Advances in Intelligent Systems Research, volume 174
Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

Information and Analytical Support for Biomedical Research in the Field of the Cardiovascular Disease Risk Prediction

Alexander A. Zakharov*
Institute of Mathematics and Computer Science
Tyumen State University
Tyumen, Russia
a.a.zakharov@utmn.ru

Alexander P. Potapov
Telemedicine Service Group
Tyumen Clinical Hospital No. 1
Tyumen, Russia
dr.potapov@gmail.com

Pavel Y. Gayduk
Institute of Mathematics and Computer Science
Tyumen State University
Tyumen, Russia
tanoker09@yandex.ru

Alexander A. Kotelnikov
Institute of Mathematics and Computer Science
Tyumen State University
Tyumen, Russia
aakotelnikov72@gmail.com

Dmitry V. Panfilenko
Institute of Mathematics and Computer Science
Tyumen State University
Tyumen, Russia
dimjke-72@mail.ru

***Abstract*—The aim of the article is to study Data mining methods to predict the cardiovascular disease risks' level, based on data from medical information systems. We propose a new approach to the development of information and *analytical* support for biomedical research. Based on the proposed approach, we have developed the special methods, technologies and services to extract and anonymize valid problem-oriented information from unstructured electronic medical records. Created data store contains more than 70,000 records of electronic medical records and provides the researcher anonymized "smart" information in accordance with the possible scenario of its use. The developed services for visualizing the values of objective indicators allow us to determine the optimal data structure for the diagnosis of a specific group of diseases. This is shown by the example of risk testers of cardiovascular diseases. Based on the selected indicators, models for personified prediction of the cardiovascular disease risks' level were developed and tested)/**

Keywords—information and analytical support of scientific research, Data mining, information extraction, machine learning, electronic medical records, digital patient phenotype, medical text processing, personified prediction, cardiovascular diseases

I. INTRODUCTION

At present, modern medical information systems (MIS) are used as the basis for electronic document management system in health care. MIS implements an automatic generation of the necessary reporting documents and provides controlled access to patient data, staff and material, technical and financial resources of the medical organization. Moreover, for scientific research, biomedical data (BMD) in the form in which they are stored in the MIS, even if anonymized, are not of particular value, since there are no adequate tools for BMD processing [1].

At the same time, preliminary data analysis is of particular importance for the design and development of information storage, in which structuredness, correctness and consistency of stored BMDs are ensured. The purpose of the analysis is to determine non-obvious relationships and patterns that are necessary to extract objective information.

Therefore, the development of appropriate tools is one of the promising areas of the artificial intelligence methodology application for the information support of biomedical research. An important role in this context is played by the analysis of EMR data, which together determine the digital footprint of the patient. Based on the digital footprint, it is possible to identify patterns of changes in the state of health, especially the course of diseases, the effectiveness of the prescribed treatment, and as a result, to form a digital patient phenotype.

Note that EMC data allow us to assess the quality of medical care, the adequacy and timeliness of medical recommendations. At the same time, the primary task of extracting valid information from unstructured EMR data (primarily, patient examination protocols) does not have universal solutions, since clinical texts in Russian are not standardized. In this study, we solved this problem in the case of extracting data from EMR that is directly or indirectly associated with the risk of the occurrence and development of cardiovascular diseases (including taking into account concomitant diseases).

Thus, the goal of this study is to develop methods for extracting valid information from EHR data and build machine-learning models for a personified prediction of CVD risk.

II. STATE OF THE ART

In studies devoted to the fullest possible accounting of EMR data for predicting various, including cardiovascular,

diseases, two main approaches based on artificial intelligence methods are presented.

1) Building predictive models based on formal rules that reflect known criteria values and dependencies of indicators that directly determine the patient's condition in the studies of Rapsomaniki et al. [2], Miotto et al. [3] Choi et al. [4].

2) Using EMR data to build predictive models based on Data mining methods, Big Data technology and neural networks, as, in particular, in the studies of Ravı et al. [5], Shickel et al. [6].

Both approaches require the extraction of valid information directly from the texts of the examination protocols included in the EMR. Luo et al. [7], Kreimeyer et al. [8], Névéol et al. [9] note the objectivity of the existence of a large number of medical record processing systems in the natural (English) language (NLP, Natural Language Processing) and the possibility of taking this experience into account when developing similar systems for texts in other languages. Researchers have repeatedly noted the difficulty in interpreting clinical texts in Russian. Baranov et al. [10, 11], Dudchenko et al. [12] emphasize that the extraction of valid information related to specific named entities of EMR records is a fundamentally important problem that has not yet found a universal solution.

The solution to the issues of normalization and standardization of EMC data supplied by various MISs is presented in the works on the SHARP project (The Strategic Health IT Advanced Research Projects) of Rea et al. [13], Pathak et al. [14]. These studies emphasize the fundamental importance of reusing accumulated BMD not only to improve the quality of treatment, but also to conduct biomedical research and epidemiological monitoring at the national health system level. The authors define BMD standardization as the most important goal and propose a platform architecture that provides data exchange services between different organizations, storing and extracting structured data from EMR texts of various formats using the Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) tools, as well as converting to standard terminology. However, as shown by Peek et al. [15], Hripcsak et al. [16] there are problems associated with the features of the transition to new methods of data analysis (requirements for secure distributed storage and processing of data), as well as with possible deviations and incorrectness of the original BMD. All this naturally determines the accuracy of possible classification or prediction models and other conclusions.

Wang et al. [17], Luo et al. [18] show that another important aspect of extracting EMR data is internal structural complexity, which determines the relationship between survey data and prescriptions, taking into account chronology and other features. This determines the demand for machine and deep learning methods not only when building predictive models, but also already at the stage of extraction and preliminary analysis of EMR data (Luo et al. [18], Wang et al. [19]). Moreover, the creation of a ontologies' systems that characterize various groups of diseases plays the most important role for the correct data extraction and the use of fundamental subject (biomedical) relationships between entities in interpreting BMD, classifying them and constructing prediction models (Haendel et al. [20], Gribova et al. [21]). Risk prediction is becoming an increasingly important tool for managing patients with various diseases, as well as for making decisions about patients who do not yet have obvious diseases, but who are at risk for the disease in the short or long term. Obviously, a personalized risk assessment for an individual patient based on clinical, laboratory data can be used to make optimal decisions about treatment or the choice of a prevention strategy. For cardiovascular diseases, there are quite a few explicit models and risk prediction algorithms (Reynolds Risk Score, Heart, Score, etc. [22]) that rely on objective information (height, weight, gender, age, laboratory data, blood pressure, indicators ECG), as well as subjective data provided by the patient himself (smoking, physical activity, etc.).

It is very important to note that in the works devoted to CVD risk analysis, researchers emphasize the close relationship of the relevant criteria with a specific population, which served as a data source for prediction. This determines the need for training machine learning models on the data of the target region.

III. PROPOSAL METHOLOGY

We propose to solve the problem, based on the hypothesis that the effectiveness and reliability of biomedical research results, in particular, aimed at personified prediction of CVD risk, can be improved if we rely on "smart" information support - a problem-oriented digital phenotype patient (CVD-DPP).

CVD-DPP refers to the optimal data structure that combines indicators that are key to diagnosing a particular disease or group of diseases. In this case, we are talking about predicting the CVD risk from the point of view of criteria (rules) accepted in cardiology, with additional features. These characteristics can be determined by analyzing the data extracted from a large array of EMR. In the future, they can be supplemented with operational information obtained from portable diagnostic devices and include additional data characterizing both the patient and the doctor.

Here we are talking about complex dynamic features that show the patterns of caring for the health of the patient himself (herself) in dynamics (the frequency of visits to the doctor, the timely delivery of tests and additional examinations, etc.), and the quality of medical services indirectly reflected in the EMR records.

The proposed approach involves solving the following problems.

1) For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) Development of methods and technologies for extracting valid problem-oriented information from structured and unstructured EMR data (original texts in natural language in patient examination protocols).

3) Creation of an information repository of anonymized BMD, in which the structuredness, correctness, and consistency of BMD are ensured.

4) Presentation of BMD in accordance with a possible scenario for their use in biomedical research (preliminary analysis and visualization, clustering / classification of medical events, patients, medical personnel and individual medical records, etc.).

Development and testing of classification models to verify the possibility of recognizing the presence of coronary heart disease by the selected features.

IV. MATERIALS AND METHODS

The source data includes more than 70,000 records of the Tyumen residents' EMRs for 2014-2016 from the MIS SAP for Healthcare (https://www.sap.com/cis/industries/healthcare.html) and MIS 1C: Medicine. Clinic (https://solutions.1c.ru/catalog/clinic), used since 2017.

The source data includes more than 70,000 records of the Tyumen residents' EMRs for 2014-2016 from the MIS SAP for Healthcare (https://www.sap.com/cis/industries/healthcare.html) and MIS 1C: Medicine. Clinic (https://solutions.1c.ru/catalog/clinic), used since 2017.

When unloading, medical records (results of laboratory tests, descriptions of ultrasound examinations and ECG, examination protocols, etc.) are stored in separate XML files. As a result, the "raw" data is represented by files, which can be conditionally divided into 3 types according to the content of the fields: tabular, partially structured text, unstructured text.

Tabular files have a clear structure with specific field names and are formed, as a rule, according to the results of laboratory tests (blood, urine tests, etc.). Each of these files contains the name of the studied parameter, the measurement result, units of measurement, the limits of the norm of the value of the studied parameter, the date of the measurement, the location, name of the person responsible for this laboratory study. One file may contain the results of measurements of several parameters. Partially structured text files along with unambiguously defined and mandatory fields (ECG, ultrasound, etc.) contain textual conclusions in arbitrary form. Files of this type may vary in structure for even one type of survey. For example, some of them may contain only examination data, the structure of which is similar to the structure of laboratory tests (for example, ECG values), and others - the doctor's opinion, presented in a partially structured text.

Finally, the files belonging to the third type are formed on the basis of the patient examination protocols and contain unstructured text, which, along with the doctor's conclusion and recommendations, may contain additional features (blood pressure, weight, smoking, drinking strong alcohol, heredity, etc.). Moreover, the diagnosis can be recorded not only in accordance with the International Classification of Diseases ICD-10 (http://mkb-10.com), but also quite arbitrarily (for example, I25, CHD, IHD, coronary heart disease, chronic ischemic heart disease). In addition, the information contained in the examination protocols requires validation. This is due to the fact that the protocol is recorded on the basis of templates with some pre-entered data that the doctor may accidentally leave unchanged, although the actual patient performance may vary. The general thing was that files of all types initially contained personal data (surname, name, patronymic of the patient and the medical worker, date of birth of the patient, address), presented in full or in abbreviated / distorted form, which determined the non-triviality of depersonalization methods. In accordance with the identified objectives, we have developed and used the following methods.

Data stored in structured files was extracted using standard parsing methods for the XML format. The coordination of the laboratory tests' results with the data from the inspection protocols was carried out according to the criterion of proximity of the relevant dates with the choice of indicators values that deviated most from the norm. The solution to the problem of extracting valid information from unstructured texts related to specific named entities (personal data, diagnosis) was based on the integrated use of natural language text analysis methods. They included tokenization, morphological analysis, and preliminary analysis of medical texts (conclusions, examination protocols) to identify patterns of recording dates (birth, doctor's visit), blood pressure values and diagnoses, as well as obtaining a dictionary of conditional synonyms for diagnoses.

To create an information storage, we developed a multilayer architecture [23] based on the PostgreSQL DBMS, the choice of which was due to the need for denormalized storage of EMR data, primarily, patient examination protocols.

Smart" data, which is used for preliminary analysis and visualization, was recorded in XML files using Web services. Due to this approach, the researcher can evaluate and select significant parameters for risk metric methods for predicting CVD.

To train the models and select the classifier, we created a dataset that contains the following prehospital data: laboratory tests (Cholesterol, Glucose, Creatinine, AST, ALT), upper (systolic) and lower (diastolic) blood pressure, and an indicator of the presence or absence of diagnosis coronary heart disease.

We used the Python programming language and the Scikit-learn, Natural Language Tool Kit, pymorphy2 libraries for implementing methods and developing services.

V. RESULTS

We developed methods and technologies for extracting valid information on a test dataset, which included EMC records for 2632 patients. The total number of XML files is 72130, including 30845 laboratory test reports, 41285 inspection protocols.

A preliminary analysis showed that the data describing one type of laboratory test (especially inspection protocols) can be represented by various structures. This can be explained by the

fact that the source data was obtained as a result of unloading from the MIS. This system is in commercial operation, but is constantly being refined and updated. Accordingly, the dataset clearly indicates: date of formation (date of visit) in 5115 files; date of birth in 5096 files (1280 patients); gender in 3350 files (1097 patients); the main diagnosis with the ICD code in 174 files (of which 103 - I20, 71 - I25).

A. Data prepropcessing

In the process of anonymizing personal data, we solved the problem of determining the missing values of the patient's gender. For this, the surname, name and patronymic were used (full name). The decision included two stages: finding the name and gender determination by name. We solved the problem taking into account possible forms of recording (surname and initials, full name, surname and name) and isolating the patient's full name and the name of the medical staff. We used regular expressions and a morphological analyzer from the pymorphy2 library (selecting words in the nominative case and determining the gender). We also introduced a special metric to show the context of this data in the document. Next, gender was determined by the value of the name, in its absence by the value of the surname.

Further, before extracting data from unstructured texts, we analyzed the contents of 1017 files in order to identify possible options for doctors to write the date of birth, date of visit, blood pressure and diagnosis. For each of these parameters, we formed a contextual set of patterns - typical expressions, which was used in the future to compose regular expressions. The size of the context set for the date of birth was 31 expressions. The context set for the date of visit included 71 expressions. The context set of patterns for recording blood pressure in the inspection protocol included a significant number of elements (36) due to the variability of both the structures as a whole and the separator characters - from ordinary punctuation to alphabetic characters or their absence.

We included 17 expressions in the context set for the diagnosis (since we are talking about CVD, these are diagnoses related to I20-I25 according to ICD-10) indicating the presence or absence of certain phrases in the conclusion, based on available data and on the recommendation of experts.

For all of the above parameters, taking into account the contents of contextual sets, we developed a regular expression class that we used to extract data from unstructured medical texts. In addition, for each parameter, we prepared a set of special tests (filtering by date, date consistency, the presence of control words in the sentence). These tests were used to validate the data found using the corresponding regular expression class.

For EMR, in which there was no date of birth, it was calculated (with cross-checking) according to inspection protocols after extracting the age and date of visit. This value was then entered into the patient's key metadata.

As a result of the extraction and completion of the missing data, the proportion of EMR for which it was not possible to correctly establish the missing values was less than 5%, with the exception of the exact formulation of the diagnosis in accordance with ICD-10. In the latter case, the uncertainty was more than 40%. Therefore, previously, we entered into the database both mandatory binary values for I20-I25 (presence / absence of the diagnosis "coronary heart disease") and optional (if he was present in the EMC records) clarifying diagnoses (angina pectoris, acute myocardial infarction, etc.).

Thus, preliminary data processing allowed the formation of a three-layer information storage of depersonalized, correct and consistent BMD.

The first layer provides the anonymization of personal information, which allows organizing a secure exchange of distributed BMD (including EMC from medical information systems of various institutions) in its original form.

The second layer (pre-processed data) is responsible for organizing the coordinated distributed storage of key metadata for individual records (structure, record creation date, doctor ID) and EMR as a whole (patient ID, date of birth, gender, etc.).

The third layer ("smart" data) provides storage filling according to the results of extracting valid problem-oriented information (at this stage it is the presence or absence of a disease associated with CVD and the value of objective indicators of health status).

In addition, this same layer saves problem-oriented knowledge bases (Reynolds Risk Score, HeartScore criteria and rules, medical recommendations corresponding to them), as well as a dictionary of synonyms for diagnoses generated during the extraction of information from patient examination protocols.

The developed Web services for working with smart data are of interest, first of all, for cardiologists. The services provide a visualized representation of the distribution of objective indicators used to predict the level of CVD risk: gender, age, body mass index, systolic and diastolic pressure, laboratory test results, including taking into account the diagnoses (main and associated diseases).

B. Prediction models

The dataset for constructing predictive models contained 401 EMR records in which all the necessary values were present: age, gender, laboratory tests (Cholesterol, Glucose, Creatinine, AST, ALT), upper and lower blood pressure, and the presence or absence of a diagnosis of coronary heart disease (CHD). In these records, 170 patients did not have a body mass index (BMI). To restore this value, we used the SciKit-Learn IterativeImputer and SimpleImputer library classes with the strategy = 'mean'.

SimpleImputer with the parameter strategy = 'mean' is the class that fills the missing values with the average value for the parameter.

IterativeImputer is a more complex approach that models the function for missing parameter values as a function of the available values of other parameters, and uses this estimate to substitute the value.

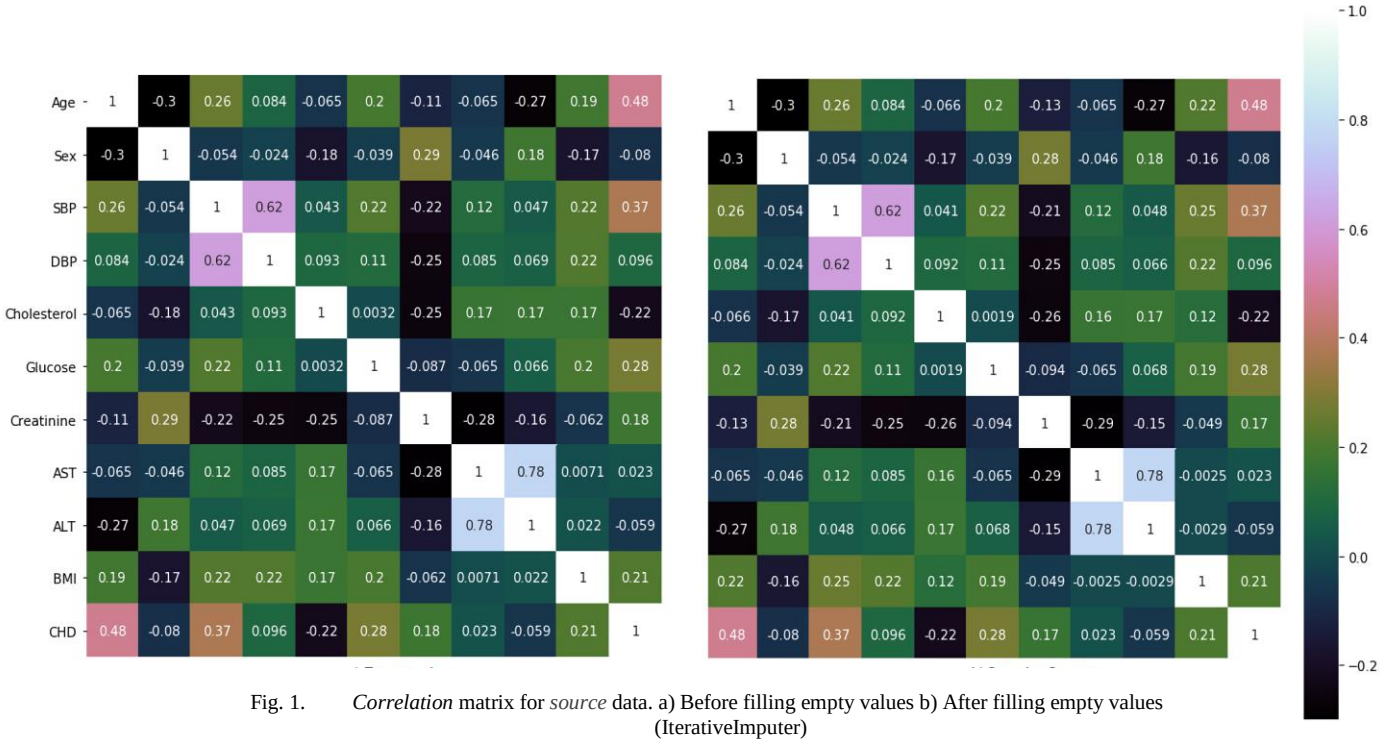

Fig. 1. *Correlation* matrix for *source* data. a) Before filling empty values b) After filling empty values (IterativeImputer)

The correlation matrix (Fig. 1) shows that the following parameters are most correlated with each other:

AST and ALT. Correlation coefficient 0.78
Upper and lower pressure. Correlation coefficient 0.62
Age and CHD. Correlation coefficient 0.48
Upper pressure and CHD. Correlation coefficient 0.37

To build the models, we used a sample with filled BMI values according to the IterativeImputer method. The correlation matrix shown in Fig. 1b demonstrates that the relationships between the parameters did not change significantly with this method of filling BMI values. Further, the data were divided into test and training samples. The training sample contained 240 records, of which 139 with CHD = 1. The test sample contained 161 records, 90 of which with CHD = 1. We selected three models for the study: Logistic regression, Random forest, Support Vector Machine.

TABLE I. MODEL QUALITY MEASURES

	Accuracy	Recall/Sen	Specificity	Precision
Logistic Regression	0.73	0.78	0.68	0.75
Random F	0.83	0.90	0.75	0.82
SVM	0.77	0.97	0.51	0.72

The results presented in Table 1 show that the Random forest and Support Vector Machine models are most suitable for predicting the risk of CVD. The SVM model is more accurate for identifying subjects with a positive diagnosis of CHD. But, as can be seen in Figure 2, this is due to over diagnosis - more cases of false positive choices. The model should be used if necessary to maximize the identification of patients with an appropriate diagnosis. The Random Forest model has such errors, but they are less common. It is advisable to use the model, for example, in the case when treatment in the absence of a disease can produce an undesirable effect.

These properties of the models are confirmed by the ROC graphs - the corresponding error curve (Fig. 3-5). They allow choosing a model with the best predictive power for a particular situation.

When choosing models, we investigated the possibility of using a one-class classification or "search for outliers." The training sample contained outliers, which are defined as observations that lie far from the rest. Outlier detection algorithms try to determine the regions where the bulk of the training data is concentrated, ignoring abnormal observations. We used the OneClassSvm class of the Sklearn library. The accuracy was 0.41. The low accuracy can be explained by the fact that the elements of the sample are located quite tightly. This makes it difficult to detect “anomalies” (CHD = 1).

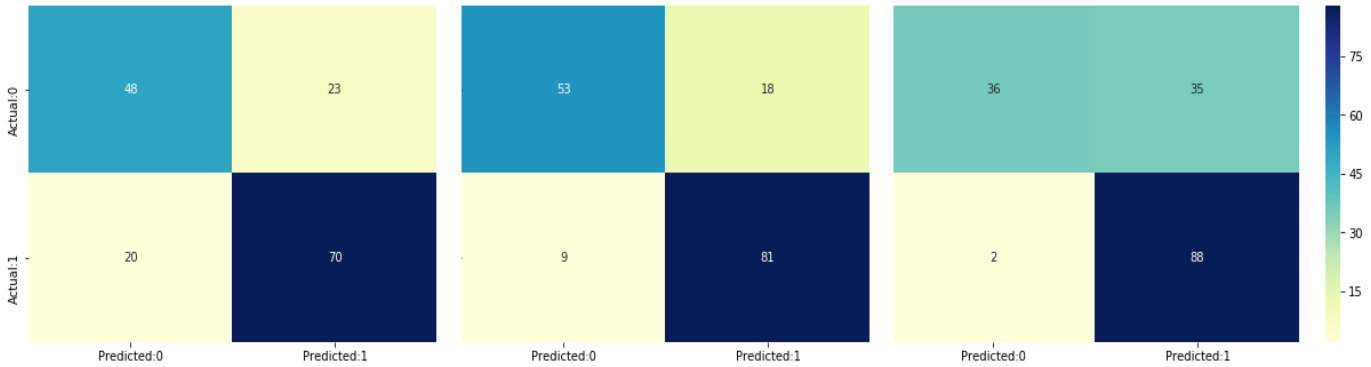

Fig. 2. Confusion *matrices*

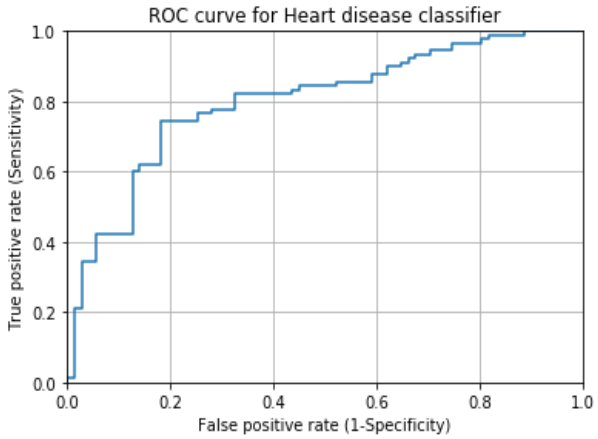

Fig. 3. Logistic regression. AUC = 0.80

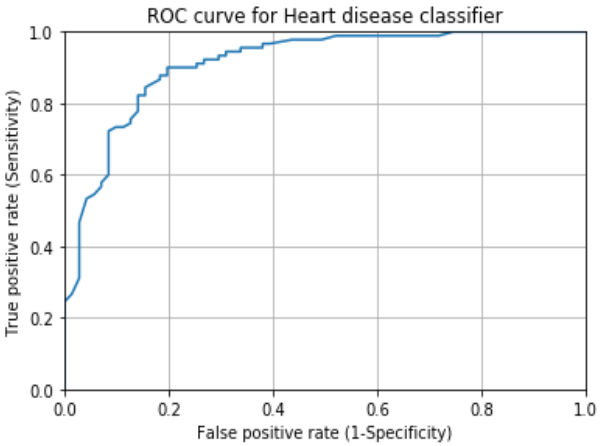

Fig. 4. Random forest. AUC = 0.91

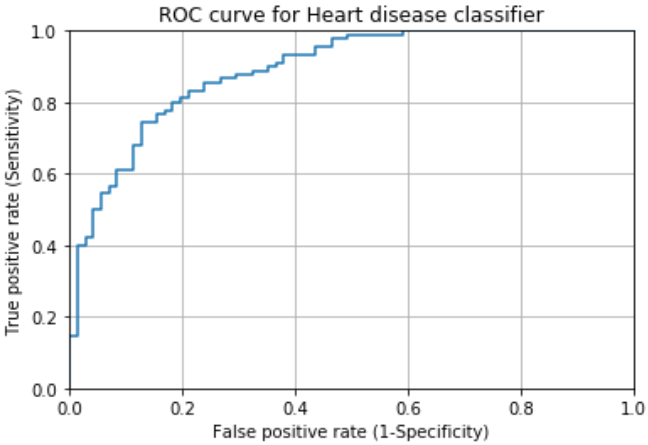

Fig. 5. *SVM. AUC = 0.88*

VI. Conclusion

The presented results on the predicting of CVD risks have shown their effectiveness in the pilot project, which allows using them as a basis for the development of external additional services of the regional IIA. By implementing data mining of electronic medical records, these services will be able to provide:

- assessing the risks of developing chronic life-threatening diseases and reducing the number of missing critical and subcritical deviations, for example, during mass screenings;
- automated monitoring of compliance with optimal medical management tactics for patients with risks of developing chronic life-threatening diseases;
- automated monitoring of the personalized follow-up plans implementation in accordance with the previously selected medical tactics for treating patients with a high risk of developing life-threatening chronic diseases (determination of reference control points, boundaries of parameter deviations and control of critical deviations).

REFERENCES

[1] Tavrovsky V. M., Gusev A. V. What should healthcare informatization lead to: an attempt to project the future // Doctor and Information Technologies. 2011. No. 5. P. 60-76. (In Russian)

[2] Rapsomaniki, E., Shah, A., Perel, P., Denaxas, S., George, J., Nicholas, O., ... & Smeeth, L. (2013). Prognostic models for stable coronary artery disease based on electronic health record cohort of 102 023 patients. European heart journal, 35(13), 844-852.

[3] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6, 26094. URL: https://www.nature.com/articles/srep26094

[4] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016, December). Doctor ai: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference (pp. 301-318).

[5] Ravı, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep learning for health informatics. IEEE journal of biomedical and health informatics, 21(1), 4-21.

[6] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics, 22(5), 1589-1604.

[7] Luo, L., Li, L., Hu, J., Wang, X., Hou, B., Zhang, T., & Zhao, L. P. (2016). A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. BMC medical informatics and decision-making, 16(1), 114. URL: https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-016-0357-5.

[8] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F.,& Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of biomedical informatics, 73, 14-29.

[9] Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: opportunities and challenges. Journal of biomedical semantics, 9(1), 12. URL: https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-018-0179-8.

[10] Baranov A. A., Namazova-Baranova, L.S., Smirnov, I.V., Devyatkin, D.A., Shelmanov, A.O., Vishneva, E.A. et al Methods and means of integrated intellectual analysis of medical data .] // Proceedings of the Institute for System Analysis of the Russian Academy of Sciences. - 2015. - T. 65. - No. 2. - S. 81-93. (In Russian)

[11] Baranov A.A., Namazova-Baranova, L.S., Smirnov, I.V., Devyatkin, D.A., Shelmanov, A.O., Vishneva, E.A. et al Technologies for the integrated intellectual analysis of clinical data // Bulletin of the Russian Academy of Medical Sciences. - 2016. - T. 71. - No. 2. - S. 160-171. (In Russian)

[12] Dudchenko P. V., Dudchenko A. V., Kopanitsa G. D. Methods of extracting data from unstructured medical records // Intelligent analysis of signals, data and knowledge: methods and tools. Collection of articles of the II All-Russian scientific-practical conference with international participation. (In Russian)

[13] Rea, S., Pathak, J., Savova, G., Oniki, T. A., Westberg, L., Beebe, C. E. Chute, C. G. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the

[14] Pathak, J., Bailey, K. R., Beebe, C. E., Bethard, S., Carrell, D. S., Chen, P. J., ... & Huff, S. M. (2013). Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American Medical Informatics Association, 20(e2): e341-e348.

[15] Peek N., Holmes J. H., Sun J. (2014) Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics //Yearbook of medical informatics. 9(1), 42-47.

[16] Hripcsak, G., & Albers, D. J. (2017). High-fidelity phenotyping: richness and freedom from bias. Journal of the American Medical Informatics Association, 25(3), 289-294.

[17] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N.,. & Liu, H. (2018). Clinical information extraction applications: a literature review. Journal of biomedical informatics, 77, 34-49

[18] Luo, Y. (2017). Recurrent neural networks for classifying relations in clinical notes. Journal of biomedical informatics, 72, 85-95.

[19] Wang, L., Wang, Y., Shen, F., Rastegar-Mojarad, M., & Liu, H. (2018, July). Predicting Practice Setting Using Topic Modeling. In 6th IEEE International Conference on Healthcare Informatics Workshops, ICHI-W 2018 (pp. 62-63). Institute of Electrical and Electronics Engineers Inc.

[20] Haendel, M. A., Chute, C. G., & Robinson, P. N. (2018). Classification, ontology, and precision medicine. New England Journal of Medicine, 379(15), 1452-1462.

[21] Gribova V.V. et al. Ontology of medical diagnostics for intelligent decision support systems [Ontologiya meditsinskoy diagnostiki dlya intellektual'nykh sistem podderzhki prinyatiya resheniy]// Design Ontology. - 2018. - T. 8. - No. 1 (27)..

[22] Petretta M., Cuocolo A. Prediction models for risk classification in cardiovascular disease //European journal of nuclear medicine and molecular imaging. – 2012. – T. 39. – №. 12. – C. 1959-1969

[23] Zakharov, A., Potapov, A., Zakharova, I., Kotelnikov, A., & Panfilenko, D. (2019, May). Infrastructure of the Electronic Health Record Data Management for Digital Patient Phenotype Creating. In 7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019). Atlantis Press. P.285-290.