

Complex Analysis of Medical Data with Data Mining Usage

**Nafisa Yusupova¹, Gyuzel Shakhmametova¹,
Rustem Zulkarneev²**

¹Ufa State Aviation Technical University, Computer Science & Robotics Dept.,
Karl Marks str. 12, 450008 Ufa, Russia, yussupova@ugatu.ac.ru, siit@ugatu.su

²Bashkir State Medical University, Faculty of General Medicine, Lenin str. 3,
450008 Ufa, Russia, pvb@bashgmu.ru

Abstract: the new approach to the medical, in particular, the toxicological data analysis is considered. For the data processing multilevel system realization, the three-stage technique for data analysis with data mining usage is offered. The results of the research are discussed.

Keywords: medical data processing; data mining; complex analysis

1 Introduction

The most important field of modern IT application for medical purposes is processing data for solving specific tasks in such areas as diagnostics, treatment and prevention of diseases. It is of paramount importance especially for those medical professionals who must manage particularly vast amounts of entrance information or execute a complex algorithm of data processing which at times represents serious difficulties for the decision-maker [1]. The importance of this task increases with the avalanche-like accumulation of information due to the improvement of technologies for collecting and storage of information today. Proceeding from understanding the need for high-quality changes in the approaches to processing of the growing volumes of medical information, in recent years the increasing development is gained by methods of the data analysis which include statistical methods as well as methods of data mining [2]. Introduction of information technologies in medical practice allows to radically turn around a situation though there is still a set of problems that need to be solved by joint efforts of information scientists and physicians.

The main idea of research consists in the realization of the new approach to the analysis of medical, in particular, toxicological data. In this article, the technique

for the analysis of medical data is offered and examples of the analysis of the results yielded with interpretation on the example of the toxicological data collected in the Republic of Bashkortostan for 2015-2016 are given. For organizing a multilevel system of information processing, it is offered to create a technique of the three-stage data analysis that allows the researcher to gain an impression about the structure of data, to understand the main regularities, to take new, earlier unknown knowledge on the basis of the considered selection, and also to increase the efficiency of the information analysis process.

2 State of Art

Today, one of the most popular medical data in the field of toxicology analysis methods is the elementary visual analysis utilizing diagrams and charts. In general, for the medical data study, statistical and intelligent analysis techniques are used. In [3], the medical data analysis methodology is considered, the research is aimed at increasing the efficiency of doctors'/clinical physicians' activities. The preliminary analysis was carried out using the charts of dispersion and density, and the applied analysis methods included correlation analysis, logistic regression, and the accidental trees nonparametric qualifier. In [4], several selection methods that may be used in medical studies with various scenarios and problems are considered. The main applied methods are sampling and randomization. A large number of researches is devoted to Data Mining applications for the hidden patterns recognition in the medical data analysis tasks. For example, in [5], the Data Mining use for obesity disease detection in children is discussed. The cluster analysis used for the definition of children's groups with similar results after the executed treatment is considered. The methods applied in the research are dispersion analysis, cluster analysis (K-averages method), and the limited search algorithm. In [6], the analysis of data on acute peroral poisonings based on the REACH data is considered, the main research direction is the separate chemicals impact definition on an organism. The data analysis is carried out by the following methods: training at examples, neural networks, and k-nearest neighbors. In [7], the research of acute exogenous poisonings in Altai Region during 1997-2013 is given. The research allows us to assess a toxicological situation in the region and includes the systematization of acute poisonings in a section of gender and age and social groups of the population. Research was carried out in MS Excel using data visual analysis methods. In [8], the research of possible risks in patients with acute alcohol poisoning is conducted using descriptive statistics and visual analysis methods.

All data analysis research considered by authors use a limited number of methods, i.e., only the visual analysis, or only the statistical methods, or only the Data mining methods. Today there is no complex technique for the medical

toxicological data analysis which could allow us to process data most fully and to get the maximum quantity of new knowledge and hidden patterns.

Medical data, from the point of view of the analyst, have the next features:

- the data are retrospective;
- as a rule, they are diverse and have quantitative and qualitative indicators;
- they are semi-structured;
- as a rule, such data do not submit to normal distribution.

In this regard, there are difficulties with processing and analysis of such data because of the heterogeneity of the data that demands the application of various methods and approaches for data analysis.

Several methods are used today for processing medical data, i.e., descriptive and inductive statistics; correlation, regression, multiple-factor discriminant, cluster analysis; method of artificial neural networks, and others. Comparative characteristics of these methods are provided in Table 1.

Table 1
Contains the result of comparing in pairs with the final result

Method	Goal	Advantages	Shortcomings
Descriptive statistics	Processing of empirical data, systematization, quantitative description by means of main statistics	Effective and rather easy way of data consideration and description; convenient way of information representation [9]	It does not make conclusions about population based on results of special cases research [10]
Inductive statistics	Check of statistical hypotheses for the law of distribution	Simplicity of method application [11]	Low level of reliability; considerable errors for small size samplings [12]
Correlation analysis	Detection of existence and level of communication between two and more variables for predicting possible value of one of them if another one is known	Possibility of creating new rules for interaction of functions and also an assessment of functions interaction [13]	The results may be used only in the immediate research field or in one close to it [12]
Regression analysis	Detection of dependence	It allows to define dependence (linear,	Application for processing of

	between independent variable and one or several dependent variables	nonlinear) quantitatively, in the form of a mathematical formula [14]	qualitative data is impossible [15]
Multiple-factor analysis	Detection of latent variables or the factors causing multiple correlative communications	Possibility of smaller number of data utilization therefore leading to more expedient model generation [15]	The subjectivity of results interpretation, complexity of the procedure; requires several cycles of conducting a procedure for obtaining qualitative result [16]
Discriminant analysis	Classification of objects, i.e., reference to one of several set groups (classes)	It allows to make the multidimensional analysis of data [15]	By improperly conducted research, the developed models will not work with new data [17]
Cluster analysis	Method of classification analysis; its purpose is splitting a set of the studied objects and signs into uniform groups	It does not impose restrictions for a type of the objects considered, allowing to work with large volumes of data and to classify objects by a number of signs [18]	Subjective interpretation of results [19]; at different introduction can give different results [20]
Neural networks	Generalization and allocation of hidden dependences between the input and output data	No need of knowledge formalization; orientation to parallel processing; possibility of multidimensional data and knowledge processing without increase in labor input [21]	Difficulties in explanation of neural network functioning results [22]; impossibility to guarantee repeatability and uniqueness of obtaining results [23]

Each of these methods has its advantages and shortcomings, and many of them have restrictions in the character of the analyzed data. The problem is that a single method may only solve a narrow task of the data analysis which is not enough for decision-making. Authors offer a complex technique for the analysis of medical data in the field of toxicology including methods of mathematical statistics as well as methods of data mining, allowing to carry out a comprehensive analysis of the data and to benefit from the largest possible amount of knowledge, interrelations, and patterns. The research novelty consists of the new approach to the toxicology data analysis which represents the complex three-stage analysis with visual,

statistical and Data Mining methods use and allows us to study the data comprehensively.

3 Data Analysis Stages

The technique of medical data analysis in the field of toxicology including complex data analysis and interpretation of results, and consisting of the next main stages is suggested as follows (Figure 1):

- 1) Primary statistical data analysis by the means of MS Excel, visualization of raw data for understanding their quantitative and qualitative structure, making hypotheses about patterns existence in data.
- 2) Statistical analysis of the data using multidimensional analysis and nonparametric methods for confirmation or denial of the hypotheses made at the first stage. This stage is also the stage of the "prospecting" analysis for the purpose of making new hypotheses and assumptions.
- 3) Data mining assumes search of hidden regularities and patterns in the data with utilizing Data mining, knowledge discovery, confirmation or denial of the hypotheses made at the previous stages.

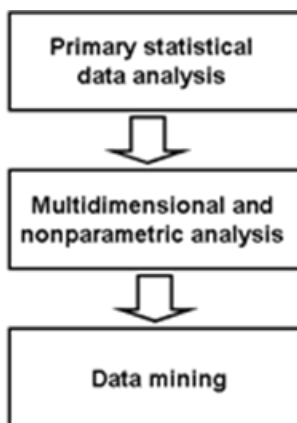


Figure 1

Stages of medical data analysis in the field of toxicology

For the analysis, the sampling of the toxicological data on the Republic of Bashkortostan for 2015-2016 has been taken. The volume of the database for the analysis is 6 338 records. The structure of the data is provided in Table 2.

Table 2
Data structure

Parameter name	Data type	Data kind
Gender	character	qualitative
Age	integer	quantitative
Social group	character	qualitative
Address (city, region)	character	qualitative
City/village	character	qualitative
Place of poisoning	character	qualitative
Date of poisoning	date	qualitative
Diagnosis	character	qualitative
MKB10 code	character	qualitative
Who has made the	character	qualitative
Health facility	character	qualitative
Number of victims	integer	quantitative
Lethal outcome	logic	qualitative
Purpose	character	qualitative
Place of obtaining poison	character	qualitative
Other	...	...

The data are diverse, and only a part of the data is quantitative (numerical); for the most part, the data are qualitative (symbolical) which, on one hand, complicates the statistical analysis, but on the other hand this creates prerequisites for data mining application.

A feature of biomedical data is primarily the heterogeneity of the data itself, they can be both quantitative and qualitative, in most cases the data is not subject to normal distribution. Medical data is generally semi-structured. Toxicological data are generally the same as medical data. This is information about the state of a person - his age, social group, objective and subjective signs of disease (poisoning), etc.

Our data can be divided into the following kinds:

- 1) Quantitative data - parameters; they can be characterized by discrete values: the age of the patient, the number of victims.
- 2) Qualitative data - characteristics; do not give in to exact assessment, though can be ranged (for example, are systematized on conditional points: one point, two points, etc.). The social group, the address, the place of poisoning, date of poisoning (data type - date), the diagnosis, the MKB10 code, who established the diagnosis, where there took place treatment, the place of poisoning, the poisoning purpose, the place of acquisition of poison, etc. Qualitative characteristics can be classified into only two categories - sex, individual and group poisoning, lethal outcome (logical data type).

Thus, the data are presented in different kinds, in different types and formats, not subject to normal distribution. It is necessary to analyze not only within each kind and type of data, but also to identify patterns between data belonging to different kinds and types.

3.1 Primary Statistical Data Analysis

The primary research of raw data is the first stage of the analysis and is carried out for the detection of the most general regularities and tendencies, character and properties of the analyzed data, and laws of the analyzed data distribution [25]. The results of the initial prospecting analysis are not used for making decisions, their purpose is to help in the development of the best strategy for the profound analysis, hypotheses making, specification of mathematical methods, and model feature application. The prospecting analysis helps to concisely describe the structure of data in a visual form, and then to research it in more detail by statistical analysis and data mining. The purpose of this stage is to visualize data and to collect the maximum quantity of hypotheses for possible interrelations and regularities in data.

To carry out the initial analysis of data, it is necessary to process data and to output them the quantitative indices. For this purpose, selections have been divided into groups on the grouping indicators. We will consider the dynamics of quantity for cases of acute poisonings with various poisons on the example of age groups in Figure 2.

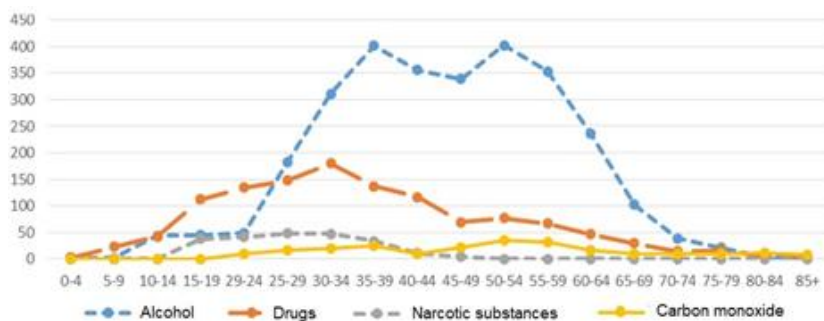


Figure 2

Dynamics for quantity of acute poisoning cases with various poisons in different age groups

The most frequent cause of acute poisonings in the Republic of Bashkortostan is alcohol poisonings. They are most prevalent in the group of 20 to 70 years of age. The maximum peak is reached for the groups from 35 to 60 years of age. Narcotic poisonings are registered in age groups from 15 to 40 years. Medicinal poisonings and also cases of poisoning with carbon monoxide are characteristic for all age

groups. The percentage ratio for the quantity of acute poisonings depending on age and gender is presented in Figure 3.

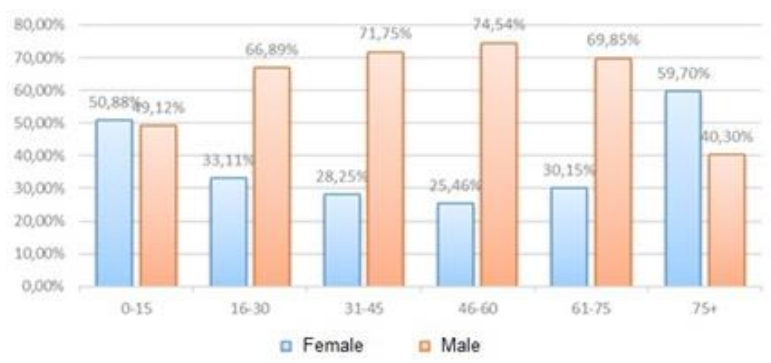


Figure 3

Structure of acute poisonings depending on age of men and women

From the chart, it is possible to see that the peak of the acute poisonings in men is reached at the age of 46-60 years, and in women, at the age of 75 and over, which, in turn, is not evident knowledge and demands further research.

For a clearer understanding of poisoning distribution in different age groups, dynamics of mortality (Figure 4) have been studied.

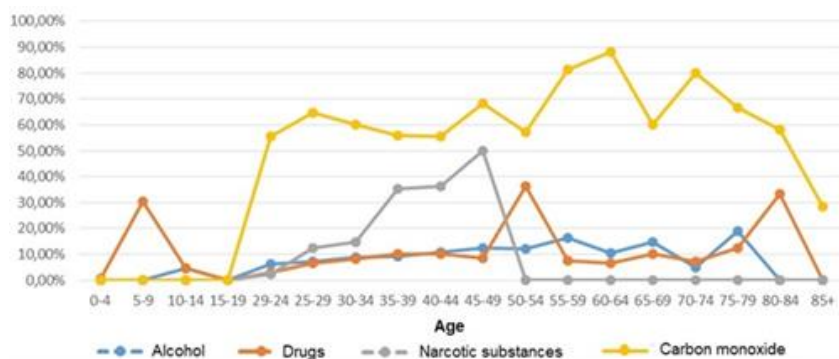


Figure 4

Mortality rate from acute poisonings with various poisons by age groups

The greatest indicator of mortality is present at poisonings with carbon monoxide. Alcohol poisonings mortality rate around 20 years of age is rather low but steadily growing. Medicinal poisonings also have an average level, with 3 peak age groups in which the death rate considerably exceeds the general. Risk groups are at the age of 5-9 years, 50-54 years, and also after 80 years. For the specification of

results, the structure of lethal cases of acute poisonings depending on the cause for men and women (Figure 5) is considered.

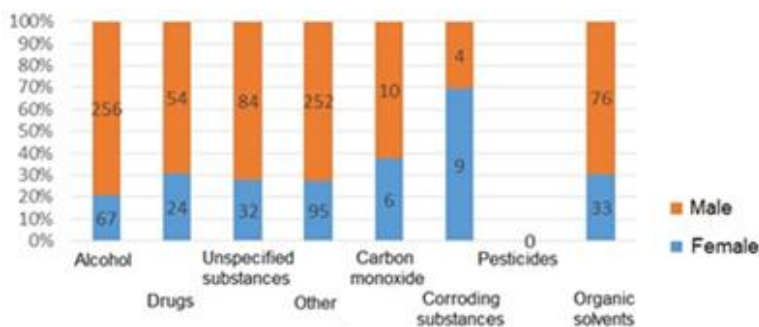


Figure 5

Structure of lethal cases of acute poisonings depending on the cause for men and women

Thus, already at the first stage of the prospecting analysis, it is possible to draw the following conclusions: the greatest number of poisonings occurs from 25 up to 60 years, this index is also high for people aged from 1 up to 3 years; the number of cases of acute poisonings for men is much higher than for women; the greatest death rate is for people aged from 25 up to 65 years; the death rate for men is higher than for women in all basic causes of poisonings, except the ones from corroding substances.

3.2 Data Statistical Analysis

The next stage of analysis is deeper data studying by means of the statistical analysis methods.

According to statistical principles used in the basis, the methods are subdivided in [12]:

- parametric - applied mainly to the analysis of normally distributed quantitative signs;
- nonparametric - applied to the analysis of quantitative signs irrespective of their distribution and to the analysis of qualitative signs.

Nonparametric methods are developed for those situations when the researcher knows nothing about any parameters of the analysed data [12]. Owing to features of the medical data that is listed above it is more effective to apply methods of the nonparametric analysis to their processing.

The examples of data processing results with nonparametric analysis methods for the main causes of poisoning executed in version 13.2 STATISTICA package are given below. In this analysis, dynamics of number of acute poisonings for the

main causes and quantity of cases of poisonings with a lethal outcome for 2015-2016 have been analysed.

From sample data, the following indicators for the number of cases of acute poisonings depending on the main causes for men and women (Table 3) have been received.

Table 3
Number of cases of acute poisonings by main causes for men and women

Poisoning cause	Number of poisonings	
	Female	Male
Alcohol	553	2346
Other	152	203
Drugs	888	811
Unspecified substances	254	727
Organic solvents	20	33
Pesticides	29	38
Corroding substances	49	72
Carbon monoxide	60	101

Then the Kruskal-Wallis test has been applied to compare the number of poisonings according to the main diagnoses (Figure 6).

Kruskal-Wallis ANOVA by Ranks; Var2 (Spreadsheet18) Independent (grouping) variable: Var1 Kruskal-Wallis test: H (7, N= 137) =73,03182 p =,0000					
Depend.: Var2	Code	Valid N	Sum of Ranks	Mean Rank	
Alcohol	101	18	1803,500	100,1944	
Other	102	18	1428,500	79,3611	
Drugs	103	18	1875,000	104,1667	
Unspecified substances	104	18	1703,500	94,6389	
Organic solvents	105	15	374,500	24,9667	
Pesticides	106	14	447,000	31,9286	
Corroding substances	107	18	827,500	45,9722	
Carbon monoxide	108	18	993,500	55,1944	

Figure 6

The results of Kruskal-Wallis test for comparing the number of poisonings according to the main diagnoses

The significance value of $p < 0.05$ signifies the statistical importance of results. The value for the criterion of H is exceeded the tabular one. Therefore, the statistical importance of distinctions is high. The greatest contribution to poisonings, from the largest to the smallest: alcohol, drugs, unspecified substances, carbon monoxide.

The indicators numbers of acute poisonings with lethal outcome depending on the main causes for men and women are presented in Table 4.

Table 4

Number of acute poisoning cases with a lethal outcome depending on their main causes for men and women

Poisoning cause	Number of poisonings	
	Female	Male
Alcohol	67	256
Other	24	54
Drugs	32	84
Unspecified substances	95	252
Organic solvents	6	10
Pesticides	9	4
Corroding substances	33	76

The results of Kruskal-Wallis test (Figure 7) have shown that the significance value lies also below 0:05 which signifies the statistical importance of results of this testing. Therefore, it is possible to say that separate indicators make significant contributions to the death rate from acute poisonings.

Depend.: Var3		Kruskal-Wallis ANOVA by Ranks; Var3 (Spreadsheet21) Independent (grouping) variable: Var2 Kruskal-Wallis test: H (6, N= 92) =33.45563 p =.0000			
		Code	Valid N	Sum of Ranks	Mean Rank
Other Quantity	101	15	646,500	43,10000	
Drugs Quantity	102	16	631,500	39,46875	
Carbon monoxide Quantity	103	16	782,000	48,87500	
Unspecified substances Quantity	104	16	1035,000	64,68750	
Alcohol Quantity	105	13	869,000	66,84615	
Organic solvents Quantity	106	10	166,000	16,60000	
Corroding substances Quantity	107	6	148,000	24,66667	

Figure 7

The results of Kruskal-Wallis test for comparison of lethal outcome according to the main diagnoses

Ranging of the leading causes of death, decreasing in-order: alcohol, unspecified substances, carbon monoxide, drugs, corroding substances, organic solvents.

The analysis of dynamics for numbers of acute poisonings for 1981-2016 (Fig. 8) had been carried out.

Kruskal-Wallis criterion has shown the statistical importance of differences in indications for selection of poisoning levels by main causes. The greatest contribution is made by alcohol poisonings, the level of poisonings with carbon monoxide and unspecified substances is also high. Indicators of narcotic and medicinal poisonings which had considerably high rates by consideration of dynamics of poisonings in the last two years throughout the long period have shown the lowest level.

Depend.: Quantity		Kruskal-Wallis ANOVA by Ranks: Quantity(Spreadsheet1) Independent (grouping) variable: Diagnosis Kruskal-Wallis test: H (7, N= 288) =187,7026 p =0,000			
		Code	Valid N	Sum of Ranks	Mean Rank
T51 Alcohol	101	36	8869,000	246,3611	
T58 Carbon monoxide	102	36	8062,000	223,9444	
T40 Narcotic substances	103	36	3721,500	103,3750	
T54 Corroding substances	104	36	4013,500	111,4861	
T52-T53 Organic solvents	105	36	2969,000	82,4722	
T36-T50 (without T40) Drugs	106	36	2762,500	76,7361	
T60 Pesticides	107	36	3341,500	92,8194	
T65 Other and unspecified substances	108	36	7877,000	218,8056	

Figure 8

The results of Kruskal-Wallis test for comparison of death rate according to the main diagnoses during the period for 1981-2016 in RB

At the second stage analysis of data, deep interrelations between data that can be used for decision-making are detected. At the same time, both stages are intended for the prospecting analysis, the best understanding of data, and hypotheses making which are preliminary steps for data mining.

3.3 Data Mining Stage

The technology of data mining allows to discover such patterns among large volumes of data which cannot be found by statistical ways of data processing but are objective and practically useful. Using these methods. the researcher can observe five main patterns in data [2], namely:

- association – several events are connected with each other;
- sequence – a chain of the events connected in time;
- classification – reference of an object to one of the classes with known characteristics;
- clustering – allocation of uniform groups of objects;
- temporary templates - dynamics of behavior of target indicators.

All listed patterns are applicable to medical data.

Results of the data processing with use of decision trees has been executed in the Deductor Studio Academic 5.3.0.88 package are given below.

Decision trees are a method of representation governed by a hierarchical, consecutive structures where the only unique knot giving the decision corresponds to each object. This method is already actively applied in medicine and biomedical studies and allows to make the diagnosis and to predict possible consequences of treatment.

Having set the MKB10 Code as an output parameter, and all other available parameters as entrance, we have received model which will allow us to estimate at the initial stage the importance of a contribution of separate parameters to a resultant indicator – the "MKB Code" describing the type and causes of poisoning. The greatest contribution to the definition of the poisoning diagnosis is made by such indicators as "the purpose of toxic agents intake ", age of the patient, and lethality; an insignificant contribution is made by "the number of victims", date of poisoning and the patient gender (Figure 9).

Target attribute: MKB10 Code				
Nº	Number	Attribute	Certainty, %	/
1	9	Purpose		51,605
2	2	Age		22,032
3	7	Letal_outcome		16,630
4	8	Number_of_victims		4,746
5	5	Date_of_poisoning		2,009
6	1	Gender		1,945
7	4	City/village		0,812
8	6	Who_has_made_the_diagnosis		0,221
9	3	Social_group		0,000

Figure 9

Analysis for the importance of separate factors influence on the diagnosis

From the received model, the following results were obtained:

- 1) At the age of more than 80 years, the most frequent causes of poisonings are an erratic drug intake or the corroding substances and also poisoning with carbon monoxide.
- 2) Up until the age of 36, men in public places become poisoned with narcotic substances, after this age, more often with drugs.
- 3) Group poisonings are most often caused by carbon monoxide.

Also, the contribution of separate parameters to the probability of a lethal outcome in case of acute poisoning was evaluated. The largest contribution is made by the Who Set the Diagnosis parameter; however, upon studying the initial selection it becomes clear that in the majority of lethal cases, the diagnosis is made by the forensic scientist. Other contributors to the probability of a lethal outcome are such indices as the cause of poisoning and the age of the patient (Figure 10).

Target attribute: Lethal_outcome				
Nº	Number	Attribute	Certainty, %	
1	9	Who_has_made_the_diagnosis		95,168
2	8	MKB10_Code		2,896
3	2	Age		1,936
4	7	Date_of_poisoning		0,000
5	11	Purpose		0,000
6	10	Number_of_victims		0,000
7	3	Social_group		0,000
8	1	Gender		0,000
9	4	Address_(city,_region)		0,000
10	6	Place_of_obtaining_poisoning		0,000
11	5	City/village		0,000

Figure 10

Analysis of the importance of influence of separate factors on the probability of a lethal outcome

In case of rendering medical assistance (both by the doctor and the paramedic), it is possible to avoid a lethal outcome in almost all cases except for those caused by the influence of narcotic substances in people over 30 years of age (Figure 11).

Antecedent	Consequent	Support	Reliability
IF		1997	1755
Who_has_made_the_diagnosis = Doctor	False	1608	1587
Who_has_made_the_diagnosis = Unknown	False	74	61
MKB10_Code	False	0	0
MKB10_Code = Drugs	False	39	39
MKB10_Code = Narcotic substances	False	25	13
Age < 30.5	False	10	10
Age >= 30.5	True	15	13
MKB10_Code = Organic solvents	False	3	3
MKB10_Code = Pesticides	False	0	0
MKB10_Code = Corroding substances	False	6	6
MKB10_Code = Carbon monoxide	False	1	1
Who_has_made_the_diagnosis = Forensic scientist	True	210	206
Who_has_made_the_diagnosis = Paramedic	False	105	103

Figure 11

Decision tree denoting medical assistance and lethal outcome

For specification of results, all non-significant parameters have been excluded from selection, and the value of a contribution of the most significant parameters to rgw probability of a lethal outcome (Figure 12) was estimated once more. The poisoning diagnosis parameter defines the probability of a lethal outcome by 80%. A small contribution is made by the age of the patient (18%). The influence of the patient belonging to a certain social group was revealed in this research in a very insignificant form.

To verify the results received above, the impact of the cause of poisoning, age, and social group on the probability of a lethal outcome has been considered (Fig. 13). The indicator of a lethal outcome is equal to "False" (false) in all groups, except for patients with the diagnosis of poisoning with narcotic substances who were older than 42.

Target attribute: Lethal_outcome				
Nº	Number	Attribute	Certainty, %	/
1	3	MKB10_Code		81,521
2	1	Age		18,026
3	2	Social_group		0,452

Figure 12

The analysis of the importance of the influence of the set factors on the probability of a lethal outcome

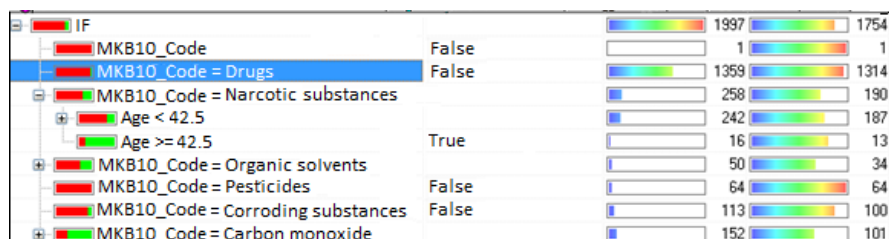


Figure 13

Decision tree with use of the set parameters

Thus, it is possible to draw a conclusion that the influence of narcotic substances becomes more considerable with age and with high probability leads to lethal cases after the age of 40. Let us check the assumption of the influence of age on the cause of poisoning. The analysis of the decision tree denoting social groups and main causes for toxic agent intake has shown that in women, cases of suicide poisonings are the most frequent cause for all main social groups. We will introduce amendments to the model and look at the influence of age on the cause of poisoning (Figure 14).

The analysis of the received results shows that the most frequent cause is drug intake for the purpose of suicide; the age, except for childhood, does not play an essential role, and the number of suicides for women is much higher than for men.

Results of the third analysis stage with the application of data mining allow us to draw the following main conclusions:

- 1) In the analysis of the main causes for poisonings in standard age groups, it is revealed that mortality from narcotic substances considerably increases after the age of 30, and in most cases, acute poisonings from drugs lead to a lethal outcome.
- 2) At this stage, alcohol poisonings have been excluded from the research. The most frequent cause of acute poisonings, as well as lethal outcome, was medicinal poisonings. For women, acute medicinal poisonings most often result from suicide intentions, but extremely seldom lead to a lethal outcome. For men, the level of suicide poisonings cases with medicinal substances is

much lower. However, mortality is much higher. Most often, medicinal poisonings are caused in men seeking to receive alcoholic intoxication.

- 3) At the retirement age, medicinal poisonings are most often caused by intake errors.
- 4) Timely assistance by the doctor or paramedic in most cases allows to avoid lethal outcome, except for cases of acute narcotic poisoning in people over 30.

The received regularities represent a new, unevident knowledge taken from the data and suitable for use in decision-making.

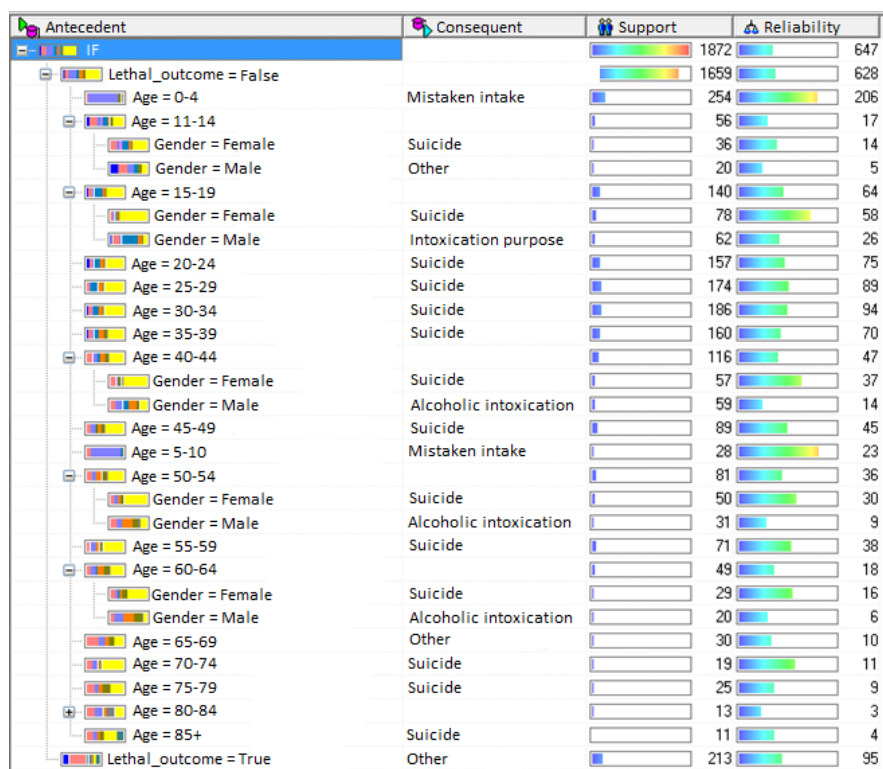


Figure 14

Decision tree denoting age groups and the main causes for the intake of toxic agents in specified groups

Conclusions

The analysis of researches in the field of toxicological data processing has shown that the majority of researches is conducted with the use of the simplest statistical methods. At the same time, traditional mathematical statistics uses the concept of

averaging on selection, i.e., operates with average characteristics which often are nonexistent whereas methods of the intellectual analysis can find unevident regularities in data. However, applying all of these methods, it becomes possible to study data in its entirety. Methods of mathematical statistics are useful mainly for check of hypotheses formulated in advance and for the prospecting analysis, representing an effective base for the subsequent application of methods of the intellectual analysis.

On the basis of the obtained information about existing methods of data analysis, the technique for medical data complex analysis in the field of toxicology has been developed. The technique includes three main stages: prospecting analysis of data with the use of visual methods of the analysis, statistical nonparametric, and also data mining.

The developed technique is applied to the analysis of data on acute poisonings in the Republic of Bashkortostan in 2015-2016. The results received during such analysis can help the leaders of medical institutions and chief specialists of governing bodies in the analysis of the indicators characterizing dynamics of tendencies in population health, distribution resource planning for health care of an area, and management of specialized health services.

The authors plan to conduct further researches in the area of data mining for the extraction of implicit regularities.

Acknowledgement

The reported study was funded by RFBR according to the research projects No. 19-07-00780, 19-07-00709, 18-07-00193 and state task No. FEUE-2020-0007.

References

- [1] E. H. Shortliffe, J. J. Cimino, *Biomedical Informatics*. Springer-Verlag, London, 2014
- [2] Herland et al., "A review of data mining using big data in health informatics", *Journal of Big Data*, 2014
- [3] Tsanas, M. A. Little, P. E. McSharry, "A methodology for the analysis of medical data" in *Handbook of Systems and Complexity in Health*, Springer, New York, 2013
- [4] Karthik Suresh, Sanjeev V. Thomas, Geetha Suresh, "Design, data analysis and sampling techniques for clinical research", *Ann Indian Acad Neurol*, 14(4), pp. 287-290, 2011
- [5] O. V. Marukhina, E. E. Mokina, E. V. Berestneva, "Using Data Mining for revealing hidden regularities in the task of analyzing medical data" in *Fundamental Research*, Vol. 4, pp. 107-113, 2015 (in Russian)

- [6] Thomas Luechtefeld et al., “Analysis of Public Oral Toxicity Data from REACH Registrations 2008-2014”, *Alternatives to Animal Experimentation: ALTEX*, Vol. 33 (2), pp. 111-122, 2016
- [7] P. Saldan, A. A. Ushakov, T. N. Karpova, “The analysis of the situation on chemical etiology acute poisonings in the Altai Region administrative center city Barnaul in 1997-2013”, *Fundamental and application-oriented aspects of risk analysis for the population health: Proc. of the All-Russian scientific and practical Internet-conference of young scientists*. Perm: Book format, pp. 121-128, 2014 (in Russian)
- [8] Joachim Gruettner, Thomas Walter, Siegfried Lang, Miriam Reichert, Stephan Haas, “Risk Assessment in Patients with Acute Alcohol Intoxication” in *Vivo*, Vol. 29, No. 1, pp. 123-127, 2015
- [9] Ross A., Willson V. L., “Descriptive Statistics” in: *Basic and Advanced Statistical Tests*, SensePublishers, Rotterdam, 2017
- [10] Igual L., Seguí S., “Descriptive Statistics” in: *Introduction to Data Science, Undergraduate Topics in Computer Science*, Springer, 2017
- [11] Le Roux et al., “Inductive Data Analysis” in: *Geometric Data Analysis*, Springer, Dordrecht, 2005
- [12] I. Kobzdar, *Applied mathematics and statistics*, Moscow: FIZMATLIT, 2012 (in Russian)
- [13] Kirch W. (eds), “Canonical Correlation Analysis” in: *Encyclopedia of Public Health*, Springer, Dordrecht, 2008
- [14] Igual L., Seguí S., “Regression Analysis” In: *Introduction to Data Science. Undergraduate Topics in Computer Science*, Springer, 2017
- [15] K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*, Springer Nature Singapore Pte Ltd., 2016
- [16] E. Filatov, *Methods of the determined factor analysis*, LAP Lambert Academic Publishing, 2012
- [17] H. K. Ramakrishna, *Medical Statistics*, Springer Science+Business Media Singapore, 2017
- [18] T. J. Cleophas, A. H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, Springer International Publishing Switzerland, 2015
- [19] Bonet et al. “Clustering of Metagenomic Data by Combining Different Distance Functions”, *Acta Polytechnica Hungarica*, Vol. 14, No. 3, pp. 223-236, 2017
- [20] G. Gosztolya et al. “Application of Fuzzy and Possibilistic c-Means Clustering Models in Blind Speaker Clustering”, *Acta Polytechnica Hungarica*, Vol. 12, No. 7, pp. 41-56, 2015

- [21] Jaeger D., Jung R. *Encyclopedia of Computational Neuroscience*, Springer, New York, NY, 2015
- [22] José Neves, Adriana Cunha et al., “Artificial Neural Networks in Diagnosis of Liver Diseases”, *Proc. of the 6th International Conference on Information Technology in Bio- and Medical Informatics ITBAM 2015*, Valencia, Spain, pp. 71-80, 2015
- [23] H-Y Li et al. “Medical Sample Classifier Design Using Fuzzy Cerebellar Model Neural Networks”, *Acta Polytechnica Hungarica*, Vol. 13, No. 6, pp. 7-24, 2016