ISSN 2664-9969

UDC 004.9:614.2
DOI: 10.15587/2706-5448.2020.210671

Mulesa O.,
Snytyuk V.,
Trombola M.,
Ivazkevych V.

DESIGN OF INFORMATION TECHNOLOGY CLASSIFICATION BASED ON MEDICAL DATA

Decision-making processes associated with assigning a person to a risk group for diseases are accompanied by the need to analyze large volumes of medical and social data. At the same time, a qualified doctor must operate both the patient's personal data and the relevant treatment protocols and instructions. Early prediction of the risks of disease occurrence allows medical workers to plan, develop a system of preventive measures, and the like. Therefore, the object of research is to support and provide information and analytical support for decision-making processes for early predicting the risks of diseases in individuals based on medical data. Such support is necessary to analyze the experience of a medical worker, which is recorded in the form of statistical data. One of the most problematic areas at the design and implementation stage of relevant information technology is the collection and analysis of statistical data on the problem under study.

The study was carried out in accordance with the systematic approach methodology. All stages of the design of information technology for forecasting based on medical data correspond to the stages of a systematic approach: systematization, formalization, goal orientation. The developed technology is based on the classification method based on Wald's sequential analysis.

The research resulted in:

– built a mathematical model of the problem of predicting the risks of disease occurrence as a classification problem;

– a functional diagram of an information and analytical system has been developed to solve the classification problem based on medical data. The analytical core of the information and analytical system is formed by algorithms for statistical data processing, as well as a classification method based on Wald's sequential analysis;

– experimental verification of the developed technology for the task of predicting the occurrence of dentoalveolar anomalies in children has been performed. Based on the available statistical data, a differential prognostic table was constructed. All calculations are performed. The examples demonstrate the effectiveness of the developed technology.

The developed information technology can be used by medical workers in the process of early forecasting of the risks of disease.

Keywords: *medical and social data, medical statistics, risk groups, medical services, information and analytical support.*

Received date: 20.04.2020
Accepted date: 21.05.2020
Published date: 31.08.2020

1. Introduction

Decision-making processes associated with assigning a person to a risk group for diseases are accompanied by the need to analyze large volumes of medical and social data. At the same time, a qualified doctor must operate both the patient's personal data and the relevant treatment protocols and instructions. However, in practice, an additional important source of data is the doctor's own experience, which is recorded in the form of medical statistics. Based on the experience of previous cases of the disease in different people, a qualified doctor can perform an early prognosis and assign a person to the risk group for the disease. Implementation of an early prognosis allows a medical professional to plan and implement a number of preventive measures, to provide a person with qualified assistance and recommendations for further action.

Considering the fact that the described decision-making processes require a doctor to promptly and accurately process data of various natures, the development and implementation of relevant information technologies for assigning persons to risk groups is urgent.

2. The object of research and its technological audit

The object of research is to support and provide information and analytical support for decision-making processes for early forecasting of the risks of diseases in individuals based on medical data. Such support is necessary to analyze the experience of a medical worker, which is recorded in the form of statistical data. Based on the analysis performed, it is possible to make a decision on attributing a person to the appropriate risk group in order to implement further medical measures on it.

At the stage of designing and implementing relevant information technology, it is important to collect and analyze statistical data on the investigated problem.

3. The aim and objectives of research

The aim of research is to analyze the stages of the information technology design process for assigning a person to a risk group for a disease.

To achieve this aim, it is necessary to complete the following tasks:

1. To carry out a systematization of problems and tasks that arise in the process of designing relevant information technology.

2. To construct a mathematical model of the problem of predicting the risks of disease occurrence as a classification problem.

3. To develop a functional diagram of an information and analytical system to solve the classification problem based on medical data.

4. To carry out experimental verification of the developed technology for the task of predicting the occurrence of dentoalveolar anomalies in children.

4. Research of existing solutions of the problem

A number of scientific studies are devoted to the problems of forecasting in medicine. In [1], a study is presented in which the forecast of the main indicators associated with the onset and treatment of renal failure is performed. Forecasting is carried out on the basis of time series by various forecasting models. The work [2] is devoted to the problem of using models and methods of Big Data for the analysis of medical data in the course of solving problems of forecasting and classification. The issue of using cloud applications for processing and analyzing medical data is separately considered here. In [3, 4], a detailed analysis of the problem of using time series forecasting methods for modeling and forecasting the needs for emergency medical care in the region is presented. Autoregressive models with different parameters were used for forecasting. The problem of forecasting future needs for medical services is devoted to [5]. Forecasting in the work is carried out on the basis of time series characterizing the main demographic indicators of the region, as well as using simulation models. A medical data management system using mechanisms for predicting medical events is given in [6]. The developed system works with both statistical and dynamic data. Forecasting is done for different cohorts. The system is based on machine learning models and methods. A comparative analysis of modern methods for predicting medical time series is given in [7, 8].

An analytical review of modern scientific publications devoted to forecasting based on medical data shows that the overwhelming majority of them contain models and forecasting methods based on time series. The problems of predicting the risks of future occurrence of such conditions requiring medical support are rarely solved for specific clinical cases. To solve this type of problems, pathometric algorithms based on the analysis of statistical data and the development of clear rules for making a forecast are successfully used [9, 10]. Such algorithms are usually based on Bayes' theorem.

Thus, it is advisable to develop information technologies for predicting the risks of diseases on the basis of probabilistic methods.

5. Methods of research

The study was carried out in accordance with the systematic approach methodology. All stages of designing information technology for forecasting based on medical data correspond to the stages of a systems approach.

The developed technology is based on the classification method based on Wald's sequential analysis.

6. Research results

The information technology of assigning a person to the risk group for diseases was designed in accordance with the stages of a systematic approach – systematization, formalization, goal orientation [9].

At the stage of systematization of problems and tasks for the implementation of which the information technology is intended, the problem of early prediction of the risk of a disease on the basis of medical data was analyzed. The input data for such a forecast is retrospective data on the incidence of the disease in the past in different individuals. An important problem in this case is the analysis of the representativeness of the available data sample.

The second problematic issue is the choice of features by which early forecasting can be performed. The case in which early prediction of the risk of disease can be made only on the basis of the results of medical examinations of the person is trivial and, as a rule, does not require additional research. In the general case, a medical professional needs to analyze the results of a person's anamnesis, which include its medical and social data.

Thus, information technology, the purpose of which is information and analytical support of decision-making processes for predicting the risk of a disease in a person, should include models and methods for analyzing statistical data.

Formally, the problem of predicting the risk of a disease in a person can be represented as a classification problem as follows [10]: let's have a set of objects:

$$O = \{O_1, O_2, ..., O_n\},$$

for each of which the values for each criterion from the set are known:

$$K = \{K_1, K_2, ..., K_m\}.$$

That is, a given set of vectors:

$$W = \{w_i = (w_{i1}, w_{i2}, ..., w_{im}), i = \overline{1,n}\},$$

where w_{ij} – the value of the j-th criterion for the i-th object. Each object belongs to one of the two specified classes A and B, with the first class corresponding to the presence of a risk of a disease, and the second – to the absence of such a risk. It is necessary to set a rule according to which for some object O', which is characterized by a vector $w' = (w'_1, w'_2, ..., w'_m)$ of the corresponding criteria from the set K, based on the data on the objects from the set O, it will be possible to make a decision about its assignment to one of the classes A or B.

ISSN 2664-9969

To solve this problem, a classification method based on Wald's sequential analysis was chosen [10]. To apply this method, at the initial stage, it is necessary to perform preliminary processing of the input data [10, 11]:

– to calculate the diagnostic coefficients of the set criteria K based on the concept of conditional probability;

– to establish the information content of each criterion based on the Kullback measure;

– to determine the reliability levels of the decisions made related to the indicators of errors of the first and second years: an error of the first kind is understood as an erroneous assignment of an object to a class B, respectively, an error of the second kind describes an erroneous assignment of an object to a class A.

Actually, the classification method based on Wald's sequential analysis is that for a given object, based on the training sample, sums are calculated that characterize its proximity to classes.

Procedures that implement the stages of data preprocessing and the classification method based on sequential Wald's analysis form the analytical core of information technology. As a result of the work of the procedures, for each object it is possible to obtain one of the following solutions:

– the object belongs to a class A with a degree of belonging μ_A;

– the object belongs to a class B with a degree of belonging μ_B;

– there is not enough information to make a decision under the given conditions.

Based on the mentioned models and methods, an information-analytical classification system based on medical data is developed. The functional diagram of the system is shown in Fig. 1.

For the experimental verification of the designed information technology (IT), the problem of predicting the risks of dentoalveolar anomalies in children and adolescents was solved.

At the stage, a database of 105 children 7–9 years old was created. In accordance with the two predicted conditions, on the basis of the database, two training samples were formed of 60 children who had dentoalveolar anomalies at the age of 13 and 45 children who did not have dentoalveolar anomalies.

The task is to choose one of two prognostic solutions for a specific child aged 7–9 years, based on the analysis of risk factors: first, the child has a high risk of dentoalveolar anomalies (condition A), the second – a low risk of anomalies (condition B). In this case, the thresholds for $A=6.4$, for $B=-6.4$.

The initial data are given in Table 1.

According to the functional diagram of IT, the initial data were processed by the fuzzy classification method based on Wald's sequential analysis, as a result of which the following diagnostic coefficients and levels of information content for the indicators were obtained (Table 2).

Table 1

Initial data for the classification problem

Indicator	Value	Examined group	
		Main ($n=60$)	Control ($n=45$)
1. Heredity	0–1	15	35
	2–6	45	10
2. Bad habits	Yes	44	12
	No	16	33
3. Injuries	Yes	20	11
	No	40	34
4. Pathological posture	Yes	32	23
	No	28	22
5. Early caries and early tooth extraction	Yes	40	12
	No	20	33
6. Improper artificial feeding, prolonged sucking of a pacifier, eating soft foods	Yes	34	10
	No	26	35
7. Incorrect setting of tooth buds	Yes	3	2
	No	57	43
8. ENT pathology	Yes	25	8
	No	35	37
9. Systemic and chromosomal diseases of the body	Yes	1	1
	No	59	44
10. Environmental and hygienic factors	Yes	25	10
	No	35	35
11. Pathology of the mother's pregnancy	Yes	30	10
	No	30	35
12. Long-term replacement of temporary teeth with permanent	>16	10	25
	0–16	50	20

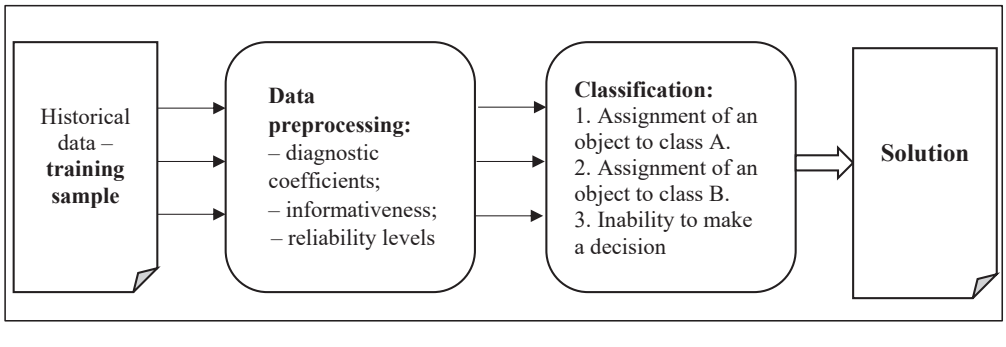

Fig. 1. Functional diagram of the information-analytical classification system based on medical data

Table 2

Differential predictive table

Indicator	Value	Diagnostic coefficient	Informativeness
1. Heredity	0–1	–4.93	2.69
	2–6	5.28	
2. Bad habits	Yes	4.39	2.05
	No	–4.39	
3. Injuries	Yes	1.35	0.08
	No	–0.54	
4. Pathological posture	Yes	0.18	0.004
	No	–0.20	
5. Early caries and early tooth extraction	Yes	3.98	1.48
	No	–3.42	
6. Improper artificial feeding, prolonged sucking of a pacifier, eating soft foods	Yes	4.07	1.18
	No	–2.54	
7. Incorrect setting of tooth buds	Yes	0.51	0.001
	No	–0.03	
8. ENT pathology	Yes	3.70	0.62
	No	–1.49	
9. Systemic and chromosomal diseases of the body	Yes	–1.25	0.004
	No	0.02	
10. Environmental and hygienic factors	Yes	2.73	0.39
	No	–1.25	
11. Pathology of the mother's pregnancy	Yes	3.52	0.76
	No	–1.92	
12. Long-term replacement of temporary teeth with permanent	>16	–5.23	1.55
	0–16	2.73	

Since, in accordance with the fuzzy classification method, only those indicators need to be taken into account, the information content of which exceeds 0.5, the classification will be carried out according to the following indicators:

– heredity;
– bad habits;
– early caries and early tooth extraction;
– improper artificial feeding;
– ENT pathology;
– environmental and hygienic factors;
– pathology of the course of pregnancy in the mother
– long-term replacement of temporary teeth.

To illustrate the operation of the algorithm, consider an example of data from a child aged 10 years, in which dentoalveolar anomalies were discovered later, at 13 years old.

Example 1. Child 10 years old. Let's apply the forecasting algorithm, considering the values of the child's indicators in descending order of their information content. The child's indicators and the results of the algorithm are shown in Table 3.

As it is possible to see from the Table 3, the algorithm stopped working at the fourth step when the sum of the diagnostic coefficients exceeded the threshold. Thus, it was decided to assign the patient to group A with a high risk of dentoalveolar anomalies.

Example 2. A child of 9 years old, in which no dentoalveolar anomalies were found in adolescence. The results of the predictive algorithm are shown in Table 4.

In this example (Table 4) it can be seen that the algorithm stopped at the stage when the sum of the diagnostic coefficients became less than the lower threshold. As a result, it was decided to assign the patient to group B of low risk of dental anomalies.

Thus, the designed information technology can be successfully used by dentists in the early stages to predict the possibility of risks of anomalies in the future.

Table 3

The results of the predictive algorithm (Example 1)

Algorithm step	Index	Value	Sum of diagnostic coefficients	Analysis
1	Heredity	4	$S=5.28$	$-6.5<5.28<6.5$
2	Bad habits	No	$S=5.28-4.39=0.89$	$-6.5<0.89<6.5$
3	Long-term replacement of temporary teeth with permanent	12	$S=0.89+2.73=3.62$	$-6.5<3.62<6.5$
4	Early caries and early tooth extraction	Yes	$S=3.62+3.98=7.6$	$7.6>6.5$

Table 4

The results of the predictive algorithm (Example 2)

Algorithm step	Index	Value	Sum of diagnostic coefficients	Analysis
1	Heredity	3	$S=5.28$	$-6.5<5.28<6.5$
2	Bad habits	No	$S=5.28-4.39=0.89$	$-6.5<0.89<6.5$
3	Long-term replacement of temporary teeth with permanent	17	$S=0.89-5.23=-4.34$	$-6.5<-4.34<6.5$
4	Early caries and early tooth extraction	No	$S=-4.34-3.42=-7.76$	$-7.76<-6.5$

ISSN 2664-9969

7. SWOT analysis of research results

Strengths. The introduction of the developed information technology will make it possible to predict the risks of the disease at the early stages. The implementation of an accurate and timely prediction will allow a medical professional to develop a set of preventive measures for each specific clinical case and thereby reduce the possible risks of dangerous conditions in a patient.

Weaknesses. A feature of the developed technology is that for the correctness of the medical prognosis, the medical worker must form representative training and control samples based on its own experience or available statistical data.

Opportunities. In the process of development of the designed information technology, it is expedient to introduce elements of the theory of fuzzy sets into the description of the main indicators by which forecasting is carried out. This approach will make it possible to make effective decisions when the indicators take values close to the limit.

Threats. The application of the developed information technology in various medical fields requires a preliminary analysis of the problem to form a set of risk factors for the occurrence of a disease, as well as intervals of values that these factors can take.

8. Conclusions

1. The systematization of problems and tasks that arose in the process of designing information technology for forecasting the risks of the disease. It is noted that the input data in the process of functioning of the technology are retrospective data on cases of the disease in the past in different individuals.

2. A mathematical model of the problem of predicting the risks of disease as a classification problem is built. Classification is carried out by assigning an object to one of two classes: a class with an existing risk of disease and a class without such a risk.

3. The functional scheme of the information-analytical system for the decision of a problem of classification on the basis of medical data is developed. The analytical core of the information-analytical system is formed by algorithms of statistical data processing, as well as the method of classification based on sequential analysis of Wald.

4. Experimental verification of the developed technology for the task of predicting the occurrence of dental anomalies in children is performed. During the verification, training and control samples were formed and risk factors (indicators) of these anomalies were selected. Two examples demonstrate the effectiveness of the developed information technology.

References

1. Sun, L., Zou, L.-X., Han, Y.-C., Huang, H.-M., Tan, Z.-M., Gao, M. et. al. (2016). Forecast of the incidence, prevalence and burden of end-stage renal disease in Nanjing, China to the Year 2025. *BMC Nephrology, 17 (1)*. doi: http://doi.org/10.1186/s12882-016-0269-8
2. Li, J.-S., Zhang, Y.-F., Tian, Y. (2016). Medical big data analysis in hospital information system. *Big data on real-world applications*, 65. doi: http://doi.org/10.5772/63754
3. Juang, W.-C., Huang, S.-J., Huang, F.-D., Cheng, P.-W., Wann, S.-R. (2017). Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. *BMJ Open, 7 (11)*, e018628. doi: http://doi.org/10.1136/bmjopen-2017-018628
4. Steins, K., Matinrad, N., Granberg, T. (2019). Forecasting the Demand for Emergency Medical Services. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. doi: http://doi.org/10.24251/hicss.2019.225
5. Lopes, M. A., Almeida, Á. S., Almada-Lobo, B. (2016). Forecasting the medical workforce: a stochastic agent-based simulation approach. *Health Care Management Science, 21 (1)*, 52–75. doi: http://doi.org/10.1007/s10729-016-9379-x
6. Park, Y., Ho, J., Vishwanath, S. (2016). *U.S. Patent Application No. 15/092,738*.
7. Amor, L. B., Lahyani, I., Jmaiel, M. (2016). Recursive and Rolling Windows for Medical Time Series Forecasting: A Comparative Study. *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, 106–113. doi: http://doi.org/10.1109/cse-euc-dcabes.2016.169
8. Kristianto, R. P., Utami, E. (2017). Optimization the parameter of forecasting algorithm by using the genetical algorithm toward the information systems of geography for predicting the patient of dengue fever in district of sragen, Indonesia. *2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 45–50. doi: http://doi.org/10.1109/icitisee.2017.8285548
9. Tymchenko, A. A. (2005). Systemnyi pidkhid do naukovoho doslidzhennia (orhanizatsiino-metodychni aspekty). *Visnyk ChDTU, 1*, 191–197.
10. Mulesa, O. Yu., Snytiuk, V. Ye., Herzanych, S. O. (2020). A fuzzy classification method based on the sequential wald analysis. *Automation of technological and business processes, 11 (4)*, 35–42. doi: http://doi.org/10.15673/atbp.v11i4.1597
11. Herzanych, S. O., Mulesa, O. Yu. (2018). Alhorytm prohnozuvannia nevynoshuvannia vahitnosti v umovakh pryrodnoho yodnoho defitsytu. *Zdorove zhenshchyni, 8 (134)*, 48–51.

Mulesa Oksana, *PhD, Associate Professor, Department of Cybernetics and Applied Mathematics, State Higher Educational Institution «Uzhgorod National University», Ukraine, ORCID: http://orcid.org/0000-0002-6117-5846, e-mail: mulesa.oksana@gmail.com*

Snytyuk Vitaliy, *Doctor of Technical Sciences, Professor, Dean of Faculty of Information Technology, Taras Shevchenko National University of Kyiv, Ukraine, ORCID: http://orcid.org/0000-0002-9954-8767, e-mail: snytyuk@gmail.com*

Trombola Mykhailo, *Postgraduate Student, Department of Cybernetics and Applied Mathematics, State Higher Educational Institution «Uzhgorod National University», Ukraine, ORCID: http://orcid.org/0000-0002-5044-799X, e-mail: mishatrombola@gmail.com*

Ivazkevych Viktoriia, *Lecturer, Department of Pediatric Dentistry, State Higher Educational Institution «Uzhgorod National University», Ukraine, ORCID: http://orcid.org/0000-0002-0080-467X*