Exploratory Analysis of Biomedical Data in Order to Construct Intelligent Analytical Models for Assessing the Risk of Cancer

Dmitry Lagerev[1], Anton Korsakov[1] and Alena Zakharova[2]

[1] *Bryansk State Technical University, 50 let Oktyabrya Ave., 7, 241035 Bryansk, Russia*
[2] *Institute of Control Science of Russian Academy of Sciences, 65, Profsoyuznaya st., Moscow, 117997, Russia*

Abstract

This article substantiates the need to use data from an integrated electronic medical record of a patient to assess the risk of cancer. An exploratory analysis of the data of the integrated electronic medical record of patients in the Bryansk region who received a diagnosis of "malignant neoplasm" is being carried out. The influence of the patient's age on the risk of oncological diseases is evaluated by the example of the nosologies C50, C61. Provides an overview of the capabilities of the Auto ML Libraries and their limitations. The article describes the result of constructing models for assessing the risk of oncological diseases based on the ML.NET and Auto-WEKA libraries. It is concluded that it is impossible to constructing models for assessing the risk of oncological diseases based on the data of an integrated electronic medical record using Auto ML libraries without preliminary preparation and preprocessing of data. And since it is required to constructing separate models for each nosology and regular retraining of these models, it is advisable to develop an add-on over the Auto ML libraries that will extract and convert the data of the integrated electronic medical record into a form suitable for analysis. In addition, to improve the quality of the model, it is advisable to use patient history data, data obtained after vectorization of laboratory tests, aggregated data on visits to specialized specialists and related diagnoses, data from online patient questionnaires filled out during the course of medical examination, as well as data on environmental pollution.

Keywords

Data mining, constructing data mining models, exploratory analysis, auto ML, biomedical data, risk assessment, socially significant diseases, malignant neoplasms.

1. Introduction

Currently, in various fields of activity, data mining models (DM) are more and more actively used. There is growing interest in the use of data mining models in medicine and in healthcare management problems [1, 2]. A prerequisite for this was the process of informatization of healthcare in the regions, which was actively started in the Russian Federation in 2012 and, in general, completed by 2018 [3]. Nowadays, in all constituent entities of the Russian Federation, medical information systems are functioning, which ensure the automation of the functioning of medical organizations and the primary input of data. There are also regional information systems that ensure the integration of regional medical information systems into a single whole and the consolidation of data from the entire region in a single data warehouse. At the federal level, regional information systems are united by the Unified State Information System in the field of health care, which currently supports 13 services. One of the major services is the Federal Integrated Electronic Medical Record, which is a subsystem of the Unified System designed to collect, systematize and process structured impersonal information about persons who receive medical care.

GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27-30, 2021, Nizhny Novgorod, Russia
EMAIL: lagerevdg@mail.ru (D.G. Lagerev); korsakov_anton@mail.ru (A.V. Korsakov); zaawmail@gmail.com (A.A. Zakharova)
ORCID: 0000-0002-2702-6492 (D.G. Lagerev); 0000-0002-4609-0246 (A.V. Korsakov); 0000-0003-4221-7710 (A.A. Zakharova)

To date, the regional information systems "Electronic medicine" of the Bryansk region have already accumulated a volume of biomedical data sufficient to start the process of constructing DM models, which, in turn, with an appropriate level of accuracy, can be used in the development of management decisions in the field of management health care at the regional level and the level of individual medical organizations. However, so far the DM models in the Russian Federation are not used in the process of health care management at the regional level, which is explained by the following reasons:

- The complexity of constructing DM models.
- The high cost of data scientists and the small number of such specialists in the regions.
- Failure to provide the level of accuracy required for quality control of all models.
- The need to construct an extremely large number of DM models for comprehensive automation of healthcare management, which is due to the number of different nosologies and the impossibility of constructing a single model for all nosologies. The International Classification of Diseases and Related Health Problems of the Tenth Revision (ICD-10) contains a description of about 14,400 different nosologies. More than a thousand of them are related to malignant and benign neoplasms.

2. Relevance of using DDM methods for assessing the degree of risk of cancer in patients

According to the latest estimates of the International Agency for Research on Cancer (IARC) at the World Health Organization (WHO) GLOBOCAN 2020 [4], the incidence of malignant neoplasms (MN) in the world has increased to 19.3 million new cases and 10.0 million deaths from them in 2020 ... One in five people in the world will develop cancer in their lifetime, and one in eight men and one in 11 women will die from the disease. The 10 most common types of cancer account for more than 60% of newly diagnosed cases and more than 70% of deaths. Cancer of the breast is the most common cancer worldwide (11.7% of the total number of new cases), followed by lung cancer (11.4%), colorectal cancer (10.0%), prostate cancer (7.3 %) and stomach cancer (5.6%) [4]. Lung cancer is the leading cause of cancer death (18.0% of total cancer deaths), followed by colorectal cancer (9.4%), liver cancer (8.3%), stomach cancer (7.7%) and breast cancer in women (6.9%).

According to forecasts [4], about 28.4 million new cancer cases will be registered worldwide in 2040, which is 47% more than in 2020. In addition, countries with economies in transition and countries classified as countries with a low or medium human development index will have the largest relative increases in cancer incidence by 2040 (95% and 64% compared to 2020, respectively).

Some scientists associate an increase in the incidence of cancer, first of all, with an improvement in diagnostics, as, for example, it happened in the United States in the 1990s. in the case of the introduction of the prostate-specific antigen (PSA) test in prostate cancer [5]. Some note that not all countries are covered by high-quality cancer registries, they are only about 15% [6], which increases the relevance of using DM models to identify risk groups. According to the research institute named after P.A. Herzen [7], the prevalence of cancer in Russia in 2019 was 2675 per 100,000 population, which is 41% higher than the level of 2009 (1897 per 100,000 population). The growth of this indicator is due to both an increase in the incidence and detection rate and an increase in the survival rate of cancer patients.

One of the main indicators that determine the prognosis of an oncological disease is the degree of prevalence of the tumor process at the time of detection. So, in Russia in 2019, 57.5% of malignant neoplasms were diagnosed in stages I and II of the disease (32.3% in stages I and 25.2% in stages II) and 42.5% in stages III and IV (17.6 % - in stage III and 24.9% - in stage IV) [7], while in 2009 more than 50% of newly diagnosed cancers were diagnosed in stages III-IV [8].

In connection with the above, the effectiveness of the functioning of the health care system in terms of detecting malignant neoplasms at an early stage (I-II) is extremely important in order to form risk groups in the process of population screening and preserve the health and quality of life of the population. An important factor hindering the timely detection of oncological diseases is the lack of clear criteria that allow doctors to identify a risk group during the period of clinical examination and conduct a set of additional examinations for patients belonging to this group. It seems promising to use a complex of data mining models for assessing the degree of risk in each patient during the period of clinical examination and subsequent informing the doctor about the degree of risk of patients

undergoing medical examination, which will, on the one hand, increase the detectability, and, on the other hand, will not overload the system of functional and laboratory diagnostics.

Among the circumstances of the risk of malignant neoplasms, there are many exogenous and endogenous factors, which are practically impossible for a doctor to take into account. According to the literature [9, 10], IARC and WHO [11, 12], among the main risk factors for cancer, the use of tobacco, alcohol, unhealthy diet, lack of physical activity, overweight, hereditary predisposition, chemical (polycyclic aromatic hydrocarbons, dioxins, pesticides, aflatoxins, arsenic, formaldehyde, nickel, asbestos, cadmium, and many others), physical (ionizing and ultraviolet radiation) and biological (infections caused by viruses, bacteria or parasites) environmental carcinogens. It should also be noted that the upward trend in the incidence of cancer in the world may reflect some general trends in an increase in the genetic load in human populations due to an increase in chemical and radiation contamination of the biosphere by “global” and “eternal” pollutants [13]. Moreover, for different types of oncology, the significance of the factors will be different, which makes it expedient to construct a separate DM model for each type of neoplasm.

3. Exploratory analysis results

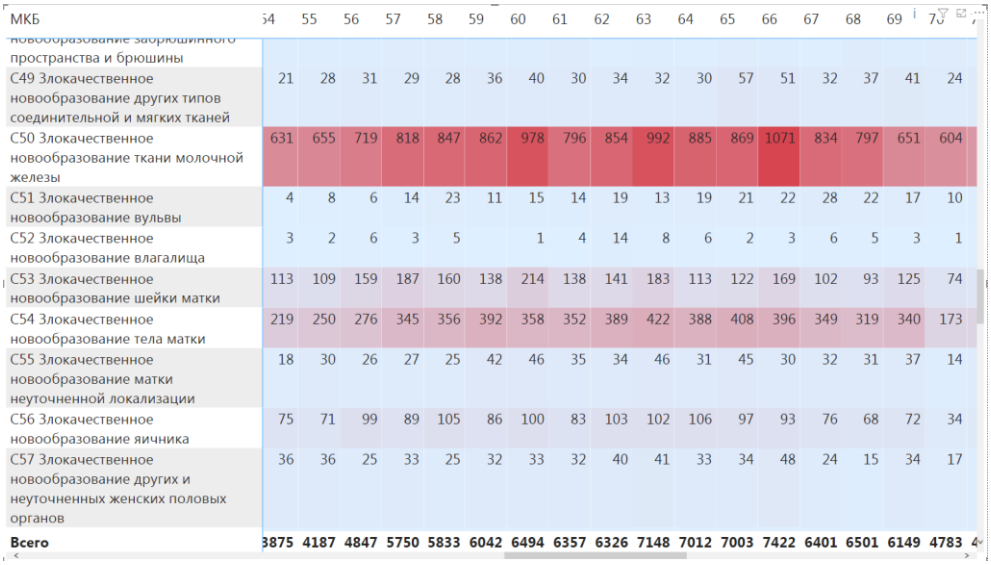
The use of technologies for the analysis of biomedical data opens up new opportunities in the development of medical information systems and improving the quality of healthcare management [14]. For the study, we used impersonal biomedical data contained in the regional information system "Electronic medicine" of the Bryansk region. The data covers the period from 2009 to March 2021 and contains fragments of integrated electronic medical records of patients.

The data volume is about 100 Gigabytes. The database contains 304 tables with a total of 2062 fields, which, in the absence of documentation, makes reverse engineering and data extraction for analysis in an integral form a very difficult and non-trivial task. 53,182,278 records of doctors' appointments and hospitalizations of patients were extracted from the database, from which 1,773,148 records were selected for exploratory analysis containing diagnoses with codes of the international classification of diseases related to oncological diseases (ICD C00-C97).

To perform exploratory data analysis, a set of dashboards was developed on the Microsoft Power BI platform (Fig. 1-11) using various visualization metaphors [15, 16]. For the analysis, we used only the cases when the patient was diagnosed for the first time, without taking into account repeated visits for the same reason. There were 195,634 such records. Figure 1 shows a matrix in which the rows contain nosologies, and the columns are the age at which the patient was diagnosed with cancer. At the intersection of a row and a column, the number of patients of this age and with such a diagnosis is given. The closer the background color is to red, the closer the quantity is to the maximum, the closer the background color is to blue, the closer the quantity is to the minimum. Fig. 2 shows a map of the distribution of the number of patients in the context of the stages at which neoplasms were diagnosed, depending on the district of the Bryansk region. The given data are incomplete, since not for all patients it was possible to correctly determine the locality of residence, encoded in the database in the FIAS format (KLADR). This issue requires further study. Fig. 3 shows a graph showing the number of patients (195,634) who were diagnosed with oncology (Y-axis) depending on the age (X-axis) at which this diagnosis was made, for all types of oncology (ICD C00-C97).

The X axis represents the quantized (10 years) age of the patient (Fig. 4) at which the diagnosis was made, the color indicates the stage at which the oncology was detected. It was found that for ICD C56 Malignant neoplasm of the ovary, almost half of the cases are detected only at the third stage, which is quite late, since at stages 3 and 4 the prognosis during treatment is much worse than at stages 1 or 2.

In the process of exploratory data analysis, an unusual dependence was found between the number of diagnosed oncologies on the age of the patient and the year of diagnosis. Fig. 5 shows the dependence of the number of diagnosed oncology cases on the patient's age for all years for ICD "C50 Malignant neoplasm of breast tissue" (25,195 patients) and C61 Malignant neoplasm of the prostate gland (indicate the number of patients 6,884). As mentioned above, these are some of the most common diagnoses. When we are analyzing Fig. 5, it can be seen that the age distribution is quite close to normal for both nosologies with a small additional peak on the right side.



МКБ	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70
новообразование забрюшинного пространства и брюшины																	
C49 Злокачественное новообразование других типов соединительной и мягких тканей	21	28	31	29	28	36	40	30	34	32	30	57	51	32	37	41	24
C50 Злокачественное новообразование ткани молочной железы	631	655	719	818	847	862	978	796	854	992	885	869	1071	834	797	651	604
C51 Злокачественное новообразование вульвы	4	8	6	14	23	11	15	14	19	13	19	21	22	28	22	17	10
C52 Злокачественное новообразование влагалища	3	2	6	3	5		1	4	14	8	6	2	3	6	5	3	1
C53 Злокачественное новообразование шейки матки	113	109	159	187	160	138	214	138	141	183	113	122	169	102	93	125	74
C54 Злокачественное новообразование тела матки	219	250	276	345	356	392	358	352	389	422	388	408	396	349	319	340	173
C55 Злокачественное новообразование матки неуточненной локализации	18	30	26	27	25	42	46	35	34	46	31	45	30	32	31	37	14
C56 Злокачественное новообразование яичника	75	71	99	89	105	86	100	83	103	102	106	97	93	76	68	72	34
C57 Злокачественное новообразование других и неуточненных женских половых органов	36	36	25	33	25	32	33	32	40	41	33	34	48	24	15	34	17
Всего	3875	4187	4847	5750	5833	6042	6494	6357	6326	7148	7012	7003	7422	6401	6501	6149	4783

Figure 1: The number of malignant neoplasms for the entire period by age

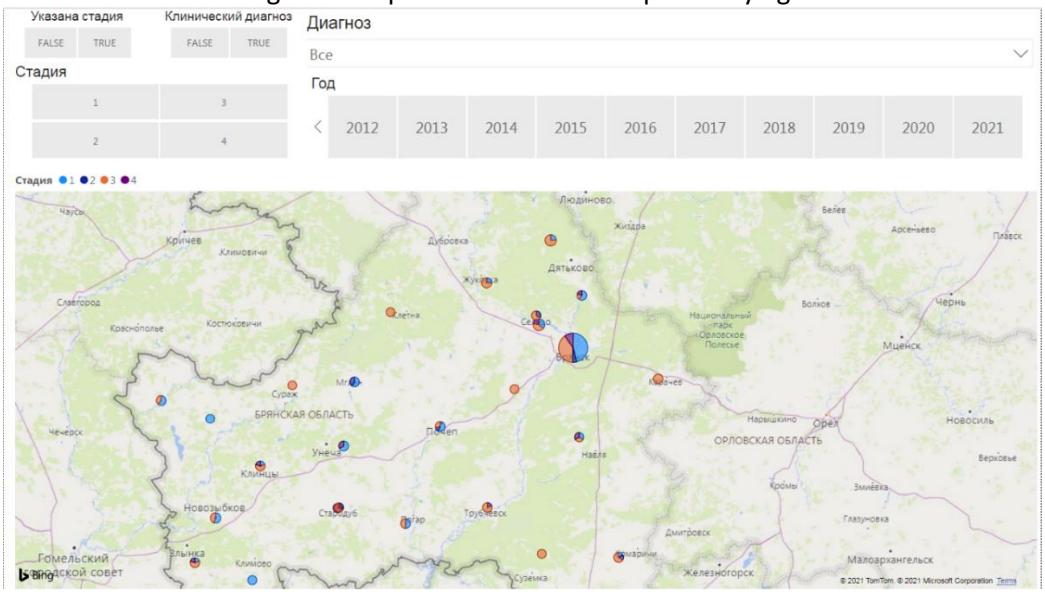


Figure 2: The number of malignant neoplasms in the context of stages and districts of the Bryansk region

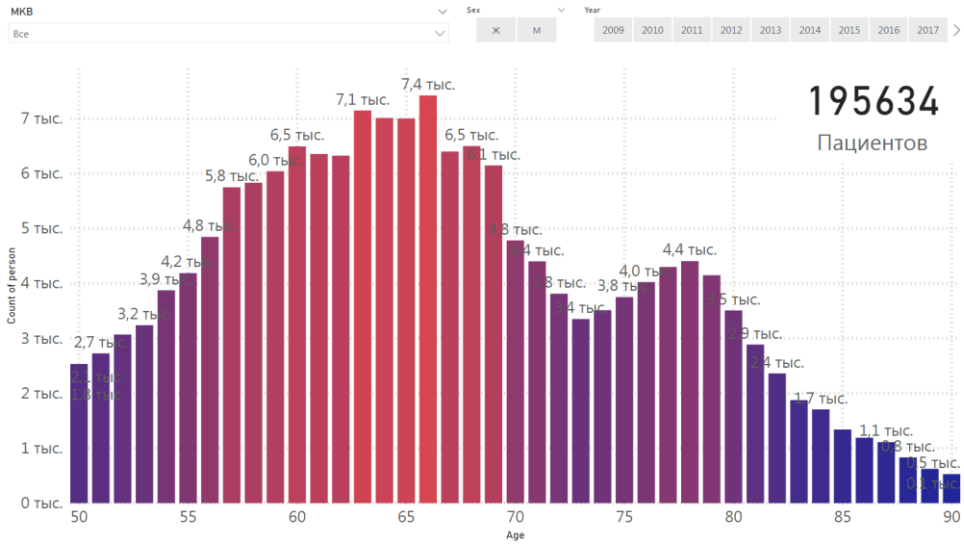


Figure 3: The number of diseases of malignant neoplasms for the entire period with grouping by age of patients

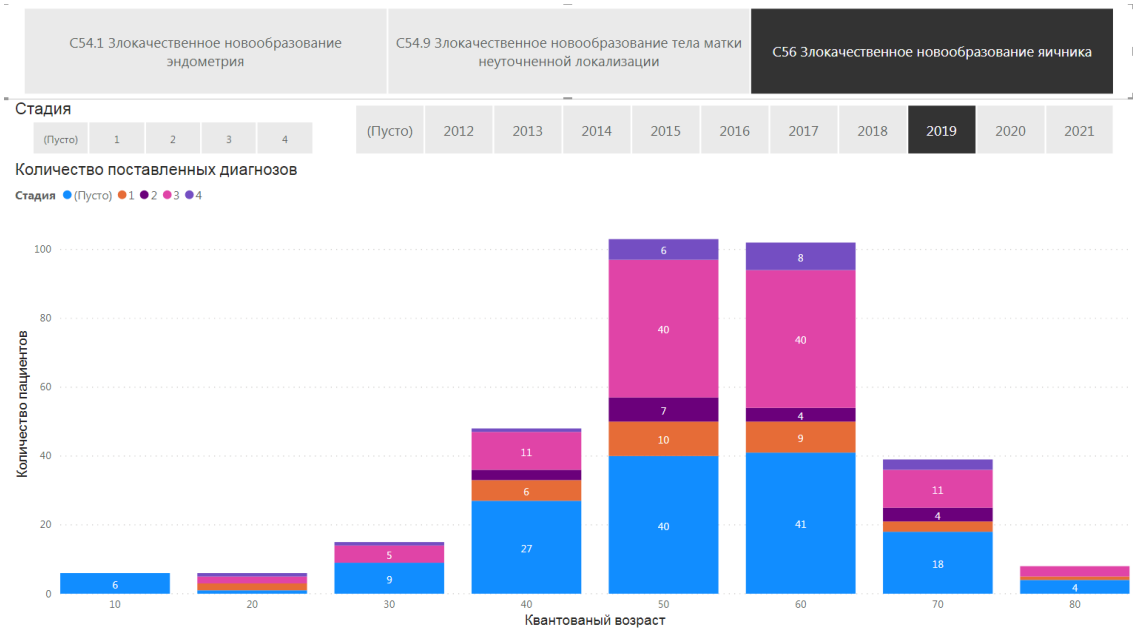


Figure 4: The number of malignant neoplasms in the context of quantized age and stages

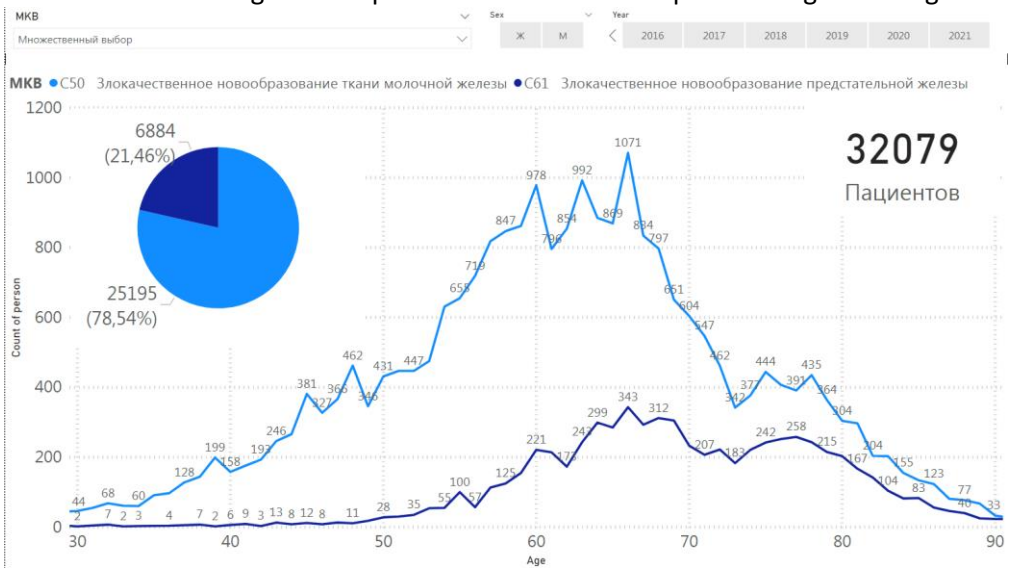


Figure 5: Malignant neoplasm of breast tissue (C50) and prostate (C61) for the entire period

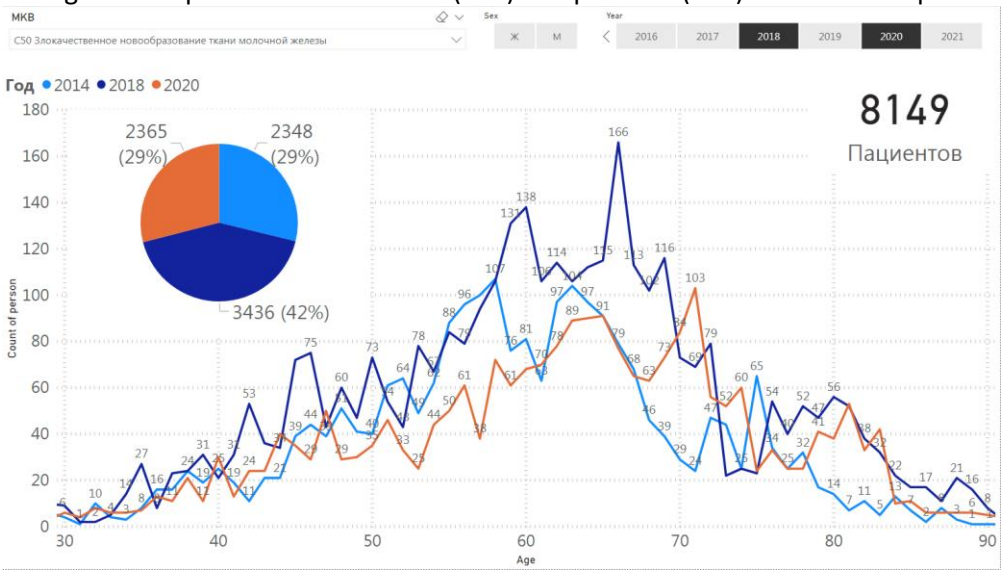


Figure 6: C50 Malignant neoplasm of breast tissue in 2014, 2018 and 2020

Fig. 6 shows graphs for "C50 Breast Tissue Malignancy" diagnosed in 2014 (2,348 patients), 2018 (3,436 patients) and 2020 (2,365 patients). In 2014, the peak was at the age of 58 years, the second peak at the age of 63, and a decline in the region of 60 years. In 2018, the peak was at age 66, the second peak at age 60, and there was a decline in the region of 61-64 years. In 2020, the peak was at age 71, the second peak at 63-65 years. As you can see, depending on the year of the diagnosis, the peak in diagnoses moves across different ages within a fairly wide range. We decided to investigate this phenomenon in more detail. Fig. 7-9 show graphs for C61 Prostate malignancy for 2014 (1,017 patients), 2015 (1,697 patients) and 2019 (784 patients). In 2014, the peak was at the age of 75 years, the second peak at the age of 66, and a decline in the region of 70 years. In 2015, the peak was at age 66, the second peak at age 74, and a decline around 70 years. In 2019, the peak was at age 70, the second peak at 66 years. The analysis shows that the degree of risk is highly dependent on the patient's age, and this dependence is not linear. Although the sampling for the entire period shows a distribution close to normal, but when analyzing for individual years, it can be seen that the peak of the distribution is shifting. In all likelihood, this process obeys certain patterns some laws, and it is extremely important to try to find them. This will allow predicting the peak incidence for a specific year or period and identifying the degree of risk of malignant neoplasms for different age groups.

Fig. 10 shows a comparison of the number of C56 malignant neoplasms of the ovary detected in 2019 in comparison with the number detected in the last year. The increase in the number of diseases is clearly visible. However, the reasons for this increase are unclear, since this may be associated with both an improvement in the quality of diagnostics and an increase in the number of people who have undergone medical examination, as well as an increase in the number of diseases.

Fig. 11 shows a comparison of the number of diseases in the context of the quantized age of patients. This allows us to see that for the listed nosologies, the peak is in the range from 60 to 69 years, and the second in terms of the number of diagnoses of this oncology is from 50 to 59 years.

Additionally, the Loginom low-code analytical platform was used to statistically evaluate the obtained data set. Figure 12 shows the result of the statistical evaluation. From which it follows that the most frequent age of oncological diagnosis is 62-66 years, and the next in importance are 67-71 and 57-61 years. For the rest of the fields, it was not possible to draw any significant conclusions based on the statistical assessment. Thus, exploratory analysis through the use of various visualizations allowed for a better understanding of the specifics of the data and identified various phenomena that need to be investigated in conjunction with medical professionals.

4. Using AutoML for the Analysis of Biomedical Data

The use of automated machine learning (AutoML) methods for the analysis (both exploratory and conventional) of biomedical data is gaining increasing popularity [17, 18], which allow in an automated mode to construct data mining models suitable for both exploratory analysis and decision making.

In article [17], the authors showed that AutoML methods can be used to analyze biomedical data and improve the productivity of specialists in solving a number of machine learning problems. It is also noted that the main limitation of AutoML at the moment is ineffective work on large datasets (dataset).

Fig. 13 shows a typical workflow for using AutoML to construct drill-down mining models. The scheme is based on the Cross-Industry Standard Process for Data Mining (CDISP-DM) methodology [19]. We reviewed the libraries of the most popular libraries that implement AutoML methods. We studied the capabilities of cloud services Google AutoML, Azure AutoML, Amazon Web Services AutoML and libraries Auto Keras, Auto-PyTorch, Auto-sklearn, Scikit-learn, Auto-WEKA, H2O AutoML, TPOT, MLBox and ML.NET. All of the above AutoML tools allow you to construct an DM model on a ready dataset in both automatic and automated modes. Many can perform selection of the optimal method from a predetermined set, selection of the optimal parameters of the selected method, limited data transformation (normalization, quantization, etc.) and optimization of the input data set (remove insignificant fields). However, none of the above libraries contains methods and tools that allow in an automated mode to select data that are optimal for constructing an DM model and perform their full preprocessing and transformation. The preparation of samples for training requires highly qualified specialists and a lot of time when processing large databases and data warehouses, which include sources of biomedical data.

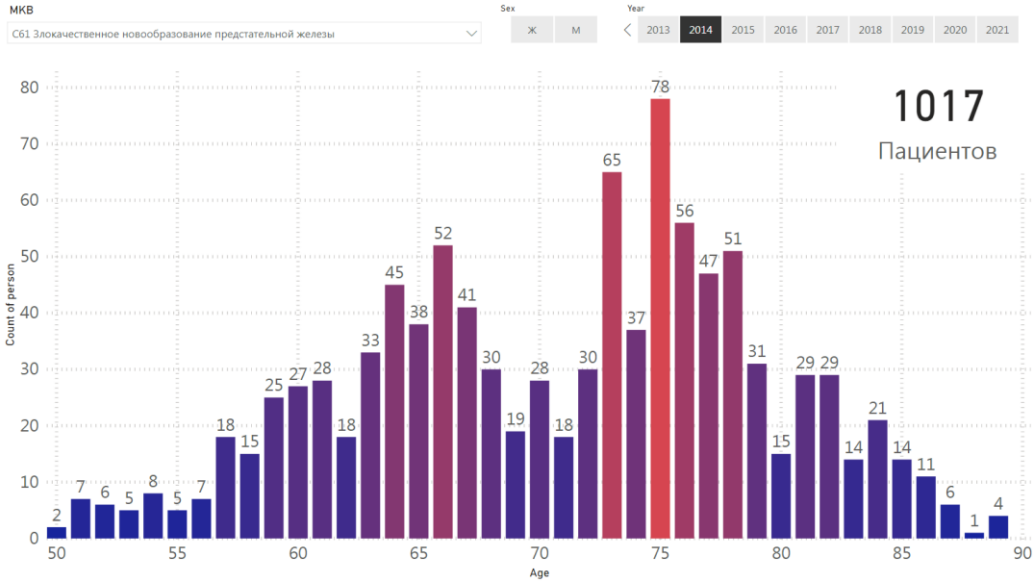


Figure 7: C61 Malignant neoplasm of the prostate in 2014

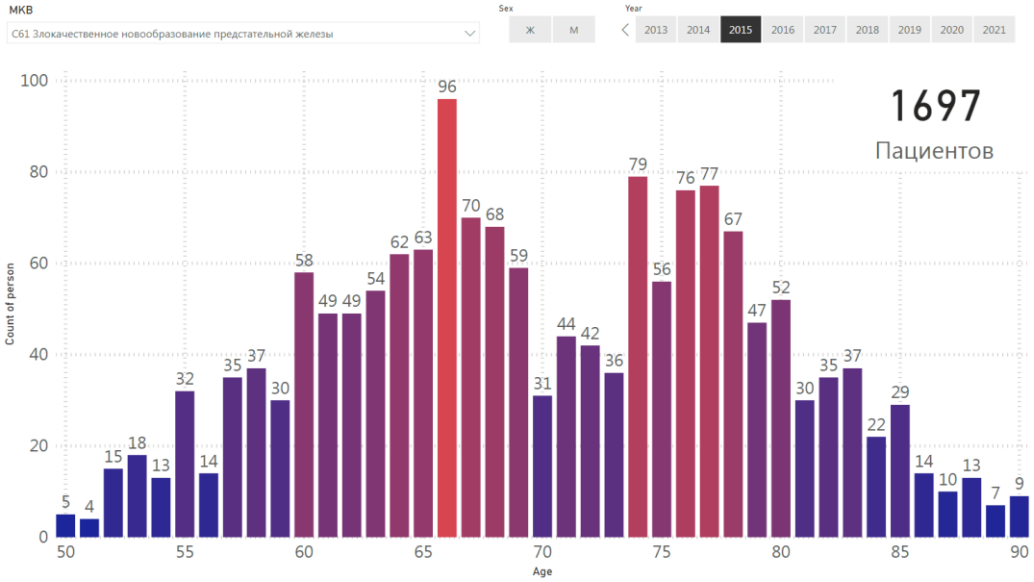


Figure 8: C61 Malignant neoplasm of the prostate gland in 2015

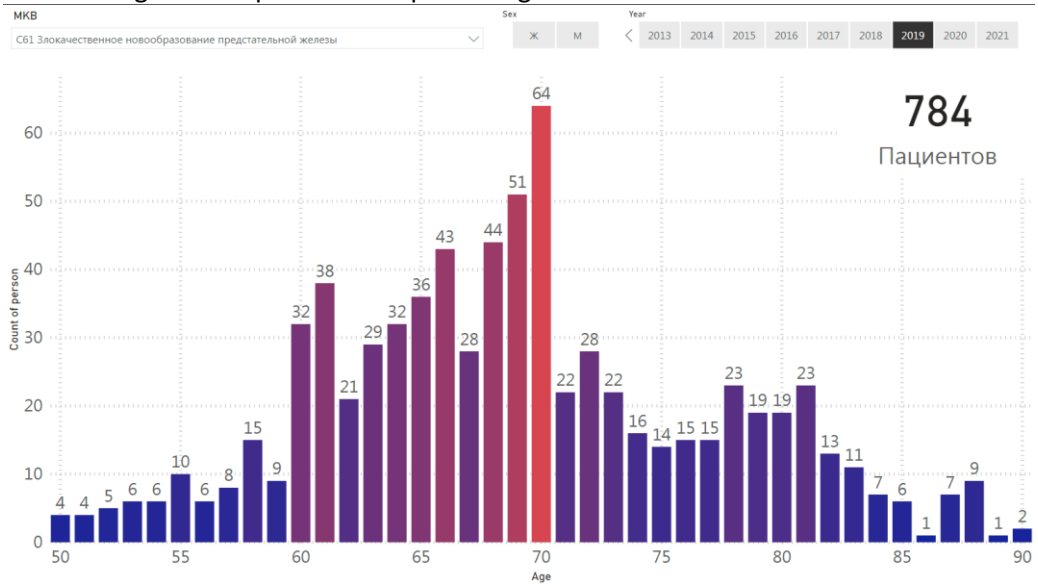


Figure 9: C61 Malignant neoplasm of the prostate in 2019

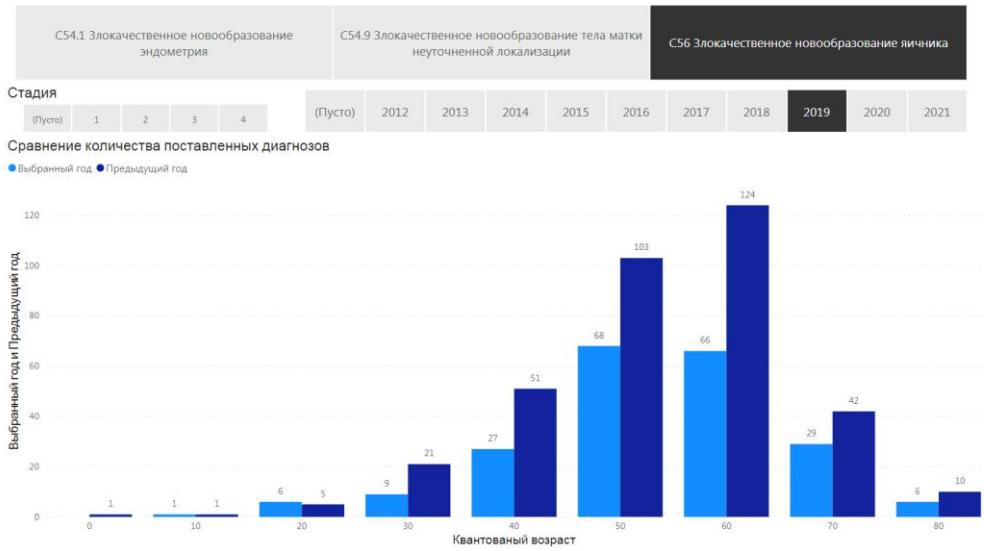


Figure 10: Comparison of the number of diseases detected in 2019 C56 Malignant neoplasm of the ovary with the last year

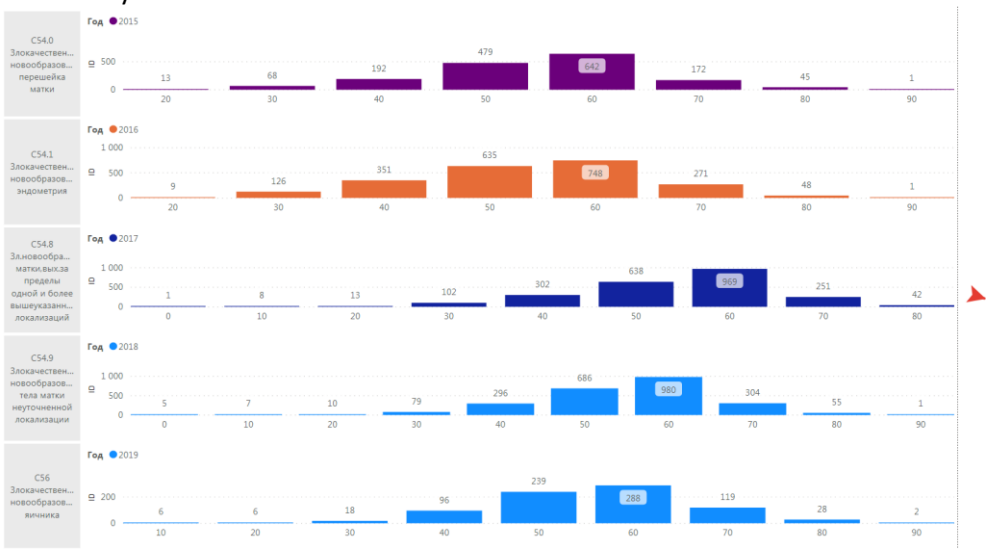


Figure 11: Comparison of the number of diseases in the context of the quantized age of patients

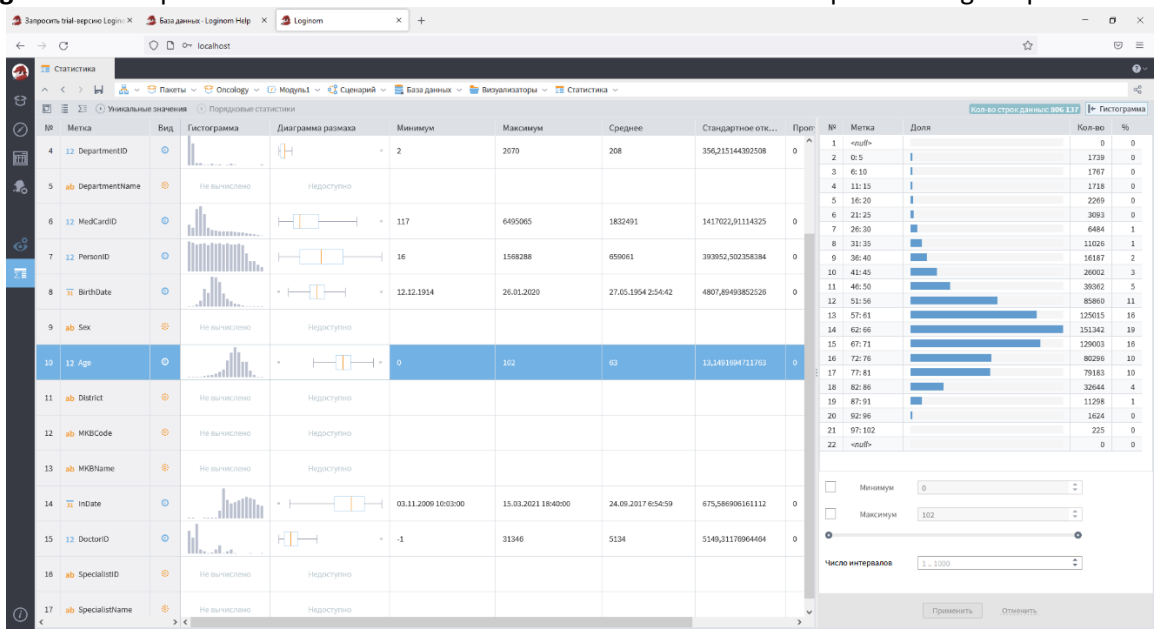


Figure 12: Statistical evaluation of a dataset using Loginom

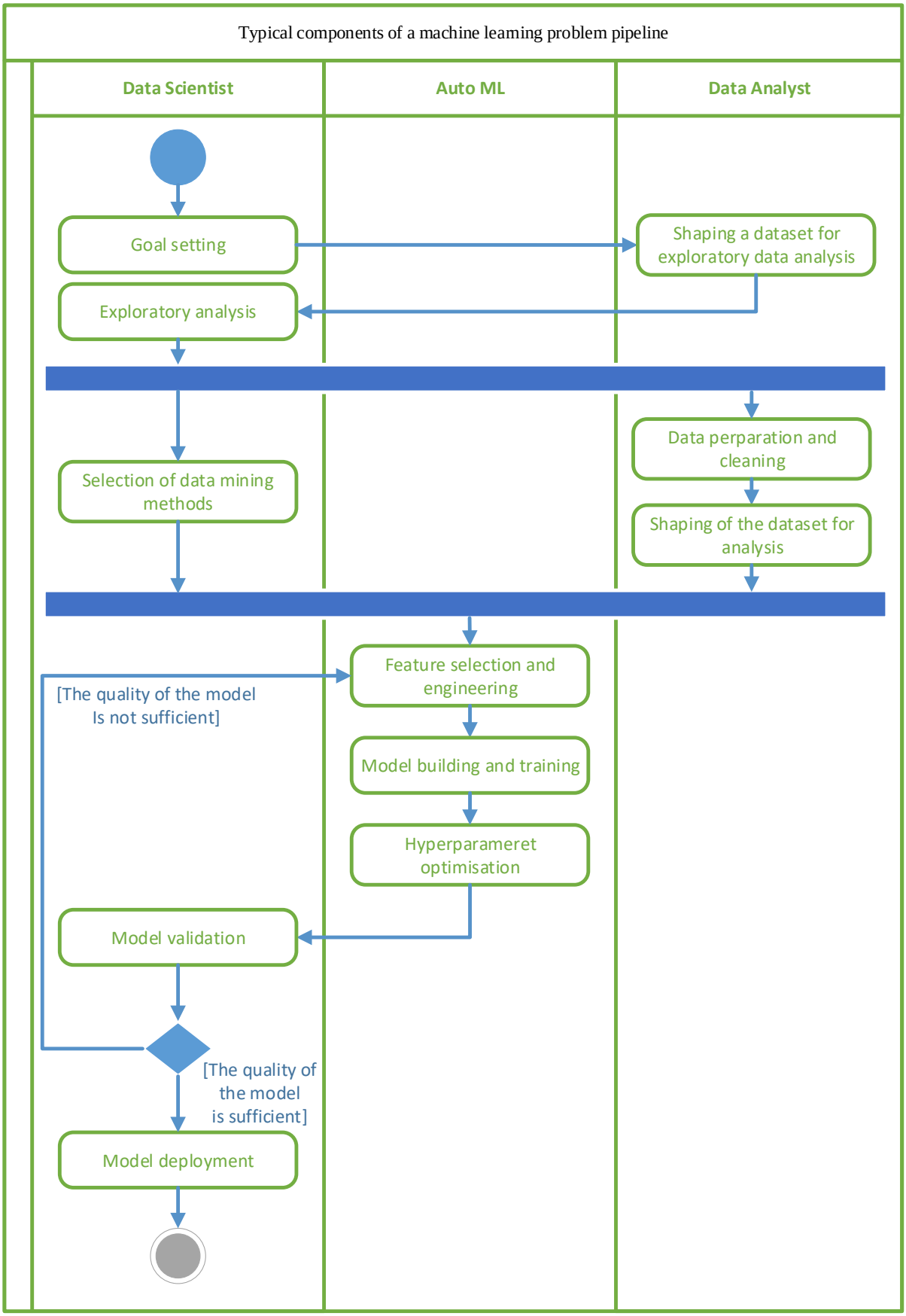


Figure 13: Typical components of a machine learning problem pipeline

The use of the ML.NET [20] and Auto-WEKA [21] libraries for constructing a model for assessing the degree of risk of oncological diseases based on integrated electronic medical records of patients in the Bryansk region did not allow constructing a model of any value. The preprocessing methods available in these libraries are intended for normalized data, which did not allow using all the information available in the database to the fullest. Most of the AutoML libraries are focused on supporting neural network models, a significant drawback of which is the inability to explain and substantiate the results obtained. To assess the risk of oncological diseases, it is advisable to use scoring models based on logistic regression, which allow an expert to assess the contribution of each of the factors to the results obtained and more reasonably prescribe additional laboratory tests to the patient. Since AutoML methods applied directly to the available dataset did not lead to significant results and are unsuitable for replicating DM models, we have developed a technique for replicating DM models based on AutoML methods. Fig. 14 shows the process of replicating the DM models for a number of nosologies based on the reference model. Before starting the process, Data Scientist constructs a reference model for any nosology according to the scheme shown Fig. 11. After that, he needs to set the parameters for replicating models: a list of ICDs for which it is necessary to construct models, acceptable data mining methods for constructing models, a set data, a list of data preprocessing methods, model optimality criteria and accuracy requirements. The Meta ML library generates parameters for constructing a basic model for the i-th nosology based on the parameters of the reference model.

After that, using the Auto ML library, the process of constructing a basic DM model is carried out. Next, Meta ML generates K sets of parameters for constructing a set of DM models and prepares data sets for constructing these models. Model constructing is done using Auto ML. All built models are saved to the repository. When the models are built for all combinations of parameters and data sets, the optimal model for the i-th nosology is selected from the repository and its parameters are compared with the parameters of the base model. The best model is published. The generation of models is repeated for N nosologies. As a result, based on the reference model, models will be built for all the indicated nosologies.

5. Conclusions

1. To effectively flight socially significant diseases, a personalized preventive approach to risk assessment, early diagnosis of malignant neoplasms in the screening process and targeted correction of negative changes at the level of the individual, cohort, and population are required.
2. It seems promising to use data mining models to increase the likelihood of detecting malignant neoplasms in the early stages during medical examination. To use biomedical data in the task of assessing the risk of malignant neoplasms oncology requires significant preprocessing and data preparation, which cannot be performed using the existing Auto ML libraries.
3. The use of various visual metaphors in the process of exploratory analysis of biomedical data allows you to better understand the specifics of the data and discover various patterns, dependencies and phenomena that must be taken into account when constructing data mining models.
4. As shown by exploratory analysis, risk assessment with an acceptable level of accuracy requires the construction of a separate predictive model for each nosology, and regular (at least once a year, or better more often) adjustment (retraining) of these models to correctly take into account the patient's age.
5. Although the overall distribution of patients by age is close to normal distribution, there is a shift in the peak incidence over time, which will have a significant impact on the estimated level of risk. In all likelihood, this process obeys certain patterns, and it is extremely important to find them. This will allow predicting the peak incidence for a specific year or period and identifying the degree of risk of malignant neoplasms for different age groups.
6. Since the total number of nosologies is large enough, a single model will have very mediocre quality indicators, and it is very costly to construct (and keep up to date) a large number of models, it is advisable to use the proposed method of replicating models and develop a special library for sampling and preprocessing data from integrated electronic medical records and subsequent launch of Auto ML methods and tools.

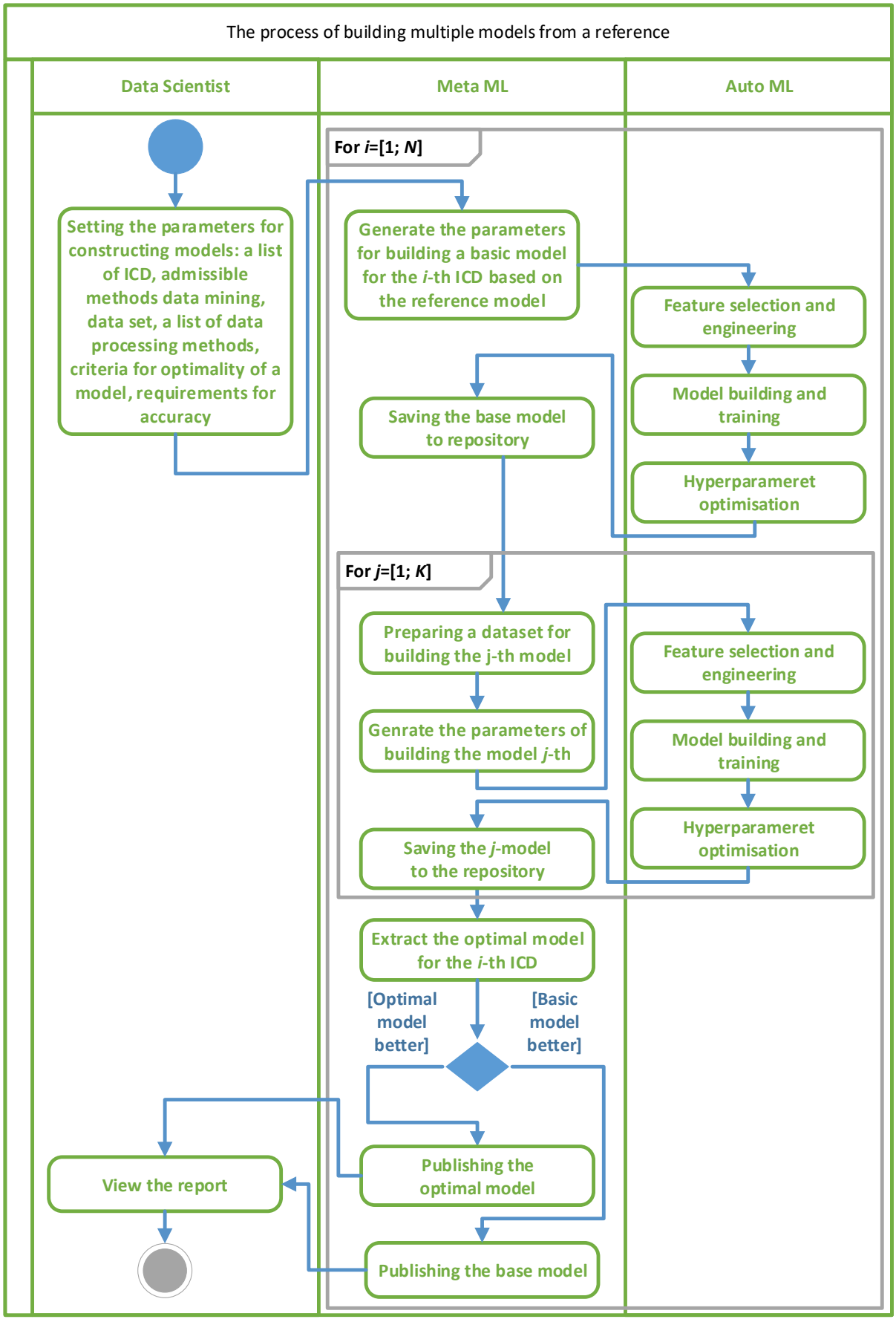


Figure 14: The process of constructing multiple models from a reference

7. When constructing models for assessing the risk of malignant neoplasms in addition to data from the integrated electronic medical record, it is advisable to use patient history data, data obtained after vectorization of laboratory tests, aggregated data on visits by specialized specialists and related diagnoses, data from online patient questionnaires filled in the process of undergoing medical examination, as well as data on environmental pollution (chemical, physical, biological contamination) and the work of the population in dangerous manufacture.

6. Acknowledgements

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

7. References

[1] O.M. Gerget, Bionic models for identification of biological systems, Journal of Physics: Conference Series 803 (2017) 012046. doi:10.1088/1742-6596/803/1/012046
[2] V.V. Danilov, I.P. Skirnevsky, O.M. Gerget, Segmentation of anatomical structures of the heart based on echocardiography, Journal of Physics: Conference Series 803 (2017) 012031. doi:10.1088/1742-6596/803/1/012031
[3] Information technologies in healthcare of the Russian Federation, Zdrav Expert 01.07.2021 (in Russian), URL: https://zdrav.expert/index.php/
[4] World Health Organization. International agency for research on cancer: Press release No. 292, 15 December 2020. URL: https://www.iarc.who.int/wp-content/uploads/2020/12/pr292_E.pdf
[5] B.F. Hankey, E.J. Feuer, L.X. Clegg et al., Cancer surveillance series: interpreting trends in prostate cancer – Part I: evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. J. Natl. Cancer Inst. 1999. vol. 91. P. 1017-1024. doi:10.1093/jnci/91.12.1017
[6] F. Bray, M. Colombet, L. Mery, Cancer Incidence in Five Continents, Vol. XI (electronic version), Lyon: International Agency for Research on Cancer, 2018. P. 67-72.
[7] A.D. Kaprin, V.V. Starinsky, A.O. Shahzadova, The state of cancer care for the population of Russia in 2019, Moscow, Scientific Research Institute of Oncology named after P.A. Herzen, 2020 (in Russian). – 239 p. URL: https://glavonco.ru/cancer_register/%D0%9F%D0%BE%D0%BC%D0%BE%D1%89%D1%8C%202019.pdf.
[8] V.I. Chissov, V.V. Starinsky, G.V. Petrova, Malignant neoplasms in Russia in 2009: morbidity and mortality, Moscow, Scientific Research Institute of Oncology named after P.A. Herzen, 2011. – 259 p. (in Russian).
[9] C.J. Stein, G.A. Colditz, Modifiable risk factors for cancer, British Journal of Cancer. 2004. 90(2), 299-303. doi:10.1038/sj.bjc.6601509
[10] Carcinogenesis: leadership, RAMS, Russian Cancer Research Center, Research Institute of Carcinogenesis, Moscow, Medicine, 2004, 574 p. (in Russian)
[11] J. Ferlay, M. Ervik, F. Ervik, M. Colombet, L. Mery, M. Piñeros, et al., Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer, 2020, URL: https://gco.iarc.fr/today
[12] WHO cancer information. 3 March 2021, URL: https://www.who.int/news-room/fact-sheets/detail/cancer
[13] A.V. Yablokov, On the concept of population load (review), Hygiene and sanitation, 2015, No. 6, p. 11-14,
[14] A.A. Zakharova, D.G. Lagerev, A.G. Podvesovskii, Multi-level Model for Structuring Heterogeneous Biomedical Data in the Tasks of Socially Significant Diseases Risk Evaluation, in: A.G. Kravets et al. (Eds.), CIT&DS 2019, Communications in Computer and Information Science, Vol. 1084, Springer Nature Switzerland AG 2019, pp. 461-473. doi:10.1007/978-3-030-29750-3_36
[15] A.A. Zakharova, A.G. Podvesovskii, A.V. Shklyar, Visual and Cognitive Interpretation of Heterogeneous Data, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W12 (2019) 243-247. doi:10.5194/isprs-archives-XLII-2-W12-243-2019

[16] A.A. Zakharova, E.V. Vekhter, A.V. Shklyar, Methods of solving problems of data analysis using analytical visual models, Scientific Visualization 9 (4), 78-88 (2017). doi: 10.26583/sv.9.4.08
[17] Jonathan Waring, Charlotta Lindvall, Renato Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, Artificial Intelligence in Medicine, Volume 104, 2020, 101822, ISSN 0933-3657. doi:10.1016/j.artmed.2020.101822.
[18] A. Mustafa, M. Rahimi Azghadi, Automated Machine Learning for Healthcare and Clinical Notes Analysis, Computers 2021, 10, 24. doi:10.3390/computers10020024
[19] Christoph Schröer, Felix Kruse, Jorge Marx Gómez, A Systematic Literature Review on Applying CRISP-DM Process Model, Procedia Computer Science, Volume 181, 2021, pages 526-534, ISSN 1877-0509. doi:10.1016/j.procs.2021.01.199.
[20] ML.NET URL: https://dotnet.microsoft.com/apps/machinelearning-ai/ml-dotnet
[21] Auto-WEKAURL: http://www.cs.ubc.ca/labs/beta/Projects/autoweka/