Methods for the Intelligent Analysis of Biomedical Data

E.V. Geger[1], A.G. Podvesovskii[2], S.A. Kuzmin[2], V.P. Tolstenok[2]
emiliya_geger@mail.ru|apodv@tu-bryansk.ru|wolv3333@mail.ru|tolstenok21@yandex.ru
[1]Bryansk Clinicodiagnostic Center, Bryansk, Russia
[2]Bryansk State Technical University, Bryansk, Russia

The paper discusses methodology of cleaning and analysis of small semi-structured samples of biomedical data. This methodology is aimed at statistical evaluation of harmful production factor correlation with workers' laboratory test data. As a result of the analysis and interpretation of the data, a deviation from the norm is observed according to some indicators of a clinical blood test in individuals whose occupational activity is associated with harmful factors. Conclusions are drawn about the need for further research in the group of people whose work is related to harmful production factors. It is necessary to employ intelligent methods for analyzing possible health risks and their negative consequences in order to make management decisions. The presented assessment methodology can be used to create an occupational health and safety information system.

Keywords*: risk assessment, data analysis, harmful working conditions, statistical methods, data cleaning, model ensembles, Kohonen self-organizing maps.*

1. Introduction

In the modern world, the necessity often arises for auxiliary methods of preliminary disease detection at early stages. This problem is especially characteristic of people whose labor activity is associated with the constant impact of harmful working conditions [6, 10].

In turn, increased level of exposure to harmful substances and related industrial hazards significantly increase the likelihood of developing occupational diseases and the risk of injury [8].

Thus, occupational risk management is a complex of organizational and technical measures which should be based on reliable results of data analysis [9].

For the tasks of occupational risk assessment, it is necessary to use medical data analysis methods that correspond to these tasks. Choice of methods affects the construction of theoretical biomedical models and characteristics of experimental studies [7, 17].

However, many experts note that biomedical data are often unsuitable for processing using traditional software not only because of their volume but also because of the variety of data types and speed at which they must be analyzed [1, 3, 15]. Therefore, a system for the intelligent analysis of medical data is required, which could aggregate and analyze heterogeneous information coming from different sources: electronic medical records, data from monitoring sensors, ultrasound and X-ray apparatus and other devices [4].

If we consider real biomedical data, they have a number of specific features that make them unstructured: presence of various data corruptions, such as omissions, extreme values, manual entry errors, incorrect information, high dimensionality and heterogeneity, a large number of noisy and duplicate data. This leads to the unsuitability of most of the data or the entire sample for existing analysis algorithms [16].

Analysts believe that today, to solve many tasks of the healthcare system, it is necessary to go in the direction of structuring information and focusing on work with small samples [5].

Particular attention should be paid to the so-called small samples, the volume of which is about 100-200 records. This, in turn, is very difficult and makes the use of existing methods of data processing and analysis ineffective.

The main obstacle to the manual analysis of small samples is the inability of the analyst to notice hidden patterns in presented data, while special analytical algorithms detect existing patterns with much greater efficiency. The main task of the analyst, in this case, is to interpret the results of these algorithms and filter out false and trivial patterns.

New analytical technologies can not only increase the efficiency of medical institutions but also make it possible to solve such health problems as identifying diagnoses, medical errors, associative connection of diagnoses with results of laboratory tests and much more.

The objective of the present research has been to assess relationship between occupational morbidity and harmful production factors. So the experimental sample consisted of individuals whose occupational activities were associated with harmful and dangerous working conditions. Another objective was development of a new methodology focused on processing and analyzing small semi-structured samples of biomedical data.

2. Existing Solutions

Among many proposed methods and tools for working with small samples, it is advisable to apply ensemble data analysis methods in combination with the basic classifier – Kohonen self-organizing maps.

An ensemble of models is a combination of several learning algorithms that, working together, help to build a model more efficient and accurate than any of the models built using a separate algorithm. That is, to find a solution for one problem or to prove a hypothesis, not one but several models are used. Besides, the overall operating result matters and not that of a single separate model.

Formerly, researchers faced the problem of combining accuracy, simplicity and ease of interpretation in one method. A simple method could be used with a relatively easy interpretation but coming short of accuracy or vice versa, a complex but accurate method could be chosen that would be difficult to interpret. Ensembles of models have become a solution to this problem as a universal way of improving the accuracy of methods.

Ensemble learning refers to the training of a finite set of basic classifiers with the subsequent combination of their forecasting results into a single forecast of an aggregated classifier. Thus, an aggregated classifier will give a more accurate result.

The goal of combining models is to improve (enhance) the solution provided by a separate model. It is assumed that a single model will never be able to achieve the efficiency that the ensemble will provide.

Self-organizing maps are a special type of artificial neural network allowing for non-linear regression and projection of multidimensional data onto a two-dimensional plane with preservation of distances in their original data space [2]. This approach can be applied in various fields, including biomedical data processing.

3. Proposed Theoretical Solution

To solve the research problems of working environment risk assessment, we formed two groups:

Group I included individuals whose labor activity was associated with exposure to harmful production factors.

Group II consisted of individuals whose professional activity lacked harmful production factor.

The studies were carried out in the laboratory of Bryansk Clinical Diagnostic Center; the results were reflected in the medical information system "MAIS DC".

The list of occupational diseases was determined in accordance with the Order of the Ministry of Health and Social Development of the Russian Federation No. 417n dated April 27, 2012 “On the Approval of the List of Occupational Diseases”. Those working in harmful labor conditions are at risk for diseases associated with exposure to occupational physical factors [12].

The studies were carried out in accordance with the Order of the Ministry of Health of the Russian Federation dated April 12, 2011 No. 302n (as amended on February 06, 2018) “On the Approval of Lists of Harmful and (or) Hazardous Occupational Factors and Kinds of Work that Require Mandatory Preliminary and Periodical Medical Examinations (Surveys), and the Procedure for Conducting Mandatory Preliminary and Periodical Medical Examinations (Surveys)” [13].

To assess the possible risk of diseases caused by harmful occupational factors, the values of clinical blood test scores were used as a source of information. The experiment was carried out in compliance with the ethical principles of biomedical research and in accordance with the Federal Law of the Russian Federation No. 152 "On Personal Data" [14].

For morbidity analysis, the "International Statistical Classification of Diseases and Related Health Problems" of the tenth revision (ICD-10) was used [11].

To solve the problem, a methodology was proposed consisting of the following stages:

1. Data cleaning.
2. Model ensemble development.
3. Result interpretation.

To carry out the first step, a special model has been designed and developed, the task of which is to clean the source data and convert unstructured data into ordered ones.

The resulting model can be roughly divided into five parts.

Data import. It includes setting field names and labels, excluding empty fields, setting data types.

Data preprocessing. It is a submodel that generates two sets of data at the output – intact data, containing no errors, and corrupted data, which are input to the next submodel.

Data cleaning. This submodel receives damaged data as input and, after appropriate processing, provides output of cleaned data that does not contain initial errors. In particular, this submodel sets correct deviation values. In addition to the cleared data, the output of the submodel contains data sets with incorrect values in deviations and in scores.

Data merge. In this part, aggregation of initially complete and cleared data takes place. In addition to this, using a code written in the JavaScript programming language, patient’s unique identifier is generated and all diagnoses are divided into several records, so that three records containing one diagnosis are obtained from one record containing three diagnoses. This, in turn, allows for data structuring. The converted data can be used when applying transaction analysis algorithm. Due to the errors in the diagnosis name, it was decided to download an excel file containing codes and names of 12257 diagnoses of the international classification of diseases of the tenth revision [11].

In the resulting data set, the initial field “Diagnosis Name” has been replaced by the name of the diagnosis from this file.

Data export. It creates a file with the extension .txt, suitable for uploading to analytical platforms and further analysis.

features of this model, in addition to the fact that it does not allow meaningless data loss, are:

1. Correction of empty values.
2. Deletion of entries containing no information.
3. Recalculation of deviations by scores.
4. Recovery of missing scores by deviations.
5. Reassignment of diagnoses by ICD codes.
6. Assigning a unique number to each patient.
7. Increasing the number of records suitable for analysis without generating synthetic patients.
8. Download and comparison with ICD certified directory of codes and their descriptions.
9. Uploading data that has been damaged and subsequently corrected for additional possible analysis.
10. Converting source data into a form suitable for intelligent analysis methods.

At the second stage, development and construction of a model ensemble takes place. To simplify the development process, the following algorithm must be followed:

1. Select a base model.

An ensemble can consist of classifiers of one type (e.g., only of decision trees or only of neural networks) or of classifiers of various types (decision trees, neural networks, regression models, etc.).

2. Define the approach to using the learning set.

This can be resampling (several subsamples are extracted from the original learning set, each of which is used to train one of the ensemble models) or the use of one learning set to train all ensemble classifiers.

3. Select a method for combining results.

Three methods are usually used: voting (the class is selected that has been produced by a simple majority of ensemble models), weighted voting (the result is delivered taking into account weights set for ensemble models) and averaging (the output of the entire ensemble is defined as the simple average value of outputs of all models; in weighted averaging, the outputs of all models are multiplied by the corresponding weights).

As a result of applying this algorithm, a basic ensemble of models has been obtained (Fig. 1), which can be further modified and complexified.

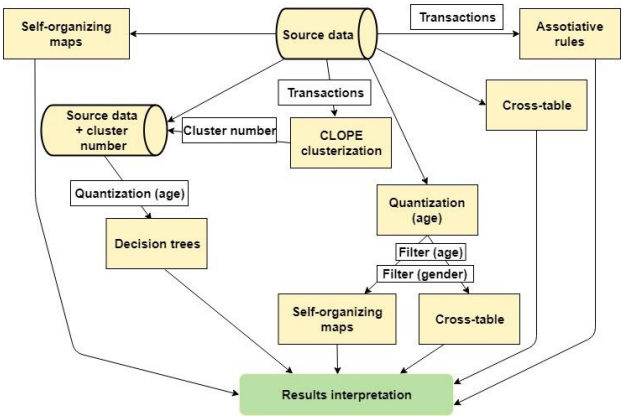

Fig. 1. Structure of the developed model ensemble

The data that has undergone preliminary cleaning at the first stage will be the input data.

At the third stage, the results obtained during the ensemble application are interpreted. In most cases, for their qualitative interpretation, it is necessary to contact specialists in the field of medicine.

For automatic analysis and visualization, ready-made analytical platform Deductor Studio was used, and, for preprocessing and data cleaning, Loginom software products developed by BaseGroupe Labs [18, 19] were used.

Thanks to the preprocessing of the initial sample, it is possible either to level out completely or to minimize the errors related to the human factor, which makes it possible to speak about sufficient correctness of the medical research. Unfortunately, small semi-structured samples cannot be called sufficiently representative. However, the identified patterns may be of interest to experts or for further verification.

4. Proposed Practical Solution

Laboratory test analysis has been carried out in workers whose occupation is related to harmful factors (Group I) and a control group (Group II).

Two small samples have been formed: the first sample includes 100 records; the second one includes 207 records for each of the scores taken into account for each group. A total of 9 scores of the general blood test were selected.

Data on the primary disease incidence in workers from Groups I and II have been analyzed. The results of diagnoses processing are presented.

The results of the study made it possible to identify the diagnoses that most often occurred in Group I and Group II:

E78.0 – Pure hypercholesterolemia;

G90 – Disorder of autonomic nervous system, unspecified;

H52.4 – Presbyopia;

I10 – Essential [primary] hypertension.

The cross-table as an interactive tool for data representation and analytical processing has allowed creating a pivot table which represents data on quantitative composition of diseases for each group.

The use of Kohonen self-organizing maps helped to identify differences and patterns between the studied groups.

Examples of Kohonen self-organizing maps used are presented in Fig. 2.

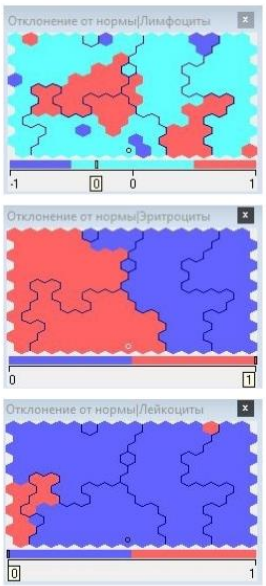

Fig.2. An example of data processing results using Kohonen self-organizing maps

It should be noted that in both groups at least 24% of employees were people aged 54 and over. In this regard, it has been suggested that these diseases may be associated primarily with age-related changes, for example, age-related decrease in the accommodative ability of the eye associated with the natural process of aging of the lens, a prolonged and persistent blood pressure increase, blood cholesterol increase.

Also, this age group is characterized by significant positive deviations from the norm in terms of hemoglobin, red blood cells, and erythrocyte sedimentation rate.

The increase in red blood cells is especially noticeable in the group of individuals whose work is associated with harmful occupational factors. Erythrocytosis could be caused by external factors.

Survey results may indicate the presence of pathological processes in the body. Additional studies are needed to diagnose the disease that has caused high content of red blood cells.

The methodology for analyzing small semi-structured samples of biomedical data considered in the article can be used with adequate efficiency as an integral part of the risk analysis of diseases related to harmful occupational factors. Application of the proposed method is demonstrated on specific actual data, collection and consolidation of which has been carried out using medical automated information system. This provides reliable evaluation of harmful occupational factor effects on working population's health indicators.

The research results will help to define and evaluate occupational risk factors that increase the likelihood of disease development and to draft proposals for the prevention of harmful effects of occupational factors on working citizens' health.

5. Conclusion

The article discusses a methodology tested on specific actual data, which allows pre-processing, cleaning and analysis of small biomedical data samples.

As a result of the analysis, there have been revealed no statistically significant difference in blood indices in individuals from groups I and II.

Also, the study of the diagnoses of the primary disease incidence in both groups has not revealed statistically significant differences between the groups.

However, the established significant deviations from the normal content of red blood cells and eosinophils in the blood of individuals from Group I may indicate the presence of certain pathological processes in the body, in particular, autoimmune processes, identification of which requires additional research.

In the future work perspective, it is advisable to conduct a research on these clinical blood test results using the traditional analysis of variance method regardless of normal intervals. This will allow for comparing the results obtained. Hereupon, it will be possible to draw final conclusions about the influence of specific occupational factors on the development of pathological processes.

The results obtained help to increase the efficiency of detecting patterns in biomedical databases and are of interest for the construction of intelligent systems designed to analyze and assess human health.

6. Acknowledgements

The reported study was funded by RFBR, project number 19-07-00844.

7. Литература

[1] Bruce McCormick (2014) Update in Anaesthesia. World Federation of Societies of Anaesthesiologists. 466 p.

[2] Kohonen T. The Self-Organizing Map // Proceeding of the IEEE. 1990. Vol. 78. P. 1464-1480.

[3] Manyika J., Chui M., Brown B. et al. (2011) Big Data: The Next Frontier for Innovation, Competition, and Productivity / McKinsey Global Institute.

[4] Baranov A.A., Namazova-Baranova L.S., Smirnova I.V. et al. Methods and Tools for Complex Intelligent Analysis of Medical Data. Trudy ISA RAN. Vol. 65. 2.2015. pp. 81-93.

[5] Barriers and Prospects for Digital Transformation: Big Data Management Issues in the Healthcare Industry [Online]. – Available: http://www.medlinks.ru/article.php?sid=83028 (Accessed: July 23, 2019).

[6] Geger E.V., Fedorenko S.I. Information Support of Decision-Making when Assessing the Risk for Occupational Morbidity Based on Analysis of Binary Samples // Proceedings of Southwest State University. Control, Computer Engineering, Information Science. Medical Instruments Engineering, no. 2 (27). 2018. pp. 101-107.

[7] Healthcare will Show the Largest Increase in Data Generation by 2025 [Online]. – Available: http://apcmed.ru/news/news-all/zdravookhranenie-pokazhet-naibolshiy-rost-v-generatsii-dannykh-k-2025-godu/ (Accessed: July 23, 2019).

[8] Izmerov N.F., Actualization of Occupational Morbidity Issues // Health Care of the Russian Federation (Zdravookhraneniye Rossiyskoy Federatsii.), no. 2. 2013. pp. 14-17.

[9] Ismailova L.N. Effective Management of Production Risks. // Economy and Business: Theory and Practice. 2016. No. 5. pp. 77-79.

[10] Kostenko N.A. Working Conditions and Occupational Morbidity as a Basis for Risk Management of Workers' Health: abstract of cand. med. sci. diss. M., 2015. 21 p.

[11] International Classification of Diseases of the tenth revision (ICD-10) [Online]. – Available: https://mkb-10.com (Accessed: July 20, 2019).

[12] The Order of the Ministry of Health and Social Development of the Russian Federation No. 417n dated 27/04/2012 “On the Approval of the List of Occupational Diseases”. [Online]. – Available: http://base.garant.ru/70177874/ (дата обращения 23.05.2019).

[13] The Order of the Ministry of Health of the Russian Federation dated 12/04/2011 No. 302n (as amended on 06/02/2018) “On the Approval of Lists of Harmful and (or) Hazardous Occupational Factors and Kinds of Work that Require Mandatory Preliminary and Periodical Medical Examinations (Surveys), and the Procedure for Conducting Mandatory Preliminary and Periodical Medical Examinations (Surveys)”. [Online]. – Available: http://base.garant.ru/12191202/ (Accessed: May 21, 2019).

[14] The Federal Law dated 27/07/2006 No. 152-FZ (as amended on 29/07/2017) "On Personal Data". [Online]. – Available: http://base.garant.ru/5635295/ (Accessed: May 21, 2019).

[15] Tsvetkova L.A., Cherchenko O.V. Big Data Technology in Medicine and Healthcare in Russia and the World // Information technologies for the Physician, 2016. No. 3. pp. 60-73.

[16] Tsygankova, I.A. Method of Intelligent Processing of Biomedical Data [Text] / I.A. Tsygankova // Software Products and Systems. – 2009. –no. 3. – pp. 120-123.

[17] Choporov O.N., Razinkin K.A. Optimization Model of Choice of the Initial Plan of Control Actions for Medical Information Systems / Control Systems and Information Technology. 2011. Vol. 46. No. 4.1. P. 185-187.

[18] BaseGroup. Data Analysis Technologies [Online] // Available: https://basegroup.ru/ (Accessed: May 20, 2019).

[19] Loginom [Online] // Available: https://loginom.ru/ (Accessed: May 20, 2019).