the list of medium informative features includes the "qualification of the doctor." The feature "visit of AG schools by the doctor" turned out to be low informative.

We have analyzed the possibilities of application of classification algorithms in cardiology, oncology, neurosurgery, and neurology and developed optimal diagnostic and prediction algorithms.

In the field of neurology, we considered four problems: t1, t2, t3.dyn, and t4.dyn. The first two problems consisted in assessing patient's condition on admission; in problem t1, patients with lethal outcome were considered as an outlier ("unfavorable" state), whereas, in problem t2, patients with "complications" were considered as an outlier, which was pointed by an expert doctor. In problem t3.dyn, a "Functional outcome at discharge (by the Rankin scale)" was assessed, while the partition into classes was performed according to "positive" or "negative" dynamics. In problem t4.dyn "Dynamics of the neurological status (estimate by the NIHSS scale at discharge)," there were also two classes–patients with positive or negative dynamics. The initial features of neurological diseases are presented in Table 4.

Table 4. Features of neurological diseases

a1	Time since the onset of the first symptoms of a disease
a2	Record number
a3	Age
a4	Sex
a5	Smoking
a6	Atrial fibrillation
a7	Diabetes
a8	Arterial hypertension
a9	Dyslipidemia
a10	Anamnesis
a11	Surgical interventions on coronary arteries
a12	Concomitant diseases
a13	Replaced cardiac valves
a14	Carotid stenoses
a15	Estimate of clinical data by the Rankin scale
a16	Estimate by the Glasgow scale
a17	Operations during hospitalization
a18	Gastrostomy
a19	Tracheostomy
a20	Neurosurgical interventions
a21	Ureteric catheterization
a22	Nasogastric tube
a23	Prescriptions during hospitalization
a24	Oral anticoagulants
a25	Antihypertensive drugs
a26	Lipid loweringdrugs
a27	Glucose loweringdrugs
a28	Unfractionated or low-molecular heparin
a29	Pathogenetic variant of ischemic stroke
a30	Drug prophylaxis of thrombosis

Problem t1. The first class ("norm") included 112 patients and the second, 13. In this and the subsequent problems, we eliminated features that characterize the treatment of a patient (a17–a22) and the purely information feature a2. The best result was obtained by the logical algorithm "voting by deadlock tests" (DT): 90.4, 92.9, and 69.2% (the percentage of correct answers of the algorithm in the sliding control mode for the entire training table for objects of the first and second classes, respectively). The features a3, a5, a10–a13, a15, a16, a24, and a27 turned out to be the most informative ones. To calculate the logical regu-

larities of classes, we used an exact (combinatorial) and approximate (relaxation) algorithms. The best exact LCR a3 < 92.5 & 13.47 < a16 for the first class was satisfied on 106 training objects (out of 112 objects), the best LRC of the approximate method was (a3 <= 92)(1 < a15 < 5) {27, 0.723214} (the curly brackets give the number of the object on which the given regularity holds and the fraction of objects from the same class that satisfy this LRC). The best exact LRC 70.5 < a3 & a8 = 1 & a11 = 0 & a13 = 0 & 4 < a15 & a16 < 11.5 & a24 = 0 & a29 < 5 of the second class includes 10 objects (nos. 115, 116, 117, 118, 119, 120, 121, 122) out of 13. The above-mentioned regularities were statistically significant for the confidence level of 0.95. We found five and three exact regularities for the first and the second classes, respectively.

Problem t2. The first class included 99 patients and the second, 26. The best results were obtained also by the DT algorithm: 69.6, 69.7, and 69.2%. The features a3, a5, a6, a10–a13, a15, a16, a24, a27, and a28 turned out to be the most informative ones. The LRC (a3 <= 88.5)(0 <= a11 <= 0.5)(0< = a30 <= 0.5) {35, 0.55} is a statistically significant LRC of the first class by the approximate method. There were no statistically significant LRCs of the second class. The calculation by the exact algorithm showed that the best logical regularity of the first class is a3 < 88.5 &13.29 < a16, which is satisfied on 86 objects, and the best logical regularity of the second class is 52.5 < a3 < 90.5 & a5 =0 & 1.5 < a15 & 10.31< a16 < 13.29 & 0.5 < a25 & a29 < 2.05, which is satisfied on 19 objects. By the exact algorithm, we obtained seven and six LRCs for the first and the second classes, respectively.

Problem t3.dyn. Here the distribution by classes is as follows: 98 + 13. The features a2, a11, and a20 were eliminated (the last two, as features having almost constant values). The best algorithms in this problem are the DT 94.6, 95.9, and 84.6% and "voting by LRCs" 93.7, 94.9, and 84.6%; the most informative features are a3, a5, a10–a13, a15–a17, a21, a22, a24, and a27. The best LRCs on the first and the second classes are 1.5 < a1 & a3 < 92.49 & a12 < 0.5 & a17< 0.5 & a19 < 0.5 (includes 89 objects of the fist class) and 73.09 < a3 & 0.5 < a8 & a13 < 0.5 & a16 < 14.5 & 0.5 < a17 & 0.5 < a21 & 0.5 < a23 & a24 < 0.5 & 0.5 < a25 & 0 <= a29 < 4.5 (includes 9 objects of the second class), respectively. The approximate method found statistically significant LRCs of the first class (1.5 <= a1)(0 <= a12 <= 0.5)(a17 = 0) (91%), (1.5 <= a1)(0 <= a21 <= 0.5) (84%), and (0 = a21)(0 = a27) (79.6%), where the brackets show the percentage of objects satisfying this LRC. For the second class, there are no statistically significant LRCs. Note that if we eliminate incorrectly recognized objects from the training sample (considering these objects as "outliers"), then we obtain the following results on the remaining 102 objects: the accuracy of the algorithms is 96.1, 96.7, and 90%; the most informative features are a1, a3, a8, a12, a13, a16–a25, and a29; the exact LRC of the first class 0 < a1 & a3 < 92.49 & a12 < 0.5 & a17 < 0.5 & a18 < 0.5 & a19 < 0.5 & a22 < 0.5 satisfies 90 objects, and the exact LRC of the second class 73.08 < a3 & 0.5 < a8 & a13 < 0.5 & a16 < 14.5 & 0.5 < a17 & 0.5 < a21 & 0.5 < a23 & a24 < 0.5 & 0.5 < a25 & 0 < a29 < 4.5 satisfies 9 objects. Using the approximate method for searching LRCs, we obtained the following shortest logical descriptions of classes: (a12 = 0)(a17 = 0) {28, 0.97}V (a21 = 0) {29, 0.95} (for the first class) and (5.56 <= a16 <= 14.00)(a21 = 1) {31, 0.8} V (70.57 <= a3 <= 78.57)(a21 = 1) {37, 0.6} (for the second class). All LRCs are statistically significant. We eliminated the features 2 and 17–22. If we apply the decision rule $\mathbf{x} \Rightarrow K_1$ if $\Gamma_1(\mathbf{x}) > 0.530$, then we have the following results of the test algorithm: 90.1, 91.8, and 76.9%. The most informative features are a3–a7, a10–a15, and a24–a30 (treatment) and, in particular, a3, a12, a15, a27, and a28. According to the exact algorithm, the best regularity on the first class was a3 < 92.5 & 13.42 < a16 (it satisfied 93 objects), while on the second class, it was 70.5 < a3 & 0.5 < a8 & 0.5 < a9 & a13 < 0.5 & 1.5 < a15 & a16 < 13.42 & 0.5 < a23 & a24 < 0.5 & 0.5 < a25 & a29 < 2.02 (it satisfied 9 objects). By the approximate algorithm, we obtained the following statistically significant LRCs of the first and the second classes: (a3 <= 72)(0 <= a13 <= 0.25) {23, 0.56} and (a15 <= 4.20)(0 <= a28 <= 0.06) {30, 0.54} and (72 <= a3 <= 79.19)(1 = a4)(4.83 <= a15)(a24 <= 0.03) {33, 0.38}, respectively.

Problem t4.dyn. The representativeness of the classes was 95 and 15 objects, respectively. The best results were shown by the logical approaches to recognition, first of all, by the DT: 96.4, 97.9, and 86.7%. The most informative features are a1, a3, a12, a13, a16–a22, a24, and a27. The shortest logical description on the first class is (1.5 <= a1)(a12 = 0)(a17 = 0) {6, 0.89} V(1 <= a15 < 5)(a17 = 0)(a27 = 0) {15, 0.78} V(a3 <= 92.17)(2 < a15 <5) {31, 0.55}. We found the LRC of the second class (4.62 <= a16 <= 11.12) {51, 0.47}. All the LRCs given above are found by the approximate method and turn out to be statistically significant. We also found six and four LRCs of the first and second classes, respectively: 1.5 < a1 & a3 < 92.49 & a12 = 0 & a17 = 0 & a19 = 0 satisfies 85 objects, and 70.51 < a3 & a8 = 1 & a13 =1 & a16 < 14.5 & a17 = 1 & a21 = 1 & a24 = 1 & 0 < a29 < 5 satisfies 11 objects. We eliminated the features 2 and 17–22. If we apply the decision rule :$\mathbf{x} \Rightarrow K_1$ if $\Gamma_1(\mathbf{x}) > 0.550$, then we have the following results of the test algorithm: 86.4, 88.4, and 73.3%. The most informative features are a3–a7, a9–a15, a24, and a27–a29 and, in particular, a3, a15, and a27. According to the exact algorithm, the best regularity on the first class was a3 < 92.5 & 13.5 < a16 (satisfied 91 objects) and, on the second class, 70.5 < a3 & 0.5 < a8 & a11 < 0.5 & a13 < 0.5 & 1.5 < a15 & a16 < 11.5 & a24 < 0.5 & a29 < 4.5 (satisfied 11 objects). By the approximate algorithm, the following statistically sig-

nificant LRCs of the first and second classes were obtained: (a3 <= 69.5) {20, 0.48} and (a15 <= 4.21)(0 <= a28 <= 0.06) {23, 0.51}, and (71.74 <= a3 <= 82.87)(0 <= a7 <= 0.12)(0.99 <= a8)(4.68 <= a15) {33, 0.4}, (71.87 <= a3 <= 79.75)(1 <= a4 <= 1.06)(4.84 <= a15)(a24 <= 0.03) {46, 0.33}.

Prediction of the risk of hypertension. We studied the concentrations of angiotensin II and HLDF24 peptides, endothelin proteins, S100b, and autoantibodies to them, as well as the concentration of atrial natriuretic peptide (ANP) in blood serum in patients with different categories of "normal" arterial pressure and hypertension. We analyzed the relationship between the concentrations of the above factors in patients' blood and estimated the possibility of predicting the hypertension risk by the above-listed biomarkers with the use of the recognition method based on machine learning technology. The efficiency of prediction of the onset of hypertension was estimated with the use of the ROC analysis. The area under a ROC curve reached a value of 0.819. Thus, we have certainly proved the predictive ability of the above biomarkers as potential predictors of arterial hypertension. These biomarkers can be recommended for personified estimate of the risk of arterial hypertension development.

The analysis of the difference between biological statuses of patients that had TIA in the history. We compared the concentrations of S-100 and VEGF neuropeptides and antibodies to NR2 in the groups of patients that had transitory ischemic attack (TIA) or acute cerebrovascular accident (ACVA) in past medical history in relation to the set of biological and biochemical indicators. The database contained observations for 55 patients with ACVA and for 33 patients with TIA: clinical laboratory indicators; biochemical indicators; coagulogram factors, hormones, and content of microelements. The accuracy of recognition of patients with TIA and ACVA by the above-listed indicators was estimated in the sliding control mode with the use of the method of optimum valid partitioning (OVP). The efficiency of the prediction of the onset of hypertension was estimated with the use of the ROC analysis. The area under the ROC curve reached a value of 0.822.

In the field of cardiology, we considered two problems (with planned and emergency intervention).

In planned examination, the original system of features is given by the features of Table 5.

In the calculations, we removed the features a2 (number of the medical history) and features containing "almost" equal values of features (a14, a18–a20, a31, a36, and a38) or too many missing data (a25). Here the distribution by classes was 48 + 6. The results of recognition were (DT) 96.3–100–66.7%. The most informative features were a3, a28, and a30 and noninformative (with weight less than 0.1), a8, a11, a22, and a37. The approximate method did not reveal any statistically significant LRCs at a significance level of 0.9. Nevertheless, we found the most informative features: a3–a5, a12–a18, a20, a30, a32, a34, and a42. The regularity (54.6 <= a3)(a24 <= 6)(0 <= a28 <= 5.7) (1 = a30) satisfying 79% of training objects turned out to be a statistically significant LRC of the first class with a significance level of 0.75.

The best LRCs of the first and second classes by the exact algorithm were the predicates 50.8 < a3 & a24< 8 & a30 < 3 & a35 < 0.5 (100% of objects) and a3 < 77.9 & 24.9 < a4 < 38.3 & 0.5 < a7 & 0.5 < a8 & a11 < 1.5 & a12 < 2.5 & a13 < 0.5 & a15 < 0.5 & a17< 0.5 & a21 < 0.1 & 0.8 < a22 & a23 < 1.3 & 0.5 < a24 & a26 < 6.5 & 0.5 < a20 & a32 < 1.5 & a34 < 0.5 & 0.5 < a37 & a40 < 2.5 & a42 < 0.5 & a43 < 0.5 (83.3% or five objects).

Under emergency examination, the original system of features is given by the features of Table 6.

Here the distribution by classes was 53 + 11. The features a2, features with constant value (a13, a16, a18, a32, a39, and a40, for which at most two objects were allowed to have different values), and features with many missing data were removed. The recognition by different methods yielded close results: high accuracy for the recognition for the first class and equiprobable recognition for the second. In this case, a typical result is that of the test algorithm: 90.6, 100.0, 45.5%.

The most informative features were a3, a4, a14, a24, a25, and a33, while, noninformative features were a7, a22, a23, a27, and a30. By the approximate algorithm, we found statistically significant LRCs with a significance value 0.9 for the first class: (52.8 <= a3)(23.3 <= a4 <= 37.0)(a24 <= 3.1)(0 <= a25 <= 3.7)(1 <= a33 <= 1.2) (satisfied on 68% of objects of the first class). The exact algorithm yielded 4 + 1 LRCs by classes, and the best regularity of the first class 45.5 < a3 < 97.0 & 20.3 < a4 & a11< 3.5 & a25 < 3.5 & a27 < 2.2 was satisfied on 96.7% of objects of a class.

In the field of oncology, we also considered two problems: prediction of the state of a patient with prostate cancer and gastric cancer. Prediction of the state of a patient with prostate cancer. We considered a dichotomic recognition problem. The first class included patients with positive outcome, and the second class–patients with negative outcome. The training sample consisted of 81 + 15 descriptions of patients for which the system of features shown Table 7.

The distribution of objects by classes was 81 and 15. From the original table we removed features with the numbers 1, 2, 8, 10, 12, 13, 30, 40, 41, 43, and 46 (those that have many missing data, "almost" constant features, class-forming, and informative features). The features 7, 9, 14, 15, 28, 39, 42, 44, and 45 were informative, while 33, 37, and 38 were noninformative. The accuracy of recognition by different methods was approximately the same. For example, the "linear machine" method calculated the following values: 86.5, 91.4, and 60.0. We found 2 + 3 LRCs by the exact method; the best LRC for the first class was a4 <2.1 & a7 < 0.1 & a9 < 0.2 & a11 < 0.1 & a14 < 2.5 & a15 < 83.1 & a19 < 10.5 & a21 < 4.5 & a38 < 3.5 &

Table 5. System of features in the problem of cardiology with scheduled intervention

a1	Complications
a2	Record number
a3	Age
a4	BMI
a5	Smoking
a6	Diabetes
a7	Arterial hypertension
a8	Dyslipidemia
a9	Burdened familial history
a10	Myocardial infarction in the history
a11	Exertional angina
a12	Chronic heart failure
a13	TIA or noncapacitating stroke in the history
a14	Stenoses of carotid arteries
a15	Peripherical vascular disease in the history
a16	Stenting in the history
a17	CABS (coronary artery bypass surgery)
a18	Chronic kidney disease
a19	Present hemodialysis
a20	COPD
a21	Left ventricular ejection fraction before PCI
a22	Variation in the left ventricular ejection fraction
a23	Results of noninvasive test
a24	By the Mayo scales
a25	By the SYNTAX scale
a26	By the EuroSCORE scale
a27	Compliance of the final decision to clinical guidelines
a28	Parameters of noncompliance to clinical guidelines
a29	Waiting for stenting for more than 2 weeks
a30	PCI access
a31	Operation
a32	Stents
a33	Ratio of introduced to calculated volume of the contrast
a34	Creatinine rise after PCI
a35	Laboratory data after PCI
a36	Result of revascularization
a37	Aspirin
a38	Clopidogrel. Other antiaggregants
a39	Anticoagulants
a40	Type of blood supply
a41	Amount of target affected coronary arteries
a42	Restenosis of a stented segment
a43	Bifurcation

a39 < 2.5 & a42< 2 & a44 < 1.5 & a45 < 0.5 (satisfying 80 objects) and, for the second class, a4 < 2.1 & a7< 0.2 & a9 < 0.2 & a11 < 0.1 & a14 < 2.5 & a15 < 83.0 & a19 < 10.5 & a21< 4.5 & a38 < 3.5 & a39 < 2.5 & a42< 2 & a44 < 1.5 & a45 < 0.5 (satisfying 12 objects). By the approximate method for searching LRCs we calculated the set of statistically significant regularities. For example, a statistically significant LRC for the first class was (a9 <= 0.05)(a15 <= 71.18)(a18 <= 4.75)(0 <= a28 <= 0.25) (satisfying 60.5% objects) and, for the second class, (2.6875 <= a39), satisfying 46.7% objects of the second class.

Prediction of the state of a patient with gastric cancer was made with the use of the Table 8.

The features a3, a4, a5, a8, a9, a12, a15, a17, a22, a25, and a27 were considered to be the most informative ones. Other features were low informative. The best algorithm was the DT algorithm with the results of 89.2, 94.5, and 60.0%. The best statistically significant regularities by the approximate algorithm were (a4 <= 68.8)(1.8 <= a17) (with quality 0.78), and, with respect to the second class (6 <= a7)(0 <= a17 <= 1.6) (with quality 0.5). The exact algorithm yielded (for example) the best LRC of the second class – (which includes 8 objects: 1, 2, 5, 6, 7, 9, 10,) and the best LRC of the first class a3 < 1.5 & 36.86 < a4 < 89.03 & 0.12 < a5 & a9 < 2.5 & a15 < 5.5 & 0.5 < a17 & 1.33 < a26 & a27 < 2.5 (which is satisfied on 54 objects). In total, we found 2 + 2 = 4 LRCs. The distribution of objects by classes was 55 + 10, and the number of features was 19.

In the field of surgery, we considered a problem of estimate of the condition of a patient after surgical intervention for "cholecystectomy." Class 1 of successful states contained 170 objects, and class 2 of unsuccessful cases contained 16. The features are presented in Table 9.

The most informative features (with weight greater than 0.6) were a2–a4, a10, a12–a15, a17, a23, a26, a28, a29, and a34. The most noninformative features (with weight less than 0.1) were 31 and 32. By the first class, we found a statistically significant regularity (0 <= a4 <= 0.25)(31.86 <= a6)(a8 <= 33.64)(a10 <= 7.5)(31.25 <= a23 <= 105) with quality 0.58. There were no statistically significant regularities of the second class. The system had 33 features. A "linear machine" had the highest accuracy: 88.7, 92.4, 50.0%.

The exact algorithm found 8 + 4 = 12 LRCs. The best logical regularity of the first class 2 < 1.4 & a10 < 30.5 & a13 < 9 & a19 < 3.5 & a21 < 2.5 & a22< 3.5 & 22.5 < a23 & a26 < 4.5 & a30 < 2.5 & 24.66 <a32 satisfied 154 objects (90.6% of objects of the class), while the best LRC of the second class was 2 < 4 & a3 < 0.5 & 784.5 <a5 < 1509 & 30.86 < a6 < 79.25 & 23.82 < a8 < 38.33 & 0.5 < a9 < 12.5 & 3.5 < a10 < 11.5 & a12 < 1.5 & a13 < 3.5 & a14<1.02 & a17 < 2 & a19 < 3.5 & 1.5 < a20 < 3.5 & 1.5 < a21 < 2.5 & a22 < 3.5 & a23 < 177.5 & 15 < a24 < 192.5 & 11.5 < a25 < 12.03 & a28<

Table 6. System of features in the problem of cardiology with emergency intervention

a1	Complications
a2	Record number
a3	Age
a4	BMI
a5	Smoking
a6	Diabetes
a7	Arterial hypertension
a8	Dyslipidemia
a9	Burdened familial history
a10	Myocardial infarction in the history
a11	Chronic heart failure
a12	TIA or noncapacitating stroke in the history
a13	Stenoses of carotid arteries
a14	Peripherical vascular disease in the history
a15	Stenting in the history
a16	CABS in the history
a17	Chronic renal failure
a18	Present hemodyalisis
a19	COPD
a20	Left ventricular ejection fraction before PCI
a21	Left ventricular ejection fraction after PCI
a22	Variation in the left ventricular ejection fraction
a23	Risk by the TIMI (STEMI) scale
a24	Risk by the TIMI (NSTEMI) scale
a25	Risk by the CRUSADE scale
a26	Risk by the GRACE scale
a27	Class of cardiac insufficiency (Killip)
a28	Compliance of the final decision to clinical guidelines
a29	Time to hospitalization from the symptoms onset of ACS with ST elevation, h
a30	Time from admission to a PCI for ACS
a31	Noncompliance parameters
a32	Waiting for stenting for more than 2 weeks
a33	Access to heart
a34	Operation
a35	Stents
a36	Ratio of introduced to calculated volume of the contrast
a37	Laboratory data after PCI
a38	Result of revascularization
a39	Aspirin.
a40	Clopidogrel. Other antiaggregants
a41	Antocoagulants
a42	Type of blood supply
a43	Amount of affected coronary arteries (more than 70% for all, and more than 50% for the proximal segment of anterior interventricular branch and the left main coronary artery)
a44	Restenosis of a stented segment
a45	Bifurcation

Table 7. System of original features for prostate cancer

a1	Patient's ID
a2	State of a patient
a3	Pathological fractures
a4	General state by the ECOG scale
a5	Complications of the initial treatment
a6	Complications of the treatment of a local recurrence
a7	Maximal degree of neutropenia
a8	Maximal degree of thrompocytopenia
a9	Maximal degree off anemia
a10	Degree of neurotoxicity
a11	Degree of nephrotoxicity
a12	Infectuous complications requiring antibacterial treatment
a13	Cardiovascular complications
a14	Effect of complications on antitumor therapy
a15	Age
a16	Yeas since the diagnosis
a17	Histological type of a tumor + code (ICD morphological code)
a18	Value of the Gleason sum
a19	Description of the primary tumor (T)
a20	Metastases into regional lymph nodes
a21	Remote metastases
a22	Implemented version of the initial treatment
a23	Type of prostatectomy
a24	Correspondence of stages before and after operation
a25	Presence of tumor at the edge of resection
a26	Hormonal therapy against early stages (all variants)
a27	Adjuvant radiation therapy
a28	Radical radiation therapy
a29	Brachytherapy
a30	Cryotherapy
a31	Ultrasound destruction (HIFU)
a32	Hormonal therapy against early stages, beginning
a33	SOD (for all types of radiation therapy), g
a34	Local recurrence
a35	Detection of remote metastases
a36	Hormonal therapy against a disseminated process, beginning
a37	Dissemination of CPG. Preparations used
a38	Orchiectomy
a39	Symptomes of metastases in bones
a40	Establishing hormone-resistance
a41	Chemotherapy, start of the first line
a42	Hormone-resistant CPG. Preparations used
a43	Number of CT courses
a44	Bringing a patient into remission
a45	Need in palliative radiation therapy
a46	Use of narcotic analgesics, including Tramal

Table 8. System of original features for gastric cancer

a3	Clinically significant complications of the first line of treatment
a 4	Age
a 5	Years since the diagnosis
a 6	Histological type of a tumor
a 7	Description of primary tumor (T)
a 8	Metastases into regional lymph nodes (N)
a 9	Remote metastases
a 10	Degree of differentiation (G)
a 11	Where the first treatment was performed
a 12	Initial estimate of the resectability of the primary tumor
a 15	Primary surgical treatment
a 16	Correspondence of stages before and after operation
a 17	Presence of tumor at the edge of resection
a 19	Adjuvant therapy
a 20	Primary nonsurgical treatment
a 22	Presence of a local recurrence
a 25	Presence of remote metastases
a 26	Where disseminated cancer was treated
a 27	Surgical removal or local destruction of individual metastases

$1.01 \& a29 < 1.01 \& 0.5 < a31 < 10.5 \& 3 < a32 < 24.67 \& a34 < 0.5$, which satisfied 13 objects (81.2%).

CONCLUSIONS

In this paper, we have presented some practical methods of the mathematical theory of recognition and the results of their application to processing medical data. We have given the main definitions related to the logical regularities of classes, their search, processing, and solution of a classification problem. We have described an algorithm for searching elementary LRCs as linear functions of features. We have considered the problem of searching for LRCs for large training samples and a statistical estimate for their reliability. We have given the definitions of the shortest and minimal logical descriptions of classes as descriptions that are equivalent to the initial logical descriptions but have minimal complexity. Finally, we have presented the results of prediction of ACVA complications obtained by the method of optimum valid partitioning. The principle of ROC analysis and its implementation in the RECOGNITION system have been presented. We have given an algorithm for smoothing an empirical distribution to remove random outliers in medical information. We have considered the general structure of the main clinical registers on neurological, cardiological, surgical, and oncological diseases. We have presented the results of application of the models for searching for LRCs, the RECOGNITION system, and the methods of classification and prediction for some diseases (cardiological, neurological, oncological, and surgical interventions) by the data of the Medical Center of the Bank of Russia.

Table 9. System of original features for the cholecystitis resection

a2	Estimate of postoperative complications by the Clavien system
a3	Repeated operation
a4	Unplanned repeated hospitalization
a5	ID of a patient
a6	Age
a7	Sex
a8	BMI
a9	Days before operation
a10	Days after operation
a11	Admission
a12	Total number of surgical risk factors
a13	Therapeutic risk factors
a14	Functional state of a patient before operation
a15	Anomalies of blood tests
a16	Procedure emergency
a17	Character of the surgical intervention performed
a18	Qualification of the operating surgeon for planned operation
a19	The main type of anesthesia
a20	Type of operation by volume
a21	Classification of a wound
a22	Severity of ASA classification
a23	Duration of operation
a24	Duration of anesthesia
a25	Main operation RVU code
a26	Category of complexity of surgical intervention
a27	Access to abdomen during laparoscopy
a28	Preparation of a cystic artery
a29	Preparation of cystic ducts
a30	Cholecystectomy
a31	Waiting time for planned surgical intervention, days
a32	Waiting time for emergency surgical intervention, hours
a33	Compliance to recommended time of operation
a34	Noncompliance to the guidelines of preprocedural preparation of a patient

In forthcoming research, we plan to significantly extend the databases on the diseases considered and to create convenient numerical schemes for the analysis of these databases by the methods of data analysis and recognition theory.

REFERENCES

1. O. V. Senko and A. V. Kuznetsova, The optimal valid partitioning procedures, InterStat. http://interstat.statjournals.net/YEAR/2006/articles/0604002.pdf
2. A. N. Dmitriev, Yu. I. Zhuravlev, and F. P. Krendelev, "On mathematical classification principles for objects and phenomena," in *Discrete Analysis. Collection of Papers* (Sobolev Institute of Mathematics, Novosibirsk, 1966), issue 7, pp. 3–11 [in Russian].
3. Yu. I. Zhuravlev and V. V. Nikiforov, "Recognition algorithms based on estimations calculation," Kibernetika, No. 3, 1–11 (1971).
4. L. V. Baskakova and Yu. I. Zhuravlev, "Recognition algorithms model with representative sets and systems of reference sets," Zh. Vychisl. Mat. Mat. Fiz. **21** (5), 1264–1275 (1981).
5. V. V. Ryazanov, "Logical regularities in recognition problems (parametrical approach)," Zh. Vychisl. Mat. Mat. Fiz. **47** (10), 1793–1808 (2007).
6. Yu. I. Zhuravlev, V. V. Ryazanov, and O. V. Sen'ko, *Recognition. Mathematical Methods. Program System. Practical Applications* (Fazis, Moscow, 2006) [in Russian].
7. N. V. Kovshov, V. L. Moiseev, and V. V. Ryazanov, "Algorithms for searching logical regularities in recognition problems," Zh. Vychisl. Mat. Mat. Fiz. **48** (2), 329–344 (2008).
8. A. V. Kuznetsova, I. V. Kostomarova, and O. V. Sen'ko, "Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients," Pattern Recogn. Image Anal. **24** (1), 114–123 (2014).
9. O. V. Senko, D. S. Dzyba, E. A. Pigarova, L. Ya. Rozhinskaya, and A. V. Kuznetsova, "A method for evaluating validity of piecewise-linear models," in *Proc. Int. Conf. on Knowledge Discovery and Information Retrieval* (Rome, 2014), pp. 437–443.
10. O. Senko and A. Kuznetsova, "A recognition method based on collective decision making using systems of regularities of various types," Pattern Recogn. Image Anal. **20** (2), 152–162 (2010).
11. V. J. Bufalino, F. A. Masoudi, S. Stranne, et al., "The American Heart Association's recommendations for expanding the applications of existing and future clinical registries. A policy statement from the American Heart Association," Circulation **123**, 2167–2179 (American Heart Association Advocacy Coordinating Committee, 2011).

Translated by I. Nikitin

Yurii Ivanovich Zhuravlev. Born 1935. Graduated from the Moscow State University in 1957. Received doctoral degree in 1965, is Professor since 1967, and Academician of the Russian Academy of Sciences since 1992. Currently is Deputy Director of the Dorodnicyn Computing Centre, Russian Academy of Sciences, Chair at the Mathematics Department of the Russian Academy of Sciences, and Head of Chair at Moscow State University. Editor-in-Chief of *Pattern Recognition and Image Analysis*. Foreign member of the Spanish Royal Academy of Sciences, the National Academy of Sciences of Ukraine, and the European Academy of Sciences. Winner of the Lenin and Lomonosov Prizes. Scientific interests: mathematical logic; control systems theory; mathematical theory of pattern recognition, image analysis, and forecasting; operations research; and artificial intelligence.

Gerasim Igorevich Nazarenko. Born 1953. Graduated from the Kirov Military Medical Academy and the Mozhaiskii Military Engineering Academy (1979). Doctor of Science in medicine, professor, academician of the Russian Academy of Sciences, Honored Scientist, and Honored Doctor. Currently is the Head of the Institute of Modern Information Technologies in Medicine. Scientific interests: medical technological processes, development of fundamental principles of information technologies for clinical medicine, control of quality and safety of medical aid, control of a medical organization by the methods if implementation of innovative projects. Author of more than 450 papers, including 35 monographs and 2 patents. Awarded 14 State prizes.

Aleksandr Petrovich Vinogradov. Born 1951. MS degree in physics from the Applied Mathematics and Control Department of the Moscow Institute of Physics and Technology, 1974. PhD in mathematical cybernetics, 1978. Senior Researcher at the Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control," Russian Academy of Sciences. Author of about 70 scientific papers. Scientific interests: algebraic and geometrical methods in pattern recognition, image analysis and processing.

Aleksandr Aleksandrovich Dokukin. Born 1980. Graduated with honors from the Department of Computational Mathematics and Cybernetics, Moscow State University, in 2002. Received candidates degree in 2008. Currently is with the Dorodnicyn Computing Centre, Russian Academy of Sciences. Scientific interests: pattern recognition and data analysis. Author of 63 papers.

Natal'ya Nikolaevna Katerinochkina. Born 1945. Graduated from the Moscow State University in 1967. Received candidates degree in 1978. Currently is senior researcher at the Dorodnicyn Computing Centre, Russian Academy of Sciences. Scientific interests: discrete mathematics, discrete optimization, recognition theory, and data analysis. Author of 45 papers.

Elena Borisovna Kleimenova. Born 1963. Graduated from the Pirogov Russian National Research Medical University in 1986. Received candidates degree in 1991 and doctoral degree in 2009. Awarded the title Honorary Doctor of Russian Federation in 2006. Currently professor at the Chair of Clinical Farmacology at the Russian Academy of Postgraduate Education. Scientific interests: technology of evidence-based medicine, practical application an inculcation aspects, information resources in evidence-based medicine, automatization of clinical instructions, automatization of diagnostic and treatment process in a multispecialty hospital, inculcation of molecular genetic and cellular technologies in a multispecialty hospital, problems of quality and safety of medical aid, approaches to the control of risks in a multispecialty hospital, management of the quality of medicine, and process approach. Author of 50 publications.

Marina Vladimirovna Konstantinova. Born 1963. Graduated from the Pirogov Russian National Research Medical University in 1987. Received candidates degree in 1993. Currently is Head of Department of Neurology at the Medical Center of the Bank of Russia Scientific interests: acute cardiovascular attacks, conservative and small-invasive methods of treatment back pain, application of information technologies and mathematical methods to increasing the quality of medical aid to patients of neurological profile. Author of 15 scientific publications.

Vladimir Vasil'evich Ryazanov. Born 1950. Graduated from the Moscow Institute of Physics and Technology in 1973. Received candidates degree in 1977 and doctoral degree in 1994. Academician of the Russian Academy of Natural Sciences since 1998 and professor since 2008. Since 1976 has been with the Dorodnitsyn Computing Centre, Russian Academy of Sciences. Currently is Head of the Department of Methods of Classification and Analysis of Data at the Dorodnicyn Computing Centre. Scientific interests: optimization methods of recognition models, algorithms for searching for and processing logical regularities by precedents, mathematical recognition models based on voting by the sets of logical regularities of classes, committee synthesis of collective clusterings and construction of stable solutions in clustering problems, restoration of missing data, restoration of regressions by the sets of recognition algorithms, development of software classification systems, and solution of practical problems in medicine, engineering, chemistry an other fields. Author of 208 publications.

Andrei Mikhailovich Cherkashov. Born 1960. Graduated from the Pirogov Russian National Research Medical University in 1983. Received candidates degree in 1989 and doctoral degree in 2002. Awarded the title Honorary Doctor of Russian Federation in 2002. Currently is Chief Doctor at the Multispecialty Medical Center of the Bank of Russia. Scientific interests: vertebrology, small-invasive interventions in case of back pains, medical information systems, application of information technologies and mathematical methods to increasing the quality of medical aid. Author of about 80 scientific publications.

Oleg Valentinovich Sen'ko. Born 1957. Graduated from the Moscow Institute of Physics and Technology in 1981. Received candidates degree in 2007. Currently is leading scientist at the Federal Research Center "Informatics and Control," Russian Academy of Sciences. Scientific interests: methods of machine learning and intelligent data analysis and their practical application. Author of more than 100 publications.