GENEALOGICAL DIAGNOSTICS OF NEOPLASMS
BASED ON ARTIFICIAL INTELLIGENCE SYSTEMS

Abstract. The compilation and analysis of the patient’s genealogies is one of the methods of population genetics, which makes it possible to identify a predisposition to a particular oncological pathology. At present, it is relevant to prove the feasibility of developing and introducing into clinical practice a comprehensive method for diagnosing and preventing tumors based on data from genetic counseling, molecular biological research and modern artificial intelligence technologies. An information-analytical system is proposed that allows analyzing the patient’s data obtained during the consultation, with the possibility of supplementing them with information from the medical history and the results of the study. The proposed information system is able to analyze the genealogy and give a preliminary conclusion about the risk of a tumor process in the patient’s family members, according to the algorithms of the morbidity accumulated in the region.

Keywords: oncology, genetic counseling, artificial intelligence, diagnostic system


© Кузнецов О. Е., 2023

Introduction. According to the latest World Health Organization (WHO) estimates, more than 60% of countries are currently implementing their own eHealth strategies, both nationally and regionally. The Ministry of Health of the Republic of Belarus has developed the concept of e-health. The concept presents the basic principles of building an e-health system in the country and the possibilities for its integration into a nationwide automated information system. It is planned that during its implementation a centralized health information system (CHIS) will be created. This will create opportunities for the formation and maintenance of a single information archive of patients and the prompt provision of medical data. In addition, a clinical decision information support system will be introduced. [1]

Oncological diseases steadily occupy a leading place in the structure of mortality of the population. Despite the fact that the incidence of death from cancer increases with age, neoplasms remain one of the leading causes of death among the young population [2].

© Кузнецов О. Е., 2023
In recent years, there have been changes in the statistics of oncological morbidity: breast cancer (11.7%) has surpassed lung cancer (11.4%) in frequency of detection, followed by colorectal cancer (10%), prostate cancer (7.3%), stomach cancer (5.6%) and liver cancer (4.7%) [3–5].

Belarus does not differ much in terms of morbidity and mortality from neighboring countries: about 50,000 patients with newly diagnosed cancer are diagnosed annually. The country is included in the group of countries with relatively low incidence rates, but at the same time, over the past decade, the incidence has tripled [6].

It is generally accepted that about 10% of neoplasms are classified as hereditary forms. These include both inherited oncological syndromes and families in which close relatives are affected by various forms of cancer. Of all new cancer cases, about 16% are associated with various types of chronic infections caused by viruses [7, 8].

With the development of molecular-genetic research in oncology, it became possible to detect and track early (preclinical) forms of cancer, which makes it possible to form and carry out timely preventive measures for people at risk. The compilation and analysis of pedigrees, taking into account etiopathogenetic factors confirmed by molecular genetic studies, is one of the most important methods of population genetics, which makes it possible to identify a predisposition to a particular pathology.

In oncology, as in no other area of medicine, the cost of a diagnostic error is high. The main difficulty lies in the early detection of a malignant process. Recently, scientists have begun to actively use artificial intelligence for diagnostics. Artificial intelligence (AI) systems, which have recently appeared in medicine, are, in our opinion, one of the promising areas in global healthcare. Artificial intelligence technologies today are changing the global healthcare system, allowing to re-evaluate and modernize the diagnostic system, as well as generally improve the quality of healthcare services while reducing costs for medical clinics [9, 10].

The use of smart systems can improve the accuracy of diagnosis, as shown in the Frost & Sullivan study, where artificial intelligence technologies increase the accuracy of diagnosis by 30–40%. Pathologist Andy Beck from Harvard Medical School believes that the continued use of AI technologies will reduce the level of errors in diagnosis by 85% [11].

Observing such changes in the medical field, many experts express concern and fear that “smart” systems will take the place of a person. Artificial intelligence is unlikely to ever replace the doctor. Thanks to the system, which has absorbed and analyzed a huge amount of information, the doctor will be able to perform his work more efficiently and successfully. A significant number of decisions in the field of oncology, in our opinion, will be based on the study of databases: criteria for the manifestation of the disease, research results, etc. It follows from this how important it is to create and expand databases, to record in them a variety of cases of disease, criteria for their manifestation and make this data set available for study. The use of such bases for the development and improvement of “smart” diagnostic systems will help specialists move towards the goal: a future where cases of cancer detection in the last stages or misdiagnosis will become nonsense [12].

The available domestic information developments in the field of artificial intelligence on this issue are not complete to date.

It seems expedient to develop and put into practice an integrated method for diagnosing and preventing malignant tumors based on artificial intelligence systems (medical information system). This scientific and practical direction is based on the proof of the contribution of genealogical counseling and molecular biology methods to the diagnosis and prevention of malignant tumors, as well as the reduction of mortality.

The aim of the study is the evaluation of algorithms for predicting oncological risk (genealogical) based on an artificial intelligence system.

**Materials and research methods.** The material of the study was data on the incidence of patients (age, year of development of the disease and year of death, gender, incidence of relatives) available in the regional oncological registry, information obtained during counseling when contacting a healthcare institution, and data from a molecular biological study of individuals, both with existing malignant tumor, and in its absence, as well as data of persons with a clinical risk of developing a tumor process, including hereditary tumors.
The study was carried out in the Grodno region (2015–2021): population, hospital screening and screening of the oncological register of patients in the region (data of the oncological register for the period 1960–2007). The study included individuals diagnosed with colorectal cancer, ovarian cancer and breast cancer. Completions of coverage – 13,182 people: 612 patients of the oncological dispensary who applied for medical help, including 196 people hospitalized for differential diagnosis of the process, 12,570 practically healthy respondents. Thanks to the survey method (data on diseases of relatives were provided), the family history of 10,706 respondents was processed (including 401 people based on the results of hospital screening). To study the clinical and genealogical features of the tumor, 865 (47.2 %) of 1833 patients with diagnosed tumors were selected for analysis in the registry.

The implementation of measures for early detection and possible prevention of tumor processes is based on the algorithm of theorems of elementary probability theory and methods of statistical analysis, which allow determining the probability of an event, provided that another statistically interdependent event has occurred [13].

To carry out activities, a computerized information and analytical system for recording and monitoring oncological diseases (SMS) was implemented, which allows analyzing patient data obtained during counseling, with the possibility of supplementing them with information from the oncological register and data from laboratory genetic examination [14].

The information-analytical system for registration and monitoring of oncological diseases (IMS) is implemented as a computer information-analytical program in the Microsoft SQL Server environment. SQL Server is a relational database management system (RDBMS) developed by Microsoft Corporation. The main query language used is Transact-SQL. Transact-SQL is an implementation of the ANSI/ISO standard for Structured Query Language (SQL) with extensions, used to work with databases ranging in size from personal to large, enterprise scale.

The IMS allows you to process accumulated data, including on the basis of drawing up pedigrees, calculate, analyze and give a preliminary conclusion about the risk of a malignant tumor in the respondent and his family members according to risk algorithms, calculate the population risk, assign the patient to clinical risk groups, export data for subsequent further analysis. Since the information stored in the IMS database is confidential, persons who are allowed access to the system are registered and receive personalized access (login, password).

Implemented in the software package, artificial intelligence methods and analysis methods based on probability theory (Bayes formula) make it possible to determine the likelihood of a possible tumor process, provided that another interdependent event has occurred, and, based on statistical analysis, calculate the population risk for a given nosological form. diseases.

To maintain and track information, the modularity of the program block was developed and taken into account: “registration”, “relative”, “family tree”, “cancer registry”, “examination”, “diseases”, “screening”, “risk”.

Results and its discussion. Based on the known risk factors for the development of tumors and taking into account the significant factors of their development (age; chronic diseases; carriage of mutations – the risk of developing a tumor throughout life is 80–85 %), “family oncological history”, which is a significant diagnostic sign and often not analyzed [15], we analyzed the above factors to evaluate possible algorithms for predicting oncological risk.

Analysis of the results of respondents during population, hospital screening and cancer registry screening made it possible to distribute patients into groups: no oncological risk, suspected presence of a tumor, and high oncological risk. The results obtained are shown in the Table.

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cancer Registry Screening:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high cancer risk</td>
<td>7</td>
<td>1.68</td>
</tr>
<tr>
<td>suspicion of a tumor</td>
<td>95</td>
<td>22.84</td>
</tr>
<tr>
<td>no cancer risk</td>
<td>314</td>
<td>75.48</td>
</tr>
<tr>
<td>total</td>
<td>416</td>
<td></td>
</tr>
<tr>
<td><strong>Hospital screening:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high cancer risk</td>
<td>13</td>
<td>6.63</td>
</tr>
<tr>
<td>suspicion of a tumor</td>
<td>53</td>
<td>27.04</td>
</tr>
<tr>
<td>no cancer risk</td>
<td>130</td>
<td>66.32</td>
</tr>
<tr>
<td>total</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td><strong>Population screening:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high cancer risk</td>
<td>10</td>
<td>0.08</td>
</tr>
<tr>
<td>suspicion of a tumor</td>
<td>52</td>
<td>0.41</td>
</tr>
<tr>
<td>no cancer risk</td>
<td>12 508</td>
<td>99.51</td>
</tr>
<tr>
<td>total</td>
<td>12 570</td>
<td></td>
</tr>
</tbody>
</table>
As can be seen, among people with a diagnosed malignant neoplasm (colorectal cancer, ovarian cancer, breast cancer), when screening the region’s cancer registry, the incidence of people with a high risk of developing a tumor process was 1.68 % of the total number ($p < 0.05$). In the analysis of persons being treated in a hospital (hospital screening), 13 patients were identified with a high oncological risk of developing a tumor and 53 with a suspected presence of a neoplasm: 6.63 % and 27.04 %, respectively.

Calculation of the population risk (for healthy children of the proband, siblings of the proband, children of the siblings of the proband), which is carried out on the basis of the Bayes formula in several stages, was carried out automatically based on the coefficients generated by the “intelligent” system. First, the degree of risk of developing tumors (in points) in a healthy proband is determined in the event that his relative suffers from an oncological disease. For this, the formula is used:

$$x = \log(\frac{ya}{zb}),$$

where $a$ – population in the region, $b$ – general oncological incidence in the region, $z$ – number of persons with a given category of relationship (in the IMS database), $y$ – the number of patients with a malignant process among persons of this category of kinship (in the IMS database), $x$ – degree of risk (in points).

The risk coefficient for a patient/proband to get sick is 0.384 if the father is sick, 0.494 for the mother, 0.398 for the sister, 0.435 for the brother, 0.043 for the mother’s sister, and 0.046 for the father’s sister.

The presence of two or more oncological diseases in one of the relatives increases the risk of disease in the proband. The degree of this additional risk is calculated by the formula:

$$x_1 = \log(\frac{y_1c}{zd}),$$

where $z$ – the number of persons with this category of relationship (in the IMS database), $y_1$ – the number of patients with two or more oncological diseases among persons with this category of kinship (in the IMS database), $c$ – number of persons of the same sex as the affected relative in the cancer registry, $d$ – the number in the oncology register of persons of the same sex as the diseased relative who suffer from two or more oncological diseases, $x_1$ – the degree of additional risk associated with the presence of two or more oncological diseases in relatives (in points).

According to our data, the degree of additional risk for a proband to get sick if his father suffers from two or more oncological diseases is 0.245, mother – 0.220, sister – 0.202, brother – 0.046, mother’s sister – 0.022.

It is also known that the risk of developing tumors in a proband increases if his relatives (one of the relatives) developed oncological diseases at a younger age. The additional risk associated with the age at which relatives fell ill is calculated by the formula:

$$x_i = \log(\frac{n_i c}{zk_i}), i = 1, 2, 3,$$

where $z$ – the number of persons with this category of relationship (in the IMS database), $n_i$ – the number of persons with a given category of kinship who developed the disease in the $i$ age range (in the IMS database), $c$ – number of persons of the same sex as the affected relative in the cancer registry, $k_i$ – the number of persons in the cancer registry of the same sex as the sick relative who developed the disease in the “$i$” age range, $x_i$ – the degree of additional risk associated with the age at which the oncological disease developed in a relative (in points), $i = 1, 2, 3$ correspond to the age range of 20–29 years old, 30–39 years old, 40–49 years old.

The additional risk for the proband to get sick, associated with the age at which the oncological disease developed in his parents, is:

- if the father was diagnosed in the age range of 20–29 years – 0.109, from 30 to 39 years – 0.070, from 40 to 49 years – 0.090.
- if the mother is diagnosed in the age range of 20–29 years – 0.030, from 30 to 39 years – 0.250, from 40 to 49 years – 0.243.

General (population) risk ($R$) is determined by the formula $R = 10^{(x+x_1+x_2)}$.

The assessment of the general risk allows to determine how many times the risk of developing tumors is higher in a patient/proband compared to persons whose heredity is not burdened with oncological diseases.
The analysis algorithm began with entering the patient/proband data into a software electronic form. The information contained in the questionnaire is entered into the program manually (with direct survey, personal computer, Windows software environment), the system automatically assigns information about the time, date of creation of the questionnaire and the number of the questionnaire. Data of the contingent that has passed the preliminary survey, screening or has already applied for medical care in an oncological healthcare institution (hospital screening, registry screening) can be imported from the oncological registry, if available. The search is performed automatically by key letters or words, which is important because it is not always possible to accurately identify the letters if the person’s handwriting is illegible. As additional search criteria, the use of the date of birth, the number of the patient’s medical or outpatient card has been implemented. If the proband is registered in the register, a request “register data” will appear on the screen, which will contain the information already available, as well as (if available) passport data of persons whose last name, first name, patronymic (or their fragments) match the search criteria.

After the choice is made, all the necessary information, including place of residence, personal number, clinical diagnosis (diagnoses), disease code according to the International Classification of Diseases of the 10th revision (ICD-10), medical record number, laboratory data (histological, cytological, molecular biological), as well as the number under which the patient is registered in the registry, are imported into the IMS database. At the next stage, the registration (questionnaire) of data on the relatives of the proband and the possible presence of oncological diseases is performed. For each of the relatives suffering from malignant neoplasms, an electronic questionnaire “diseases of relatives” is filled out, similar to the “passport data” form, and data from the register is also imported, if available.

The clinical risk group for the presence of a predisposition to the development of, for example, breast or colon cancer, as well as the results of a laboratory study on possible gene mutations, are entered into the appropriate fields of the database using ready-made directories. The system also carries out intelligent control and does not allow entering dubious data, as well as duplicating the registration of the same patient/proband/relative.

The IMS software package provides and implements the ability to visualize the pedigree, which greatly facilitates the process of assigning a proband and/or its relative to the group of clinical risk of developing a tumor process. When constructing a pedigree, generally accepted symbols are used (see Figure).

The proposed intellectual information system is capable of analyzing pedigrees and issuing a preliminary conclusion on the risk of a malignant tumor in family members according to planned and programmed algorithms.

So, for example, the applied diagnostic criteria:
A. Colorectal Cancer (CRC):
   in the pedigree 3 cases of CRC. Conclusion: “high risk of colorectal cancer”;
   in the pedigree 2 cases of CRC. Conclusion: “medium risk of colorectal cancer”; 
   in the pedigree 1 case of CRC and 1 or more cases of cancer of another localization that does not meet the previous criteria (that is, not breast cancer, not OC, not CRC). Conclusion: “low risk of colorectal cancer”;

“Pedigree”, section “screening”, screen view
in the pedigree 1 case of CRC that does not meet the previous criteria. Conclusion: “population risk of colorectal cancer”.

B. breast cancer (BC):
- in the pedigree 3 cases of breast cancer and/or even one case of breast cancer in a man. Conclusion: “high risk of breast cancer”;
- in the pedigree 2 cases of breast cancer. Conclusion: “medium risk of breast cancer”;
- in the pedigree 1 case of breast cancer under the age of 40 years, and/or breast cancer is combined with other tumors (primary multiple variant). Conclusion: “low risk of breast cancer”;
- in the pedigree 1 case of breast cancer that does not meet the previous criteria. Conclusion: “population risk of breast cancer”;
- any of the above options in combination with the presence of mutations in the BRCA 1/2 gene. Conclusion: “very high risk of breast cancer”.

C. ovarian cancer (OC):
- in the pedigree 3 cases of OC. Conclusion: “high risk of ovarian cancer”;
- in the pedigree 2 cases of OC. Conclusion: “medium risk of ovarian cancer”;
- in the pedigree 1 case of OC under the age of 40 years, and/or OC combined with other tumors (primary multiple variant). Conclusion: “low risk of ovarian cancer”;
- in the pedigree 1 case of OC that does not meet the previous criteria. Conclusion: “population risk of ovarian cancer”;
- any of the above options in combination with mutations in the BRCA 1/2 gene. Conclusion: “very high risk of ovarian cancer”.

D. Breast cancer and ovarian cancer (BC/OC):
- in the pedigree 2 cases of breast cancer and/or even one case of breast cancer in a man, along with one case of ovarian cancer. Conclusion: “high risk of breast/ovarian cancer”;
- in the pedigree 1 case of breast cancer + 2 cases of ovarian cancer. Conclusion: “high risk of breast/ovarian cancer”;
- in the pedigree 1 case of breast cancer and 1 case of ovarian cancer. Conclusion: “medium risk of breast/ovarian cancer”;
- any of the above options in combination with the presence of mutations in the BRCA gene. Conclusion: “very high risk of breast/ovarian cancer”.

The system automatically generates conclusions on all entered questionnaires and algorithms, which not only facilitates the search for persons in need of medical advice, but also forms certain risk groups for further analysis. The final decision to assign the proband and/or relatives to any risk group is made by a specialist after studying the pedigree.

Almost all indicators used to calculate population risk, with the exception of the population in the region, can be obtained directly from the IMS. This allows you to make the system self-learning, intelligent. Using the directory system, you can set almost any conditions for searching for information in the database (arbitrary query). The IMS uses generally accepted standard reference books: for making a diagnosis – ICD-10, for indicating the place of residence, etc.

To provide information support for outpatient (dispensary) monitoring of persons with an increased risk of developing a tumor process, the software package provides for a “monitoring” section. The section allows you to create a list of people who need to be monitored using targeted queries (selection for monitoring of healthy people, for example, relatives of the proband who have an increased risk of developing a tumor), people suffering from cancer and registered during hospital screening.

**Conclusion.** Organizational and scientific and methodological forms of providing consultative medical care to the population remain, although studied, but relatively in demand, so any experience in this area is of undoubted value.

It is advisable to introduce into clinical practice an integrated approach in monitoring and identifying individuals with tumor pathology, as well as in diagnosing and preventing tumors based on modern information “intelligent” systems, taking into account precision molecular biological markers.

The proposed algorithms, based on the coefficients generated by the “intelligent” system, made it possible to identify a high risk of developing a tumor process (colorectal cancer, ovarian cancer, breast cancer) in almost 0.5 % of the region’s population. The frequency of identifying individuals with a high
risk of developing cancer is higher in hospital screening than in the regional cancer registry screening \( (p < 0.002) \). According to the results of the study (hospital, population screening and registry screening), out of the total number of persons, the share of those included in the high-risk group for developing cancer was 8.39 %. These persons (including relatives) need further dynamic monitoring for the purpose of prevention, early diagnosis and effective treatment in case of detection of cancer forms, as well as an in-depth study of the clinical, morphological and molecular biological features of tumor forms.

The democratization of innovative technologies, in particular artificial intelligence, in such an important area as healthcare is a key goal for IT-developers in the world. “Smart” algorithms will make it possible to achieve a high level of accuracy in diagnosing diseases in the early stages, and, consequently, will prevent the disease or increase the effectiveness of treatment. All “smart” algorithms are created by people. Therefore, forms of such assistance to the population are valuable.

**Conflict of interests.** The authors declare no conflict of interests.

**References**


**Information about the author**

Aleh E. Kuzniatsou – Ph. D. (Biol.), Associate Professor, Director. Institute of Biochemistry of Biologically Active Compounds of the National Academy of Sciences of Belarus (7, Antonii Tyzengauz Sq., 230012, Grodno, Republic of Belarus). https://orcid.org/0000-0002-1348-8688. E-mail: olegkuznetsov@inbox.ru

**Информация об авторе**

Кузнецов Олег Евгеньевич – канд. биол. наук, доцент, директор. Институт биохимии биологически активных соединений НАН Беларуси (ул. Антония Тызенгауза, 7, 230012, г. Гродно, Республика Беларусь). https://orcid.org/0000-0002-1348-8688. E-mail: olegkuznetsov@inbox.ru