MEDICAL INFORMATICS

Tutorial
for foreign English-speaking students
of medical universities

PUBLIC HEALTH MINISTRY OF UKRAINE
Kharkiv national medical university

MEDICAL INFORMATICS

Tutorial
for foreign English-speaking students
of medical universities

МЕДИЧНА ІНФОРМАТИКА

Навчальний посібник
для іноземних англомовних студентів
медичних університетів
(*англійською мовою*)

Kharkiv
KhNMU
2019

УДК 61:004(075.8)
М 42

Approved by the Scientific Council of KhNMU. Protocol No 8 of 19.09.2019.

Reviewers:

Timanyuk V. A. – PhD in Physics and Mathematics, Professor, Department of Physics, National University of Pharmacy.

Berest V. P. – PhD in Physics and Mathematics, Associate Professor, Head of the Department of Biological and Medical Physics, V. N. Karazin Kharkiv National University.

Authors:

Knigavko V. G., Zaytseva O. V., Bondarenko M. A., Batyuk L. V., Rukin A. S.

М 42 Medical informatics : tutorial for foreign English-speaking students of medical universities / V. G. Knigavko, O. V. Zaytseva, M. A. Bondarenko et al. – Kharkov : KhNMU, 2019. – 60 p.

The tutorial was created in accordance with the updated working curriculum of the discipline "Medical Informatics". The structure of the tutorial suggests the possibility of its use in practical classes in this discipline. The manual is adapted for foreign students of medical universities studying in English.

Автори:

Кнігавко В. Г., Зайцева О. В., Бондаренко М. А., Батюк Л. В., Рукін О. С.

М 42 Медична інформатика : навч. посібник для іноземних англомовних студентів мед. ун-тів (англійською мовою) / В. Г. Кнігавко, О. В. Зайцева, М. А. Бондаренко та ін. – Харків : ХНМУ, 2019. – 60 с.

Навчальний посібник створено відповідно до оновленої робочої навчальної програми дисципліни «Медична інформатика». Структура навчального посібника дає можливість його використання на практичних заняттях з даної дисципліни.

УДК 61:004(075.8)

TOPIC 1. Basic concepts of Medical Informatics

The discipline "Medical Informatics" is taught to familiarize students with the use of information and communication technologies (ICT) in the field of healthcare, medical and biological data processing through the ICT, as well as to provide the information competence in future doctors.

The subject of "Medical Informatics" discipline is the information processes which involve the ICT application in the field of healthcare.

1.1. Basic concepts of Medical Informatics

Medical informatics (MI) is a scientific discipline that studies the processes of receiving, transmitting, processing, storage, distribution, presentation of information using information technique and technology in medicine and healthcare.

The main objective of MI is to optimize information processes in medicine through the use of computer technology, which ensures an increase in the quality of public healthcare.

Information is a collection of knowledge (new, previously unknown information) obtained during data processing.

Such concepts as “data”, “message”, “signal”, “communication channel” are associated with the concept of information.

Data are information presented in a formalized form and intended for processing by technical means, such as computers.

A message is an ordered collection of signals that can carry information.

A signal is any process that affects sensor systems.

A communication channel is a medium through which signals are transmitted. For example, in oral conversation, the signal is speech, and the communication channel is air, in the nervous system the signal is nerve impulses, and the channels are nerve fibers.

Signals can be **discrete** and **continuous**. An example of a discrete signal is the transmission of numbers by current pulses, an example of a continuous signal is a change in voltage in a circuit corresponding to a change in temperature.

Every message consists of a combination of a small number of simple signals of a certain physical nature. A complete set of such signals is called the **alphabet**; one signal from this set is called a **letter** of the alphabet.

The description of a message using a certain alphabet is called **encoding**, the translation of this message into another alphabet is called **recoding**, and the decryption of a message is called **decoding**.

Any information can be encoded using a *two-character alphabet (0; 1).* This code is called a **binary code**. This is the most common coding method in modern information systems. There are other coding methods, including those created by nature. For example, coding of information in the DNA of a cell is realized using 4 different "letters" of the alphabet (A, G, T, C) – nitrogenous bases (adenine, guanine, thymine, cytosine), and in the primary structure of proteins – using 20 amino acids.

1.2. Amount of information

One of the basic concepts of information theory is the **amount of information (*I*)**. The more different possibilities an event has, the more information about it carries a

message. The amount of information about event *A* changes in relation reciprocal to the **probability** $P(A)$ of this event, i.e. the greater the probability of the event, the less information there is in the message that this event occurred, and vice versa.

The dependence of the amount of information on the probability of an event is described by the **Hartley formula** (derived by the American scientist Ralph Hartley in 1928):

$$I(A) = -\log_2 P(A).$$

The amount of information is a positive quantity. Since the probability of any event $P(A) \le 1$, the value $\log_2 P(A)$ is always negative, therefore, the Hartley formula has a minus sign.

If the probability of the event $P(A) = 1$, then such an event must necessarily occur, and the message about this event does not carry information (the amount of information is zero): $I(A) = 0$.

If the probability is very small, i.e. $P(A) \to 0$, then the amount of information tends to infinity: $I(A) \to \infty$.

The unit of measurement of the amount of information is **bit**. **1 bit** is the amount of information contained in the message that an event whose probability is equal to $\frac{1}{2}$ has occurred:

$$I(A) = -\log_2 \frac{1}{2} = -(\log_2 1 - \log_2 2) = -(0-1) = 1 \text{ (bit).}$$

In other words, **1 bit** is the information contained in the message that one of two equally possible events has occurred.

Here is an example of calculating the amount of information contained in a message about event. Let us find the amount of information in the message that when throwing a dice, the number 5 fell out. The probability of this event is $P(5) = \frac{1}{6}$. Therefore,

$$I(5) = -\log_2 \frac{1}{6} = -(\log_2 1 - \log_2 6) = \log_2 6 = \frac{\ln 6}{\ln 2} \approx 2.58 \text{ (bit).}$$

Derived units of the amount of information: **1 byte = 8 bits**, 1 kbyte = 2^{10} byte = 1024 bytes, 1 Mbyte = 1024 kbyte, 1 Gbyte = 1024 Mbyte, etc.

If the events are not equally possible, it is important to know the **average amount of information** ($\overline{I}$) per message. This value is the mathematical expectation of the value $I(A)$ and is calculated by the **Shannon formula** (the formula was obtained by the American mathematician Claude Shannon in 1948):

$$\overline{I} = -\sum_{i=1}^{n} P_i \log_2 P_i,$$

where *n* is the number of all possible events that are reported, P_i is the probability of the *i*-th event (*i* = 1, 2, ..., *n*).

The average amount of information is also called **information entropy (*H*)** and is a measure of the uncertainty of information in a message.

Example. There are 2 white and 5 black balls in the basket. Find the average amount of information in a message about the color of a randomly drawn ball.

Probability to take out a white ball at random is $P(A)=\frac{2}{7}$, for a black ball it is $P(B)=\frac{5}{7}$. Using the Shannon formula, we obtain

$$\bar{I}=-\sum_{i=1}^{n}P_i\log_2 P_i=-\left(\frac{2}{7}\log_2\frac{2}{7}+\frac{5}{7}\log_2\frac{5}{7}\right)=0.863 \text{ (bit).}$$

1.3. Information Technology. Computer

Information technology is a computerized way of processing, storing, transmitting and using information.

A set of devices designed for automatic or automated data processing is called **computer technology**, a specific set of interacting devices and programs is called a **computing system**.

The central device of most computing systems is a **computer**. Typically, a computer contains the following devices:

- **an arithmetic-logical device** that performs arithmetic and logical operations;
- **a control device** that organizes the program execution process;
- **storage devices** for storing programs and data;
- **external (peripheral) devices** for **input-output** of information.

In modern computers, the arithmetic-logical device and the control device are combined into one device – **the central processor**.

Storage devices (**memory**) of a computer are divided into several types:

1. **Random access memory** (**RAM**). It stores data and programs that the computer is currently working with. It has a small volume and high speed. When you turn off the computer, the information in RAM is *erased*.
2. **Long-term (external) memory**. It has a low speed, but a large capacity, and the information stored in it is *not erased* when you turn off the computer. This type of memory includes media on magnetic disks (for example, hard disks), compact disks (CD, DVD), flash drives, etc.
3. **Cache memory** is a special ultra-fast random-access memory that stores copies of the most commonly used areas of RAM. In modern computers, it is usually a part of the central processor.
4. **Read-only memory** (**ROM**) contains data recorded in the computer during its manufacture, and these data cannot be changed.
5. **CMOS-memory** (by the name of the structure of CMOS microcircuits, complementary metal-oxide-semiconductor) is designed to store computer configuration parameters. Usually uses independent power to store information, i.e. a battery.
6. Other types of memory (for example, video memory).

Peripheral devices (input-output devices) are intended for input and output of information.

The main **input devices** in the computer are the **keyboard** and the **mouse. Input** devices are also a **scanner, video camera,** etc. **Analog-to-digital converters** (**ADC**s) for inputting information in the form of electrical signals, for example, for recording sound using a microphone or for receiving information from sensors, belong to input devices.

The main **output devices** are a **monitor** and a **printer**. The computer is also able to output information in the form of electrical signals to any actuators through **digital-to-analog converters** (**DAC**s). Sound output is also carried out using a DAC.

Input-output devices also include **network devices** (network adapters, modems, Wi-Fi wireless devices, etc.). They are used for information exchange between computers through various communication channels.

1.4. Software. File systems

Software is a set of programs that ensure the normal operation of a computer, it is used by programmers and computer users for solving various problems.

The software is divided into three types.

1. **System software**. This type includes operating systems, shells and auxiliary programs (utilities).
2. **General-purpose application software**. This type includes software of wide application: programming systems, database management systems (DBMS), text editors and publishing systems, universal office suites, etc.
3. **Special-purpose application software**. This group includes specialized programs: accounting programs, specialized databases, financial analysis programs, automatic design systems, statistical programs, expert systems, and many others.

Information in modern computers is usually stored as files.

A **file** is a unit of storage of data of arbitrary size with a unique name. Files provide a way to save information to disk and read it again. At the same time, such details as the method and place of information storage are hidden from the user.

File systems are designed to store data on disks and provide access to them. The file system determines how data are organized on a disk or other medium.

When working with files, **structuring mechanisms** are usually introduced. As a rule, they have **hierarchical relationships**. To organize such relationships, the concept of a **directory** is introduced, which is now often referred to as a **folder**. The directory contains information about data organized as files. The directory may contain other directories called **subdirectories**. With their help, we get the opportunity to build an almost unlimited hierarchy.

TOPIC 2. Healthcare information resources

2.1. Internet information resources and access

Information resources in the broad sense of the term are sets of data organized to effectively obtain reliable information; they are *arrays of documents* in information systems: libraries, archives, banks and databases, and other types of information systems.

The most common way to obtain access to information resources today is **Internet services**. Internet information services provide users with the ability to access certain information resources stored in Internet. Such resources are either files of standard formats, or various types of documents that can be viewed, printed, saved.

The *main information services* include the **file transfer service** provided with file transfer protocol (FTP) and the **World Wide Web** (WWW) provided with two protocols 1) HTTP – hypertext transfer protocol, and 2) HTTPS – hypertext transfer protocol secure – for secure connections. The World Wide Web accounts for more than 80 % of the world's networked information reserves. WWW has such a leadership due to the form of information presentation adapted for the Internet – hypertext.

Hypertext, unlike regular text, contains a system of links through which you can conveniently and intuitive structure very large volumes of information. The information base of the WWW service is a network of documents (**Web pages**) stored on Web servers of the Internet and interconnected by hyperlinks. A collection of interconnected pages belonging to one person or organization is called a **Web site**.

A web server is a program or service running on a computer whose task is to provide access to the data that are hosted on it using the HTTP and HTTPS protocols. A web server is both software of this kind and the computer on which it is installed. There are millions of Web servers in the world, and each of them can have one or more sites. Most of the world's web servers are functioning under Linux and UNIX operating systems, and the most popular web server today is Apache. Sites and pages are in a specific folder on a computer with a running Web server, and when we type the address of a Web page, we simply open the files in this folder.

2.2. Internet information search technologies

The problem of providing a convenient information search for Web users is solved in two ways: the first is the creation of Internet directories, the second is the creation of search engines (or search systems).

Internet resource catalogs (directories) are constantly updated and hierarchically replenishing catalogs of links to various information resources, containing a lot of individual Web pages, Web sites or their categories (groups) with a brief description of their contents.

Directories are easy to use: the way to search for information in a directory involves “moving down the stairs,” that is, moving from more general categories to more specific ones. On the main page of the Internet resources catalog site there is a topic list (*Fig. 1a-b*). By clicking on the topic name, you get to the list of subsections related to the selected topic, and then you have the opportunity to view the list of sites and get acquainted with their contents. This method of information searching is the most convenient when you find it difficult to clearly formulate the purpose of your search, or want to get a general idea of the topic.

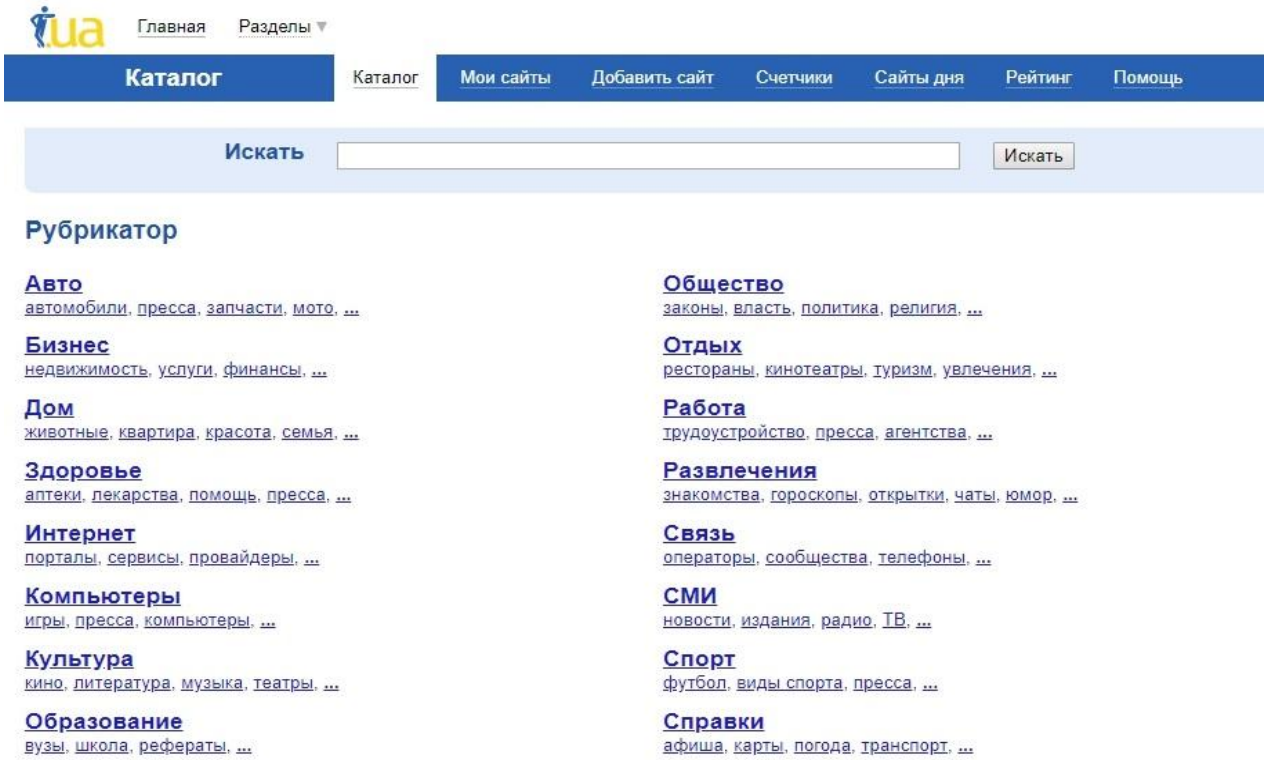

Fig. 1a. Internet resource catalog i.ua (catalog.i.ua) interface

Fig. 1b. Internet resource catalog www.ukraine-today.net interface

One of the advantages of thematic directories is that the *creators of the site* included in the catalog *themselves* provide explanations for the links and fully reflect its contents, which allows you to determine more accurately, how much the site's content matches the purpose of your search. However, there is a significant drawback: links to whole servers, sites or large sections of sites are entered into directories, so links to all pages, where the information you need could be found, do not fall into the directory. Directories are not suitable for a detailed, subtle search.

Examples of some thematic Ukrainian and Russian-language general-purpose catalogs: catalog.i.ua, favorites.com.ua, www.ukraine-today.net etc.

Each of the general-purpose Internet resource directories contains a section on medicine and health. For example, i.ua – http://catalog.i.ua/catalog/7/.

Search engines (search systems, or **Web search engines)** are servers with a huge database of Internet addresses that automatically access Web pages at all these addresses, examine the contents of these pages, form and register keywords from the pages in their database data (index pages). Moreover, search engine robots follow the links found on the pages and reindex them. Since almost any Web page has many links to other pages, with this kind of work, the search engine can theoretically go around all sites on the Internet. Therefore, search engines are the most suitable tools for a detailed, subtle search for information by keywords.

The most popular search system today is **Google** (www.google.com). Equally famous are Yahoo (www.yahoo.com), Bing (www.bing.com).

The technology of using a search system is simple. The user types a key phrase and activates the search, thereby receiving a selection of documents according to the formulated request. This list of documents is ranked by the search system according to certain criteria so that at the top of the list are those documents that most closely match the user's request. Each of the search systems uses different criteria for ranking documents, both in the analysis of search results and in the formation of an index database of Web pages. Therefore, if you specify the same request in the search bar of different search systems, you can get different search results.

The main disadvantage of search systems is the relatively low percentage of information correspondence in the list of search results. This means that not every address in the list of search results meets the user's request. Sometimes a user's request is formulated too broadly, and the search result is immensely large (tens of thousands of pages), and then the search system incorrectly interprets keywords.

A lot of the sites mentioned above combine the functions of a search engine and a directory, also providing access to other Internet resources: e-mail, instant messaging in social networks, access and management of remote (cloud) data storages, etc. (*fig. 2*). These sites are called **Web portals**.

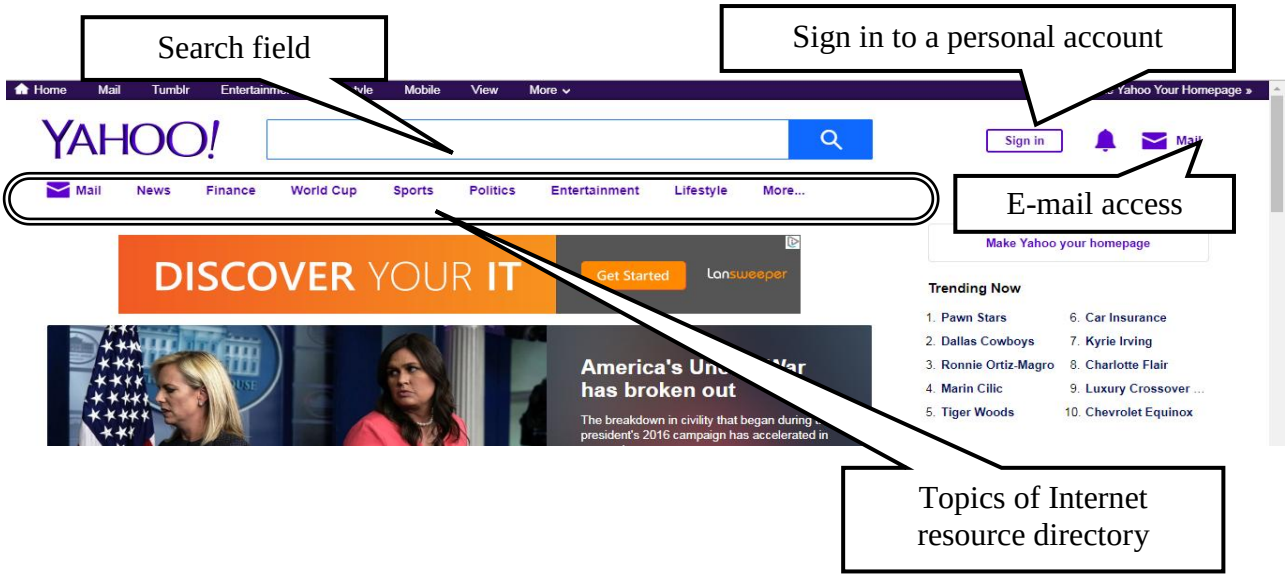

Fig. 2. Yahoo Web portal interface (www.yahoo.com)

2.3. Medical Internet resources

Health-related issues are reflected in many sites. On the Internet you can find materials of interest to patients, practitioners, healthcare providers, researchers, etc. There are special resources that are interesting for each individual group, but there are those that are necessary for a wide range of users. Two main areas of classification of medical resources of the Internet can be defined: by type of visitors and by purpose of visit.

The following resource groups can be distinguished by **type of expected visitors**:

- for patients (resources offering reference medical information about various diseases, their symptoms, methods of prevention; about doctors and institutions providing appropriate medical care);
- for doctors (specialized medical information for practitioners and researchers);
- for specialists in the organization of health care (laws and regulations, reference materials necessary for organizing work and preparing reports; outsourcing resources, i.e. medical services provided by other institutions and commercial enterprises under contracts);
- for specialists in financial and economic services and entrepreneurs whose activities are related to healthcare (equipment, medicines, supplies, tools, communications, transportation, etc.);
- for specialists in personnel services and job search (services that allow to view resumes of medical specialists, search for vacancies, etc.).

For the purpose of visiting the Internet resource, the following resource groups are allocated:

- to search for specialized information;
- to search for medical services;
- to search for therapeutic and prophylactic agents;
- for training;
- for business and provision of medical institutions;
- for job and staff search;
- for communication.

Examples of specialized medical sites:

- http://www.moz.gov.ua – the official website of the Ministry of Health of Ukraine;
- http://www.medicusamicus.com – "Medicus Amicus" – a site for doctors and pharmacists;
- http://www.morion.ua – the site of the MORION company, which specializes in the development and implementation of computer technologies in medicine and pharmacy, publishing business, research and analysis of the pharmaceutical market of Ukraine;
- http://www.health-ua.org – "Health of Ukraine" – medical portal;
- http://www.compendium.com.ua – On-line compendium;
- http://www.medicinform.net – "Medical Information Network" – search engine and catalog, news.

2.4. MedLine medical publishing database

Any scientific research starts with a search for known results obtained earlier by other researchers in this field of knowledge. Such a search allows to analyze and determine the novelty of the study, its relevance, evaluate the significance of the results, choose the relevant direction for development. Since most scientific results are published in specialized journals, a high-quality search involves viewing all periodicals on a selected topic. This problem can be solved using various bibliographic systems and databases containing information on scientific articles published in the world and monographs. One of the most famous bibliographic systems for medical publications is MedLine. This database was created at the National Medical Library of the US National Institute of Health in the early 1980s. It contains abstracts and bibliographic data of all publications from the late 1960s to the present day from more than 4000 world scientific journals. The benefits of the MedLine database include the following:

- ability to quickly select bibliographic data on articles on a given topic;
- abstracts of articles provide a general idea of the content, materials and results of publication;
- for approx. 80% of the journal publications in the bibliographic catalog of the library, full-text versions are available;
- ability to search for related topics.

The MedLine database and the PubMed utility serving it are freely available on the server http://www.ncbi.nlm.nih.gov/pubmed or http://pubmed.gov.

Search for publications is performed by keywords and phrases. There is an input field for keywords on the search page. However, keywords and phrases are only entered in English. The request is processed after clicking on the Search button (*Fig. 3*). The search result is a list of publications that contain the given keywords and phrases.

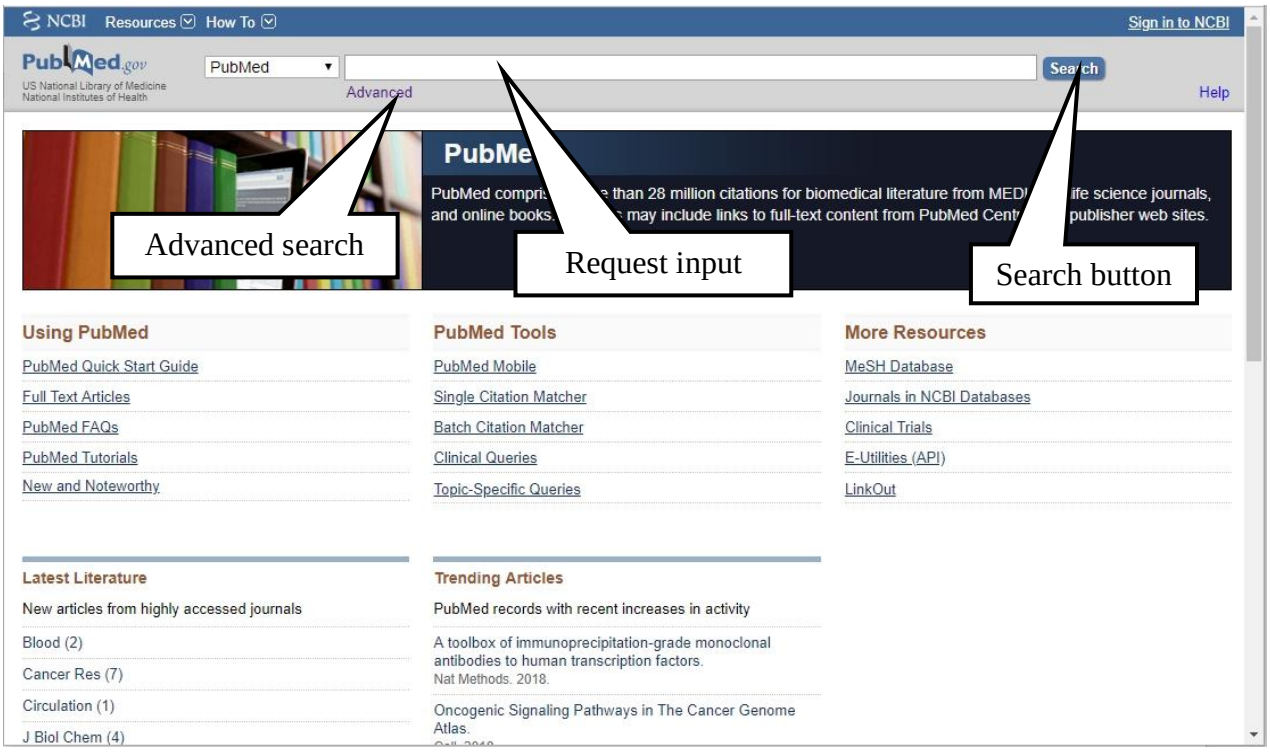

Fig. 3. PubMed homepage interface

TOPIC 3. Creation and maintenance of medical records

3.1. Classification of programs for processing texts

To create documentation, including medical one, we currently use various applications for working with texts. These applications are classified into text editors, word processors and publishing systems.

Text editors are designed to create and modify text data in general and text files in particular. They allow to view the contents of text files and perform various actions on them: inserting, deleting and copying text, context searching and replacing, sorting strings, viewing character codes and encoding conversion, printing, etc.

For serious document processing, **word processors** are used. Word processors differ from text editors by additional abilities for *text formatting*, i.e., designing (using different methods of aligning text, several fonts, etc.). Modern word processors, in addition to formatting fonts and paragraphs and spellchecking, include features previously present only in desktop publishing systems, e. g. creating tables and inserting graphic objects. Currently, most text editors are word processors. The most popular word processors are Microsoft Word, LibreOffice Writer, WordPerfect.

A **publishing system** is a set of equipment for preparing the original layout of the publication for transfer to the printing house. A publishing system includes one or more personal computers with the necessary software for creating a layout, recognition, typing and typesetting, image editing, preprint preparation of the original layout.

3.2. Word processor features

In the process of preparing a document, the user has at his disposal a set of tools and procedures for entering, editing and formatting text and embedded objects.

The main functionality for working with a document include:

- using fonts of various sizes and styles, various ways of highlighting;
- setting paragraph parameters;
- setting line spacings;
- spellchecking, grammar and synonyms;
- search and replacement of characters, words and text fragments;
- automatic hyphenation of words;
- automatic pagination;
- printing the top and bottom page headers (headers and footers);
- creating footnotes;
- creating tables of contents, indexes;
- typing in several columns;
- creating tables, figures and plotting charts;
- viewing documents before printing;
- setting paper sizes and print options;
- undoing and repeating previous user actions;
- inserting fields with standard information (date, time, copyright data, etc.);
- creating macros and hypertext links;
- embedding various objects into the document (files, formulas, etc.);
- import of documents created in other applications, etc.

Word processors also offer a wide range of tools for making the document more attractive: autoformatting, applying styles, libraries of styles and document templates. Using templates, you can automate the preparation of standard documents such as fax messages, standard business correspondence, and documentation.

Let us dwell on some of these features.

1. **Font selection**. The word processor allows to select any of the fonts available on the computer, set its size (the font size is usually specified in points, 1 point is 1/72 inch = 0.3528 mm), apply various highlighting methods (bold, italics, underline, strike-through, etc.), indicate the font and background colors. The width of characters (without changing the height) and the spacing between characters can also be specified.

2. **Paragraph formatting**. Paragraph formatting includes specifying the alignment method (left or right, center, justified); indentation from the left and right page borders, indentation of the first line; setting line spacing and paragraph spacing; specifying other properties of the paragraph (for example, prohibition of "hanging lines", i.e. tearing off the first or last line when moving to the next page).

3. **Styles**. A style is a set of formatting options that are applied to the text of a document to change its appearance quickly. Styles allow to apply simultaneously the entire set of formatting attributes (usually font and paragraph formatting) to the text. Word processors contain a set of built-in styles for creating plain text, headings, lists, etc., and also allow to modify existing styles and create new ones. Applying styles allows achieving uniformity in formatting different parts of one document or different documents of the same type, and also makes it possible, if necessary, to quickly reformat the entire text of a document or its individual elements.

4. **Creating tables**. Word processors allow to insert tables into a document. Creating tables starts by specifying the number of rows and columns in the table. The table consists of cells into which one can enter text and insert objects (symbols, formulas).

Next, the user can perform various actions with the table: merge and split cells, rows, columns; format text in the cells, set text alignment (horizontal and vertical); resize cells (height and width), including adjusting sizes automatically according to the content; set the appearance of cell borders, table fragments or the entire table (line type, thickness, color, visibility), the cell background and the text color.

5. **Creating bookmarks and hyperlinks** in the document. Word processors allow you to create bookmarks in a document, that is, indicate places in the document that you can quickly jump to using links (like on a web page). A link (hyperlink) is a piece of text that is associated with a previously set bookmark. The hyperlink is automatically highlighted in a special style (usually marked with a different color and underlined). Clicking the link sends the cursor to the place where the bookmark was created. Hyperlinks are useful for creating tables of contents and pointers to specific places in a document.

6. **Creating graphic objects**. Graphic objects include lines (straight lines, curves, broken lines, arrows) and shapes (rectangle, circle, ellipse, triangle, flowchart elements, callouts, etc.). For each object, its dimensions (height, width), position on the page (relative to the page margins, relative to the paragraph, directly in the text),

rotation angle, colors, borders, background patterns, transparency, additional properties (for example, arrow types) are specified. The finished shape can be copied to another place in the document. It is also possible to align graphic objects relative to each other, relative to the page or to the grid, group several shapes into one object for further copying, place text labels inside the shapes.

In addition to the built-in graphic objects, word processors are able to insert other types of objects into the document, i.e. pictures, diagrams, formulas, etc.

7. **Creating a document template**. A template is a model for creating a new document. The template stores the elements that form the basis of the document:

- permanent text, graphic objects along with formatting;
- fields for entering information (text fields, lists, date fields, etc.);
- parameters of the printed page of the document;
- the list of available styles;
- macro commands (sequences of actions that automate the work with a document).

When loading a template, a new document is automatically created containing a copy of all the information present in the template (text, graphics, tables, etc.) with formatting, and it is also proposed to enter data into the fields using dialog boxes. Templates are convenient to create forms of standard documents.

TOPIC 4. Databases. Database management systems

4.1. Basic concepts

Modern medicine is impossible without using databases. For example, a local database may be a database of patient records, which is associated with the concept of "registry". All medical histories, test results, ECG, radiographs and other information that may be available to the doctor at any time without outside participation can be stored here. The electronic way of keeping records of patients allows to quickly solve the problems of transmitting information to another medical institution (due to moving or referring a patient for treatment), and provide protection against unauthorized access.

A **database** (**DB**) is a named set of structured data related to a specific subject area.

A **database management system** (**DBMS**) is a set of software and language tools needed to create databases, keep them up to date and organize the search for the necessary information in them.

Data structuring is introduction of conventions on how data are presented. Creating a database, the user seeks to *organize* information about various features of objects and quickly get a sample of data with an arbitrary combination of features. This can only be done if the data are *structured*.

4.2. Database classification

There are various ways to classify a database, in particular, by data processing technology, by the method of accessing data, and by the degree of versatility.

1. According to **data processing technology**, databases are divided into centralized and distributed.

A **centralized** database is stored in the memory of one computing system. If this computing system is a component of a computer network, distributed access to such a database is possible. This method of using databases is often used in local networks.

A **distributed** database consists of several, possibly overlapping or even duplicated parts, stored on different computers. Operation of such a database is carried out using a distributed database management system.

2. By the **method of accessing data**, databases are divided into standalone databases with local access and centralized databases with remote (network) access.

Standalone local databases are the simplest ones. They store their data in the local file system on a computer where they are installed. The DBMS accessing them is located on the same computer.

Systems of **centralized databases with remote (network) access** involve various architectures of such systems – file server and client-server.

File server. The architecture of database systems with remote (network) access involves allocation of one of the network machines as a central file server. A shared centralized database is stored on such a machine. Database files in accordance with user requests are transferred to workstations, where processing is mainly performed.

Client-server. This concept implies that, in addition to storing a centralized database, the central machine (**database server**) must provide a principal part of data processing volume. A request for data issued by a client (workstation) generates a

search and retrieval of data on the server. The extracted data, *but not the files*, are transported over the network from the server to the client.

A specific feature of the client-server architecture is the use of the **SQL query language** (Structured Query Language). This is a universal language designed for creating and executing queries, processing data both in the application's own database and in databases created by other applications that support SQL. An SQL query consists of one or more statements, one after the other, separated by a semicolon.

3. By the **degree of versatility**, two classes of DBMS are distinguished – general-purpose systems and specialized systems.

4.3. Database information units

The unit of information stored in the database is a **table**.

Each table is a collection of **rows** and **columns**, where the **rows** correspond to the **instance of the object** (a specific event or phenomenon), and the **columns** correspond to the **attributes** (features, characteristics, parameters) of the object, event, phenomenon. In terms of a DB, table columns are called **fields**, and its rows are called **records**.

DBMS processing objects are the following database information units.

A **field** is an elementary unit of logical data organization, which corresponds to an indivisible unit of information – property.

A **record** is a collection of logically related fields.

A **table** is an ordered structure consisting of a finite set of records of the same type.

The **primary key** is a field (or group of fields) that allows to identify each row in the table uniquely. The primary key must have two properties:

- unique identification of the record: the record must be uniquely determined by the value of the key;
- no redundancy: no field can be removed from the key without violating the unique identification properties.

4.4. Database organization models

There are certain relationships between fields and records. Depending on the nature of these relationships, there are three types of database organization models: hierarchical, network, and relational.

1. The **hierarchical model** of a database is a collection of elements arranged in the order of their subordination from the general to the particular and forming an upside-down tree (graph). A node of this graph is a collection of data attributes describing an object. Each node at a lower level is linked to only one node located at a higher level.
2. The **network model**: with the same basic concepts (level, node, connection), each element can be linked to any other element.
3. The **relational** database model is currently the most common. This database model is based on the concept of relation. Relations are presented in the form of two-dimensional tables.

The relation (table) is represented in the computer as a data file. A row of the table corresponds to a record in the data file, and a column corresponds to a field. In relational database theory, rows are called tuples, and columns are called attributes. A

list of relation attribute names is called a relation schema. In each relation, there is one special attribute which is called a key attribute or simply a key. The key attribute must be unique, i.e. it must uniquely identify the tuples.

Various operations can be performed on relations (tables), similar to performing arithmetic operations (for example, joining tables). This makes it possible to obtain other relations from the relations stored on the computer.

The relational database model is used to create electronic medical records.

4.5. DB development stages

1. **Problem statement**. At this stage, the task of creating a database is formed. It describes in detail the composition of the database, its functions and purpose of its creation, and also lists what types of actions are supposed to be carried out in this database.
2. **Analysis of the object**. At this stage, it is considered what objects the database can consist of, what are the properties of these objects. All this information can be stored in the form of separate records and tables.
3. **Choosing the database model**. At this stage, it is necessary to choose the type of database organization model. Next, one needs to compose a diagram indicating the relationships between tables or nodes.
4. **Choice of methods for presenting information** and software tools.
5. **Creating a database**. In the process of creating the database, we can distinguish some stages typical for any DBMS:
   - DBMS launch, creation of a new database file;
   - creation of the source table or tables;
   - creation of screen forms, i.e., a graphical interface for entering and displaying data;
   - filling in the database.
6. **Working with the created database**. The work with the database includes the following actions:
   - data search;
   - data sorting;
   - data selection;
   - printing;
   - updating and adding data.

All these actions are usually performed by submitting queries to the DBMS. A **query** is an instruction for selecting the necessary information from a database.

Queries in modern DBMS, as a rule, are formed in two ways: using a graphical interface (the so-called query designer) or in SQL.

Main types of queries:

- queries for a **selection** that return data from one or more tables and display them in the form of a table;
- queries with **parameters**, i.e. queries that display a dialog box for entering data;
- queries for **modifying records**, in particular, **adding**, **deleting** or **updating** records;
- queries for **creating** or **deleting tables**.

TOPIC 5. Medical information systems. Electronic medical records

5.1. Basic concepts

An **information system** (IS) is an organized set of documents (arrays of documents) and information technologies (including using computer technology and communications) that implement information processes.

Information systems used in medicine and healthcare are called **medical information systems** (MIS).

There are various approaches to classifying MIS. The healthcare system is a multi-level structure and is built on a hierarchical basis. In accordance with this principle, medical information systems are divided into:

- **basic** level MIS;
- **institution** level MIS;
- **territorial** level MIS.

5.2. Basic level MIS

The goal of a basic level MIS is computer support for the work of a *medical specialist* (clinician, hygienist, laboratory assistant, etc.).

Basic level MIS groups:

- medical reference systems;
- medical consultative and diagnostic systems;
- medical hardware and software systems;
- doctor's automated workplace (AWP).

Medical information and reference systems are designed to search and provide medical information at the user's request. Information arrays of such systems (databases and data banks) contain medical *reference information* of various nature. This includes scientific information on various medical disciplines, reference statistical and technological information of a wide profile, accounting and documentary information.

Systems of this type *do not process information*, but provide quick access to the required data.

Medical consultative and diagnostic systems (MCDS) are intended for *diagnosing pathological conditions* (including prognosis and development of recommendations on treatment methods) for diseases of various profiles and for different categories of patients. The input information for such systems is the data on the symptoms of the disease, which are entered into the computer in an interactive mode or in the format of information cards.

In general, MCDS contains:

- a database (DB);
- a knowledge base (KB);
- a logical inference mechanism (machine) (LIM);
- user interface.

The **database** is designed to store the set of facts, specific data about objects in the field of MCDS.

The **knowledge base** contains knowledge related to a specific application area, including individual facts, rules, and, possibly, heuristics related to solving problems in this application area.

The **logical inference mechanism**, using the rules and methods of the KB, converts specific information about the object to the form corresponding to the purpose of the MCD (diagnosis, plan of action, etc.).

The **user interface** provides a continuous exchange of information between the user and the system; it also gives the user the opportunity to observe the problem-solving process taking place in the LIM.

According to the method of realizing the LIM, **probabilistic** MCDS and **expert** MCDS are distinguished, that is, the methods of probability theory or the methods of artificial intelligence can be the basis of LIM.

Medical hardware-software complexes (MHSC) are intended for information support and/or automation of the diagnostic and medical process, carried out in direct contact with the patient or the object of study.

According to their purpose, MHSC can be divided into a number of classes:

- systems for functional and morphological studies;
- monitor systems;
- medical process control systems;
- laboratory diagnostic systems;
- systems for biomedical research.

Automated workplace (AWP) of a doctor is a computer information system designed to automate the entire technological process of a doctor of corresponding specialty and provides information support when making diagnostic and tactical (medical, organizational, etc.) medical decisions. All the clinical-level information systems discussed above can and should be included in the workplace structure, providing automation of the entire medical process of a physician.

5.3. MIS of medical institutions

Information systems of this level are designed to work with **information flows of medical facilities**. They are represented by the following groups:

- IS of consultative (advisory) centers;
- information banks of medical institutions and services;
- personalized registers;
- screening systems;
- healthcare facilities information systems;
- IS for research institutes and universities.

IS of consultative centers are designed to ensure functioning of relevant units and information support for doctors in consulting, diagnosing and decision-making in emergency situations.

Information banks of medical institutions and services contain summary data on the qualitative and quantitative state of employees of the institution, the attached population; basic statistical information, characteristics of service areas and other necessary information.

Personalized registers (databases and data banks) are a type of information and reference systems containing information on attached or observed population based on a formalized medical history or electronic medical record.

Screening systems are intended for pre-medical prophylactic examination of the population, as well as for medical screening (formation of risk groups and identifying patients who need the help of a specialist).

Information systems of medical facilities are information systems based on the integration of all information flows into a single system and providing automation of various types of activities of the institution.

IS for research institutes and universities are designed to informatize the learning process, research and management activities of research institutes and universities.

5.4. Territorial level MIS

Territorial level MIS provide management of specialized and profiled medical services, clinical, hospital and ambulance services at the level of territories (city, region, state, etc.). Such MIS can also be global (worldwide).

At this level, MIS are represented by the following groups:

- **administrative and management MIS** for the administrations of territorial medical services;
- **statistical MIS** for processing territory-based summary information;
- **MIS for specialized services**: ambulance and emergency, drug provision, registers (TB, psychiatry, infectious diseases, etc.);
- **computer telecommunication networks** that create a single information space in the healthcare sector.

5.5. Electronic medical record

An electronic medical record (EMR, electronic medical history) is a medical record of a patient in a medical facility in electronic form.

An electronic medical record can be used both in outpatient and inpatient facilities, taking into account the nature and characteristics of medical care in them. EMR is compiled and stored in an automated information database of a medical institution. The record contains data on the patient's medical history, information on vaccinations and his desire to become a donor.

Electronic health records should gradually replace traditional health insurance cards. Germany was one of the first EU countries to introduce electronic medical records in October 2011.

The electronic medical record is a single information resource that allows to process personal data of patients, as well as exchange such data with other medical institutions for compiling, recording and storing medical information. Medical documents (information) with EMR can be transferred to competent organizations: insurance companies, institutions for monitoring the provision of medical care, law enforcement agencies, etc.

In Ukraine, the electronic medical card was introduced in test mode on March 1, 2019. It is available to family doctors, therapists and pediatricians of medical

institutions connected to the electronic health system and have a contract with the National Health Service of Ukraine.

Electronic medical record is an information system that relies on *relational database* technology. This method of storing information allows conveniently, in automatic mode, selecting data by a certain characteristic or set of characteristics, ordering their display by various columns of the table, for example, by date, patient name, diagnosis.

The electronic form of a medical record facilitates the solution of many problems:

- documentation (accumulation, reliable storage, ability to conveniently view) of arbitrary medical information about the patient with reference to the calendar date;
- search (filtering by a set of attributes) of the necessary information;
- tracking the time dependence of individual diagnostic parameters;
- performance studies.

Such an information system assumes the availability of convenient *means of inputting and displaying information*. To enter text and digital information, *forms* are displayed on a computer screen. If it is possible to formalize the data (forming a fixed or supplemented list of values of the input parameter), the choice of value is made from the provided list. All sorts of electronic directories can be added to the system (classifiers of diagnoses, drugs; address books, etc.). Tools for importing hardware survey data are provided. Many modern medical devices (tomographs, fluorographs, ultrasound scanners, cardiographs, etc.) generate the result directly in electronic form. International standards, for example, DICOM (Digital Imaging and Communications in Medicine), exist for instrumental diagnostic information. To protect data from accidental loss as a result of a hardware accident or deliberate hacking of data storage facilities, such information systems are located on well-protected servers that have mechanisms of data backup and status logging.

TOPIC 6. Processing medical information using spreadsheet processors

6.1. Basic concepts

When processing various types of information (including medical), **tables** and **diagrams** are convenient visual forms of data presentation. Tables and diagrams allow to store and analyze information, make decisions in accordance with the results of data processing.

Software systems used to process data in the form of tables are called **table processors**, or **spreadsheet processors**. This class of applications includes Microsoft Excel, LibreOffice Calc, Quattro Pro and many others. This section will cover the LibreOffice Calc table processor.

When the spreadsheet processor loads (or a new file is created in it), a table (**spreadsheet, worksheet**) appears on the screen. A single file can contain several such tables, usually called **sheets**.

As in DBMS, in spreadsheet processors any table consists of **rows** and **columns**. In most spreadsheet processors, including Calc, columns are denoted by one, two or three letters (A, B, ..., Z, AA, AB, ... ZZ, AAA, ...), and rows are indicated by numbers.

Columns / Rows	A	B	C	D	E	F	G
1							
2							
3							
4							
5							

6.2. Cell and its characteristics

Data in tables are stored in **cells**. Each cell is characterized by its address, format and status.

The cell **address** is usually written as a combination of the column name and row number, for example, A1, B20, AX157.

The cell **format** determines the display of the cell contents in the table. It includes formatting options for numerical values (as well as date and time), font, alignment, border appearance, and background color.

The cell **status** allows you to protect the cell from changes, hide the formula, hide the cell when displayed or when printing, etc.

Cell content. A cell in LibreOffice Calc can contain three types of data: text, numeric, and formulas.

Text data can be used for headings, explanations, etc., and are strings of text of arbitrary length.

Numeric data can be represented as integers (for example, 1; –6), decimal numbers (for example, 10.6; –0.8) or exponential numbers (5.6E–4).

Formulas are instructions for calculations and consist of operands connected by symbols of arithmetic operations (for example, “+”, “–”, etc.). **Operands** can be

numbers (integer, decimal, exponential), cell addresses and functions (arithmetic and trigonometric, date and time, etc.).

When entering formulas, they must **begin with an equal sign** (=).

The concept of **cell value** is closely related to the concept of a formula. The cell value is the **result** of operations performed by the table processor based on the content. For example, if the content of a cell is the formula =5*2+1, then its value will be 11. The value of the cell is displayed in the cell itself, and its contents is displayed in the input line.

6.3. Relative and absolute cell addresses

Cell addresses in formulas (also called **references** or **links**) can be relative and absolute. The difference between relative and absolute references is significant when copying a formula from one cell to another.

By default, LibreOffice Calc uses **relative** references. When copying a formula with a relative reference, such a reference is automatically **adjusted** in accordance with the new location, i.e., the *relative location* of the cell with the formula and the cell that the reference points to remains unchanged. In this case, the reference contained in the copied or moved formula refers to a new cell. Relative references are convenient to apply a formula by autofilling to a large array of data, for example, to a row, column or rectangular range of cells.

When copying a formula with an **absolute** reference, LibreOffice Calc copies the absolute reference *in the same way* as it looks in the original formula. Absolute references are used, in particular, to insert constants stored in separate fixed cells or data from a specific row or column into formulas.

In LibreOffice Calc formulas, an absolute reference is created by adding the *dollar sign* **"$"** in front of the address. It can appear before the column name (letters), before the row number (digits), or in both places, that is, each reference element can be absolute or relative, regardless of the other:

Reference	Column A	Row 1
$A$1	Unchanged	Unchanged
A$1	Modified	Unchanged
$A1	Unchanged	Modified
A1	Modified	Modified

Example. Let cell B3 contain a formula with a reference to cell A1. Then this formula is copied or moved one cell to the right and down, i.e., to cell C4. In this case, if the formula contains relative references, then they will also shift: A → B, 1 → 2, and the absolute references will remain unchanged. The result of copying depending on the type of reference will be as follows:

Original formula in B3	Formula after copying into C4
=$A$1	=$A$1
=A$1	=B$1
=$A1	=$A2
=A1	=B2

References in spreadsheets can point not only to one cell, but also to a **range of cells**. The range of cells can be:

- a row or its part;
- a column or its part;
- multiple rows or columns;
- a rectangular area.

A range of cells is denoted by indicating *two addresses of the same type* (columns, rows or cells) separated by the **":"** symbol as follows:

- several rows (for example, from 1 to 5 inclusive): **1:5**;
- one row (for example, row 2): **2:2** (in this way, to distinguish from a numerical value!);
- a part of a row: **A1:D1**;
- several columns (from column B to column F inclusive): **B:F**;
- one column (for example, column D): **D:D**;
- a part of a column: **B2:B5**;
- a rectangular area (addresses of cells located in the upper left and lower right corner of the region are indicated): **B1:G5**.

References in the description of cell ranges, as well as for individual cells, can be both relative and absolute. In many cases, for example, when using the "Function Wizard", cell ranges can be specified by selecting them directly with the mouse in the table.

6.4. Functions in Table Processors

For calculations and other types of data processing, spreadsheet processors are equipped with a set of built-in **functions**. Functions are divided into categories: mathematical, statistical, financial, logical, text, for working with date and time, etc.

Each function has a unique **name** and one or more **arguments**, separated by the **";"** symbol. Function arguments can be:

- fixed values (numeric or text);
- cell references (relative or absolute);
- cell ranges;
- other functions (nested functions).

To insert functions into the formula, you can enter the function name and its arguments manually or use the special dialog "**Function Wizard**", which allows you to find a function by category, by description or by name, view the function description with a list of its arguments and enter the arguments in the corresponding fields. The figure shows an example of using the Function Wizard to enter a logical function IF.

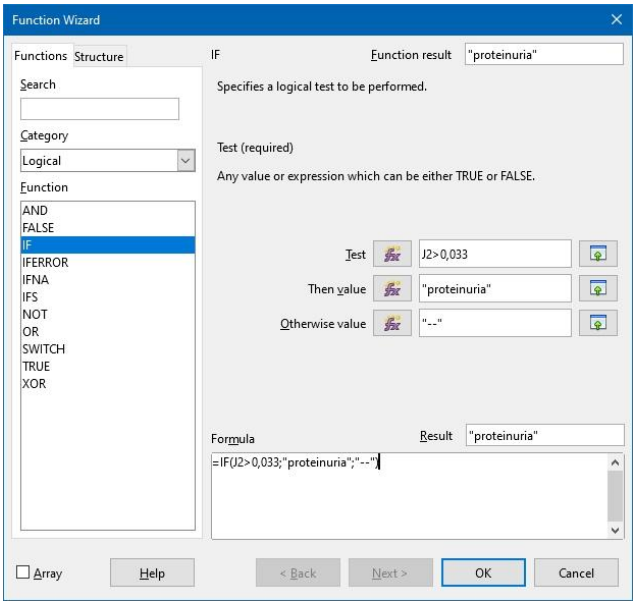

Fig. 4. Window "Function Wizard"

6.5. Plotting charts

Modern table processors allow you to display data from tables **graphically** in the form of **diagrams (charts)**. In LibreOffice Calc, there are various types of diagrams: bar and ribbon charts, circular chart, points, lines (graph), etc.

To create a chart, use the special tool "**Chart Wizard**". To build a chart, you must provide the following information:

- one or more **ranges** of cells with data (the requirements for ranges depend on the chart type);
- chart **type**;
- **location of data** in ranges (data series in columns or in rows), the presence or absence of **labels** in the table;
- quantity, order and properties of **data series** (if there are several);
- other elements of the diagram: headings, axis labels, grid appearance, legend.

Charting tools in LibreOffice Calc also allow you to process data directly on the chart, in particular, find average values, plot trend (regression) lines, indicate errors, etc.

TOPIC 7. Methods of biostatistics. Statistical analysis of biomedical data

7.1. Basic concepts

Mathematical statistics is a science that develops mathematical methods for organizing and using statistical data for scientific and practical applications. This goal is achieved by solving two main tasks.

1. **Indication of methods** for **collecting and grouping** statistical information obtained as a result of observations or as a result of specially designed experiments.

2. **Development of methods** for statistical **data analysis** depending on the objectives of the study.

Data analysis methods include:

- **estimation** of unknown probability of an event; estimation of an unknown distribution function; estimation of parameters of a known distribution; assessment of the dependence of a random variable on one or more random variables, etc.;
- **verification of statistical hypotheses** about the form of an unknown distribution or about the values of the distribution parameters if the form of this distribution is known.

A **random variable** is a quantity that takes one of many possible values as a result of an experiment (trial), and the appearance of one or another value of this quantity is a random event.

A **discrete** random variable is a random variable with a finite or countable set of possible values. As a rule, a discrete random variable describes the number of objects (events, phenomena), the serial number of an object in a list, etc.

A **continuous** random variable is a random variable that can take any of the values belonging to the interval (or intervals) in which it exists. A continuous random variable always has an infinite number of values. Most physical quantities (mass, length, concentration, etc.) are continuous variables.

The first step in statistical analysis is to **classify the data type**, i.e. to assign them to a particular **measurement scale**.

Measurement scales are classified:

1) by the type of random variables: **continuous** (temperature, hemoglobin in the blood) and **discrete** (outcome of the disease, blood type).

2) by the set of valid operations with values: nominal, ordinal, interval, absolute.

The **nominal scale (scale of names)** is used to group objects according to a qualitative criterion (for example, by color, gender, blood group). This scale makes it possible to compare objects by this criterion (i.e., only equivalence relations $x = y$ and $x \neq y$ are defined for it). The nominal scale does not imply ordering of values and quantitative relationships between values.

The **ordinal (rank) scale** is an ordered sequence of values and allows not only to establish the fact of equality or inequality of the measured objects, but also to determine the nature of the inequality in the form of judgments: “more-less”, “higher-lower”, “worse-better”, etc. n. For the ordinal scale, the relations $x = y$, $x \neq y$, $x < y$, $x > y$ are defined, but arithmetic operations (addition, subtraction, etc.) are not defined. Examples of ordinal scales: performance assessment (unsatisfactory,

satisfactory, good, excellent), the outcome of a patient's treatment (recovery, improvement, no improvement, worsening, death).

The **interval scale (scale of differences)** allows you to describe the properties of the object quantitatively by comparing the feature of the object with the standard. For interval scales, all the above comparison operations and arithmetic operations of addition and subtraction make sense. The reference point for such a scale is set arbitrarily. An example of an interval scale is the Celsius scale, on which the reference point is chosen according to the ice melting temperature.

The **absolute scale (scale of ratios)** is an interval scale in which there is an additional property - the natural and unambiguous presence of a zero point. For example, the number of people in a lecture room. This is the only one of the four scales that has an absolute zero. The zero point characterizes the absence of measurable quality. Using such scales, physical quantities can be measured – mass, length, strength, etc. An example of an absolute scale is the Kelvin scale, on which temperature is measured from absolute zero. The operations of multiplication and division make sense only for the absolute scale.

7.2. Distributions of random variables

To specify a random variable, it is necessary to indicate the **distribution law** of this quantity, which can be represented in the form of a table, formula or graph.

The distribution law of a discrete random variable X is given in the form of a table in which all possible values x_i of this quantity and the corresponding probabilities $P(x_i)$ of the appearance of these values are indicated.

The distribution law of a continuous random variable X is given by a **distribution function** $F(x) = P(-\infty < X \le x)$ that is equal to the probability that the random variable X will take a value less than or equal to x. The other way to specify the distribution of a continuous random variable is the **probability density** $f(x) = \frac{dF(x)}{dx}$.

The concept of a quantile is closely related to the concept of distribution function. A **quantile** (x_α) is a value which a given random variable does not exceed with a fixed probability. For continuous random variables, the α-quantile (α level quantile) of the distribution $F(x)$ is the solution of the equation $F(x_\alpha) = \alpha$. A quantile can be considered as a function inverse to the distribution function, i.e. an argument of the distribution function for which its value is equal to α. Quantiles of normal distribution are widely used in statistics to construct interval estimates of distribution characteristics.

Quantiles with levels that are multiples of 0.25 are called **quartiles**:

- quantile of level 0.25 – the lower (first) quartile,
- quantile of level 0.5 – **median** (the second quartile),
- quantile of level 0.75 – the upper (third) quartile.

The minimum and maximum values of a random variable are sometimes called the zero and fourth quartiles, respectively.

The law of distribution of a random variable is the most complete characteristic of this random variable, but often enough information is provided using the numerical characteristics of the random variable. Among them, the most commonly used are: mathematical expectation, variance, standard deviation.

The **mathematical expectation** $M(X)$ of a random variable X has a meaning of the average value of this quantity and is calculated by the formulas

$$M(X) = \sum_{i=1}^{N} x_i P(x_i); \qquad M(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

for discrete and continuous random variables, respectively.

The **variance** $D(X)$ characterizes the magnitude of *scatter* of values of the random variable X around its mathematical expectation $M(X)$. The variance is calculated as the mathematical expectation of the squared deviation of the random variable from $M(X)$:

$$D(X) = \sum_{i=1}^{N} (x_i - M(X))^2 P(x_i); \qquad D(X) = \int_{-\infty}^{+\infty} (x - M(X))^2 f(x)dx.$$

The **standard deviation** $\sigma(X)$ is equal to the square root of the variance: $\sigma(X) = \sqrt{D(X)}$.

The **main types of distributions** used in mathematical statistics are: for discrete quantities – binomial (Bernoulli) distribution and Poisson distribution, for continuous variables – normal distribution (Gaussian distribution).

The **binomial distribution**, or **Bernoulli distribution**, is the distribution of the number of occurrences of an event (A) in a series of n independent trials, if in each of these trials the probability of the event p is constant. The probability of occurrence of the value m ($0 \le m \le n$) has the form

$$P(m) = C_n^m p^m q^{n-m},$$

where $q = 1 - p, \quad C_n^m = \dfrac{n!}{m!(n-m)!}$.

For this distribution, $M(X) = np$, $D(X) = npq$.

The **Poisson distribution** (the law of rare events) is the limiting case of the Bernoulli distribution at $n \to \infty$, $p \to 0$ and a constant product $np = a$.

$$P(m) = \frac{a^m}{m!} e^{-a}.$$

For the Poisson distribution $M(X) = D(X) = a$.

Most **continuous** physical quantities in the nature have a distribution close to the normal one.

The **normal distribution** (Gaussian distribution) is given by the formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}},$$

where a and σ are distribution parameters.

For a normal distribution, $M(X) = a$, $D(X) = \sigma^2$.

A normal distribution with parameters $a=0,\ \sigma=1$ is called the **standard normal distribution** and has the form $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$.

7.3. Sample distribution. Graphical representation of the distribution

Let it be required to study a quantitative characteristic of the **general population**, i.e. the largest population, the elements of which possess at least one common property. Suppose that, from theoretical considerations, it was possible to establish which particular distribution the feature under study has. Naturally, the problem arises to estimate the parameters which determine this distribution.

To evaluate the parameters of the general population, from this population, by conducting experiments (trials), we obtain a certain set of *n* values of a random variable. This set of values is called a **sample** of volume *n*.

The distribution of values of a random variable in the sample can be represented in the form of a table containing all the values x_i of the quantity (usually sorted in ascending order) and the number of observations of each value (frequencies) f_i. This type of distribution is called a **variation series**.

The statistical distribution represented by the variational series can be depicted graphically, in particular, in the form of a polygon, histogram, and cumulate.

The **polygon** and the **histogram** of the variational series are graphs of **empirical probability density**. Note that the **sum of the areas** of all the columns forming the histogram is equal to **unity**.

A **cumulate** is a graph of the **empirical distribution function** of a random variable.

Statistical graphs make it possible to evaluate the type of distribution and proceed to finding the parameters of this distribution. In particular, if the distribution is close to normal, it is necessary to find sample estimates of the mathematical expectation and variance. They are expressed by the formulas:

$$\hat{M}(X)=\bar{x}=\frac{1}{n}\sum_{i=1}^{k}f_i x_i\ ; \qquad \hat{D}(X)=\hat{S}^2=\frac{1}{n-1}\sum_{i=1}^{k}(x_i-\bar{x})^2\cdot f_i\ .$$

In addition to these values, for a sample distribution (especially if it does not correspond to the normal one), one often finds its characteristics such as mode (the most common value in the sample), median, and quartiles.

7.4. Testing statistical hypotheses

One of the tasks of mathematical statistics is **verification of statistical hypotheses**, which are formulated during the study of a sample of values $x_1, x_2, ..., x_n$ of a random variable *X*.

A **statistical hypothesis** (denoted by *H*) is a hypothesis about the expected form of the studied probability distribution or about the values of parameters of this distribution. For example, hypotheses about the independence of two random variables, the equality of distribution parameters, etc. can be tested.

The algorithm for testing a statistical hypothesis.

1. A special function of values $x_1, x_2, ..., x_n$ called **statistics** (T) is calculated from the sample $x_1, x_2, ..., x_n$: $T = T(x_1, x_2, ..., x_n)$.

2. A criterion for the statistical hypothesis is formulated, that is, a rule that allows you to reject or accept hypothesis H on the basis of sample data. The criterion determines the critical range of values of statistics T.

3. Hypothesis H is rejected if the value of T belongs to the critical region, and is accepted otherwise.

The described rule for accepting or rejecting a hypothesis does not unambiguously determine the correctness or falseness of the hypothesis. Four cases are possible here.

1. Hypothesis H is *true* and *accepted* according to the criterion. The probability that hypothesis H is correctly accepted is called **confidence probability** (α).

2. Hypothesis H is *true*, but is *rejected* according to the criterion. This case is called a type 1 error. The probability of a type 1 error (P) is called the **significance level** of the criterion and is equal to $P = 1 - \alpha$.

3. Hypothesis H is *false*, but is *accepted* according to the criterion. This case is called a type 2 error. The probability of this error is denoted by β.

4. Hypothesis H is *false* and *rejected* according to the criterion. The probability of this case is equal to $1 - \beta$ and is called **statistical power**.

The listed cases can be presented as a table.

Result of the criterion	Hypothesis H	
	True	False
Accepted	Hypothesis H correctly accepted	Hypothesis H incorrectly accepted (*type 2 error*)
Rejected	Hypothesis H incorrectly rejected (*type 1 error*)	Hypothesis H correctly rejected

Testing the hypothesis about the parameters of the normal distribution.

When testing statistical hypotheses with a known type of distribution of a random variable, one proceeds as follows: from the distribution tables of statistics T one finds a critical value depending T_0 on a predetermined significance level P (in biomedical research it is usually taken equal to 0.01 or 0.05), and checks the inequality $T \le T_0$. If the inequality is true, then hypothesis H is accepted. If it turns out that $T \ge T_0$, then hypothesis H is rejected.

This method is used for testing the hypothesis that the independent observation results $x_1, x_2, ..., x_n$ obey the normal distribution law with an average value $a = a_0$ at known variance σ^2. To test this hypothesis, the following actions are performed.

1. The sample mean is found.

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}.$$

2. The statistics T is calculated by the formula

$$T = \sqrt{n}\,\frac{\overline{x} - a_0}{\sigma}.$$

3. At a given significance level P, the critical value T_0 is found in the corresponding tables of normal distribution.

4. If it turns out that $T > T_0$, then the hypothesis that the sample is taken from the population with an average value a_0 is rejected, otherwise the hypothesis is accepted.

Testing the hypothesis of significance of the difference between two sample means.

This hypothesis is verified using a similar method. The averages (sample means) of two samples with a normal distribution law are compared. The corresponding criterion is called the **Student's criterion** (Student's T test).

Suppose that there are two samples $X_1\{x_{1i}\}$, $X_2\{x_{2i}\}$ with volumes, respectively, n_1 and n_2. The following calculations are performed for these samples:

1. Average values $\overline{x}_1$, $\overline{x}_2$ and variances S_1^2, S_2^2 are calculated.

2. Student's statistics is calculated by the formula

$$T = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

3. Student's distribution has an additional parameter called the number of *degrees of freedom*. In the case $n_1 = n_2 = n$ the number of degrees of freedom is calculated by the formula

$$\nu = n - 1 + \frac{2n - 2}{\frac{S_1^2}{S_2^2} + \frac{S_2^2}{S_1^2}}.$$

4. From the tables of critical values of the Student's criterion for a given value of the significance level P (or confidence probability $\alpha = 1 - P$) and the found number of degrees of freedom ν, one finds the critical value T_c.

5. If $T > T_c$, the hypothesis of a significant difference in sample means is accepted with a given level of significance, otherwise it is rejected.

Correlation between random variables. Pearson and Spearman correlation coefficients

Correlation dependence (correlation) is a statistical relationship between two or more random variables, when changes in the values of one or more of these variables lead to a systematic change in the values of another or other variables. The mathematical measure of correlation of two random variables is the **correlation coefficient** R.

The **method for calculating the correlation coefficient** depends on the type of scale to which the variables belong. So, to measure variables with a quantitative scale, the Pearson correlation coefficient is used. If at least one of the variables has an ordinal scale or is not normally distributed, it is necessary to use Spearman's rank correlation.

The linear **Pearson correlation coefficient** of two random variables *X* and *Y* is calculated using the formula

$$R(X,Y)=\frac{M\left[\ x-M(X)\quad y-M(Y)\ \right]}{S(X)S(Y)}=\frac{M(XY)-M(X)M(Y)}{S(X)S(Y)}.$$

The correlation coefficient can take values in the interval $[-1;1]$. If the random variables are independent, the correlation coefficient is zero. The closer the $|R|$ value to unity, the more reason to consider the correlation between random variables as a linear function. The sign of the correlation coefficient shows the nature of this function: if $R>0$, it is increasing (positive correlation), if $R<0$, it is decreasing (negative correlation).

Spearman's rank correlation coefficient (r_S) is used to compare two quantities with an arbitrary distribution law or to compare qualitative indicators having an ordinal scale. It is calculated as follows.

1. The values x_i, y_i of random variables *X* and *Y* are sorted (individually) in ascending order.

2. A **rank**, that is, a serial number in a sorted sequence, is assigned to each of the values. If among the values of any of the variables there are duplicates, their ranks are **averaged**.

3. For each pair of values (x_i, y_i), the rank difference $R_{xi}-R_{yi}$ is calculated.

4. Spearman's correlation coefficient is calculated by the formula

$$r_S=1-\frac{6\sum_{i=1}^{n}(R_{xi}-R_{yi})^2}{n^3-n},$$

where *n* is the sample size.

As for the Pearson correlation coefficient, the value of the Spearman coefficient varies from −1 (sequences of ranks are completely opposite) to +1 (sequences of ranks coincide completely). A zero value indicates that the random variables are independent.

7.5. Nonparametric methods for testing statistical hypotheses. Wilcoxon and Mann – Whitney tests

The **Wilcoxon test** (*W*), or the Wilcoxon signed rank test, is a nonparametric statistical test (criterion) used to check the differences between two samples of paired or independent measurements by the level of any quantitative characteristic, measured on a continuous or ordinal scale.

The essence of the method is that the absolute magnitudes of shifts (differences in the values of the two quantities) in one or another direction are compared.

To apply the Wilcoxon test, you need to:

1. Calculate the differences $y_i - x_i$ between the individual values in the second and first measurements; find their absolute values $|y_i - x_i|$ and signs $\text{sgn}(y_i - x_i)$. Zero differences are excluded from the sample.

2. Sort the absolute values of the differences in ascending order, assigning them ranks R_i.

3. Calculate the statistics $W = \sum_{i=1}^{N} \text{sgn}(y_i - x_i) R_i$.

4. Determine the critical values for the given sample size $W_c(N)$ from the tables. If $|W| > W_c$, then a shift in one direction reliably prevails, i.e., the results of two measurements are different.

There is another version of this criterion (Wilcoxon *T*-test). Statistics is the smaller of the sums of ranks calculated for different signs of shifts. In this version, the difference hypothesis is accepted when $T < T_c$.

The **Mann-Whitney test** (*U*) is an analogue of the Wilcoxon test for two independent samples. To apply the criterion, you need to:

1. Compose a single ranked series of both tested samples (of volumes n_1 and n_2), arranging their elements in ascending order and assigning a lower rank to a lower value.

2. Separate the ranked series into two parts, consisting respectively of the values of the first and second samples. Separately calculate the sum of ranks for elements of the first sample, and separately for elements of the second sample. Determine the larger of the two rank sums (R_x) corresponding to the sample with n_x elements.

3. Determine the value of the Mann – Whitney *U* test by the formula

$$U = n_1 n_2 + \frac{n_x(n_x + 1)}{2} - R_x .$$

4. From the table for the chosen significance level *P* determine the critical value of the criterion for given n_1 and n_2. If the obtained value of *U* is less than or equal to the table critical value, then the hypothesis about a significant difference between the level of the attribute in the samples under consideration is accepted.

TOPIC 8. Cluster analysis in medical research

8.1. Basic concepts

Cluster analysis is a multidimensional statistical procedure that collects data containing information about a sample of objects and then organizes the objects into relatively homogeneous groups. In other words, cluster analysis is a combination of methods and algorithms for **classifying** objects.

A common question asked by researchers in many areas is how to *organize* the observed data into visual structures – *taxonomies*. Mathematically, a taxonomy is a tree structure of classifications of a certain set of objects. At the top of this structure is a unifying single classification – the *root taxon*, which applies to all objects of a given taxonomy. Taxa below the root are more specific classifications that relate to subsets of the general set of classified objects.

This type of analysis can be used in biology (e. g. for classification of animals), psychology, medicine and in many other areas of human activity. For example, in the medical field, clustering of diseases, methods of treating diseases, or symptoms of diseases leads to widely used taxonomies. In the field of psychiatry, the correct diagnosis of clusters of symptoms, such as paranoia, schizophrenia, etc., is crucial for successful therapy.

Application of cluster analysis involves the following **steps**:

1. Sampling for clustering. It is understood that it makes sense to cluster only quantitative data.
2. Definition of a set of features, that is, variables by which objects in the sample will be evaluated.
3. Calculating the values of a measure of similarity (or difference) between objects.
4. Applying the cluster analysis method to create groups of similar objects.
5. Validation of the results of the cluster decision.

The **goals** of clustering:

1. Understanding data by identifying a cluster structure. Dividing the sample into groups of similar objects makes it possible to simplify further data processing and decision making by applying a specific analysis method to each cluster.
2. Data compression. If the initial sample is excessively large, then you can reduce it by leaving one of the most typical representatives from each cluster.
3. Detection of novelty. Atypical objects are selected that cannot be attached to any of the clusters.

8.2. Determining a measure of similarity between objects

Cluster analysis is based on combining objects into large enough clusters by determining some **measure of similarity (distance)** between objects. The result of such clustering is a *hierarchical tree*.

The method of tree clustering is used to form clusters taking into account the distances between objects. These distances can be determined in one-dimensional or multidimensional space. The task of researchers is to choose the correct method for calculating the distance.

There are several different **methods for calculating distance**: Euclidean distance, squared Euclidean distance, Manhattan distance, power-law distance, Hamming distance. In problems of cluster analysis in biology and medicine, the Euclidean distance and the Hamming distance are most often used.

The most direct way to calculate distances between objects in multidimensional space is to calculate **Euclidean distances**.

The **Euclidean distance** $p_E(X,Y)$ between an object X with features $x_1, x_2, ..., x_n$ and an object Y with features $y_1, y_2, ..., y_n$ is determined by the formula:

$$p_E(X,Y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}.$$

Euclidean distance is an analogue of the geometric distance between objects in the case of two-dimensional or three-dimensional space. However, the clustering algorithm does not "care" about whether the "distances" provided for analysis are real geometric distances or some other distance measures.

The **Hamming distance** $d(X,Y)$ is the number of positions in which the corresponding characters of two sequences of characters of the same length are different.

For example, compare two sequences of characters: 10<u>1</u>1<u>1</u>01 and 10<u>0</u>1<u>0</u>01. As you can see, these sequences have 2 positions in which the characters do not match. This means that the Hamming distance *d*(10<u>1</u>1<u>1</u>01, 10<u>0</u>1<u>0</u>01) = 2. Similarly, you can calculate the Hamming distances for decimal numbers, words and other sequences of characters. For example: *d*(7<u>3</u>62<u>9</u>604, <u>5</u>35<u>2</u>2<u>2</u>04) = 4; *d*(m<u>a</u>x<u>i</u>mum, m<u>i</u>n<u>i</u>mum) = 2.

For nucleic acids (DNA and RNA), the possibility of hybridization of two polynucleotide chains with the formation of a secondary structure – a double helix – depends on the degree of complementarity of the nucleotide sequences of both chains. With increasing the Hamming distance, the number of hydrogen bonds formed by complementary base pairs decreases and, accordingly, the stability of the double chain also decreases. Starting from a certain critical value of the Hamming distance, hybridization becomes impossible.

For evolutionary divergence of homologous DNA sequences, the Hamming distance is a measure by which one can judge the time elapsed since the homologues diverged, for example, the length of the evolutionary segment separating homologous genes from the precursor gene.

8.3. Cluster merging methods

At the first step, when each object is a separate cluster, the distances between these objects are determined by the chosen measure. However, when several objects are linked together, the question arises: how to determine the distances between the clusters. In other words, you need a merging or linking rule for two clusters. There are various **methods of cluster merging**.

1. The **nearest neighbor** method: the degree of similarity is estimated by the distance between the nearest objects of the clusters.

2. The **most distant neighbor** method: the degree of similarity is determined by the largest distance between objects in different clusters.

3. The **centroid** method: the distance between the clusters is determined by the distance between their “centers of mass”. A variation of the centroid method is the weighted centroid method (median method), which uses weighting factors to account for cluster sizes.

4. The **average link method** (unweighted pairwise average): the distance is defined as the arithmetic average of all pairwise distances between the representatives of the groups in question. For clusters that vary significantly in size, the weighted pairwise average method is used, which takes into account the size of the clusters.

TOPIC 9. Formal logic in solving problems of diagnosis and prevention of diseases

9.1. Concept of knowledge

Modern information technology (IT) is increasingly bringing medicine closer to exact sciences. Information technologies allow to automate the processes of diagnosis, prognosis and the choice of treatment methods, make it possible to determine patterns during the course of the disease, comparing at the same time many of its signs, provide effective processing of a large amount of information.

Currently, medicine already knows more than 10 thousand diseases and about 100 thousand symptoms that can manifest themselves in various combinations. Due to such a huge flow of information, the process of making a diagnosis is becoming more complicated every year, which leads to medical errors. Here information technologies come to the aid of a doctor in the form of, for example, **artificial intelligence systems** (AIS).

Artificial intelligence (AI) refers to the ability of automatic or automated systems to take on the functions of human intelligence, for example, to make optimal decisions based on analysis of external influences and taking into account previous experience.

AIS are divided into several classes: expert systems, pattern recognition systems, robotics, systems for communicating with computers in a natural language.

Currently, the **field of practical application** of AIS includes **difficult-to-formalize tasks**, which are characterized by the following features:

- the task cannot be defined in a numerical form (requires a symbolic representation);
- the algorithmic solution of the problem is unknown (although it may exist) or cannot be used due to limited resources (computer memory, performance);
- the goals of the task cannot be expressed in terms of a well-defined objective function or there is no exact mathematical model of the problem.

All this is just characteristic of the tasks of medical diagnosis, treatment and prevention of diseases.

The main concept used in AI is the concept of "knowledge".

Knowledge is a collection of statements about the world, properties of objects, laws of processes and phenomena, as well as the rules for the logical inference of some statements from others and the rules for using them to make decisions. The main difference between knowledge and data is their structure and activity: appearance of new facts in the knowledge base or establishment of new relationships between them can become a source of changes in decision making.

Knowledge can be classified in several ways.

1. **Factual** and **strategic** knowledge. Factual knowledge is knowledge about the main laws of the subject area, allowing to solve specific productional, scientific and other problems, that is, facts, concepts, relationships, estimates, rules, heuristics. Strategic knowledge contains decision-making strategies in the subject area.

2. **Facts** and **heuristics**. Facts indicate circumstances that are well-known in a particular subject area. Such knowledge is also called *textual* knowledge, bearing in mind its sufficient coverage in specialized literature and textbooks. Heuristics are based on the individual experience of a specialist (expert) in the subject area,

accumulated as a result of many years of practice. This category of knowledge often plays a decisive role in the development of intelligent programs.

3. **Declarative** and **procedural** knowledge. Declarative knowledge is knowledge such as "A is B". Such knowledge is characteristic for databases. As a rule, these are *facts* of the type "scarlet fever is an infectious disease". Declarative knowledge characterizes the objects on which the actions are performed. Procedural knowledge includes information on how to transform declarative knowledge, i.e., *actions* to obtain a result.

9.2. Knowledge properties

Knowledge is characterized by a number of **properties** that distinguish them from traditional data models. We list these properties.

1. **Structuring**. Knowledge consists of separate information units between which classifying relationships can be established: genus – species, class – element, type – subtype, part – whole, etc. Information units can, if necessary, be divided into smaller ones and combined into larger ones according to the principle of "nested dolls".

2. **Internal interpretability**. Together with an information unit representing the data element itself, a computer system stores a **name system** associated with such an information unit. The presence of a name system allows the information system to "know" what is stored in its memory, and, therefore, to be able to respond to queries about the contents of memory that can be generated during execution of programs inside the system or arrive from users or other systems.

3. **Coherence**. It is possible to establish a wide variety of relations between information units, reflecting the connections of phenomena and facts. When a system of relations arises between information units in the system's memory, new information units can be determined by fragments of this structure.

4. **Semantic metric.** On the set of information units stored in the memory, some *scales* (relationships) are introduced, allowing to evaluate their *semantic proximity* (i.e., the strength of the *associative connection* between them). This allows to find knowledge that is *close to already found* one in the information base.

5. **Activity**. This property emphasizes the fundamental difference between knowledge and data. The performance of certain actions in the AIS is initiated by the state of the knowledge base. Emergence of new facts and relationships can activate the system, i.e., a certain structure of declarative knowledge turns out to be an activator for procedural ones. The activity of the knowledge base allows the AIS to form motives, set goals and build procedures for their implementation.

9.3. Logical model of knowledge representation. Algebra of Logic

The central issue in building knowledge-based systems is the choice of the form of knowledge representation. The most common are the following knowledge representation models:

- logical models;
- production models;
- network models;
- frame models.

The most common form of knowledge representation is the **logical model**. The main apparatus (system of rules) for working with logical models is the apparatus of the **algebra of logic**. The task of the algebra of logic is the optimization of logical expressions, i.e., the reduction of expressions to a form containing the least number of arguments and operations on them.

A **statement** is a sentence whose content can be evaluated as true or false. Statements are denoted by logical variables *A, B, C, ..., Z*. Logical variables can take only two values: "**true**" (T) and "**false**" (F), which are called **truth values**. In the algebra of logic, they correspond to the numerical values of **1** and **0**.

Complex statements are formed from simple statements using logical connectives (or logical operations).

Basic logical operations.

1. **Negation** (**inversion**, logical **NOT**). Denoted $\bar{X}$ or $\neg X$. Changes the truth value of the statement to the opposite.

2. **Disjunction** (logical addition, logical **OR**). Denoted $X \vee Y$ or $X + Y$. True when *at least one* of the variables *X* or *Y* is true.

3. **Conjunction** (logical multiplication, logical **AND**). Denoted $X \wedge Y$ or $X \& Y$. True when *both* variables *X* and *Y* are true. Disjunction and conjunction operations are defined similarly for a larger number of variables.

4. **Logical NOR** (OR-NOT, Peirce function). Denoted $X \downarrow Y = \overline{X \vee Y}$. True when both variables are false.

5. **Logical NAND** (AND-NOT, Sheffer function). Denoted $X \mid Y = \overline{X \wedge Y}$. False only when both variables are true.

6. **Equivalence**. Denoted $X \sim Y$. True when both variables are true or both are false.

7. **Inequivalence** (**exclusive OR**, **XOR**). Denoted $X \oplus Y$. True if *only one* of the two variables *X* or *Y* is true (but not both).

8. **Implication**. Denoted $X \to Y$ ("*X* implies *Y*", "if *X*, then *Y*"). False when the statement *X* is true and *Y* is false, true in all other cases.

The properties of logical functions and their expressions through conjunction, disjunction, and negation functions are presented in the **truth table**.

X	Y	$X \vee Y$	$X \wedge Y$	$X \downarrow Y$	$X \mid Y$	$X \sim Y$	$X \oplus Y$	$X \to Y$
				$\bar{X} \wedge \bar{Y}$	$\bar{X} \vee \bar{Y}$	$(\bar{X} \wedge \bar{Y}) \vee \vee (X \wedge Y)$	$(\bar{X} \wedge Y) \vee \vee (X \wedge \bar{Y})$	$\bar{X} \vee Y$
0	0	0	0	1	1	1	0	1
0	1	1	0	0	1	0	1	1
1	0	1	0	0	1	0	1	0
1	1	1	1	0	0	1	0	1

Examples of calculating the truth value of a logical expression:

1) at $X = 0$, $Y = 1$, $Z = 1$ expression $X \vee (Y \wedge Z) = 1$;

2) at $X = Y = Z = 0$ expression $X \to (Y \sim Z) = 1$.

Complex statements written not with words, but using logical variables and signs of logical operations, are called **formulas**. The algebra of logic *does not consider the specific content of statements*, but performs analysis and synthesis of formulas and studies relations between formulas.

For example, the complex statement "if 30 is divided by 2 and by 3, then 30 is divided by 6" can be written as a formula $A \wedge B \to C$. This formula will correspond not only specifically to this statement, but also to the set of all other statements that have the same structure.

9.4. Expert systems

An **expert system** (ES) is an information computer system capable of partially replacing an expert in resolving a problem situation. Expert systems belong to the class of artificial intelligence systems in which the decision-making logic of an experienced specialist is implemented. Expert systems are widely used in medicine to support decision-making in the field of diagnostics, prognosis, treatment, management, training, etc.

Many **types of medical expert systems** exist, among which are the following:

- *data interpretation* expert systems that determine the content of data, in particular, data from medical observations and experiments;
- *diagnostic* expert systems that determine the nature of the deviation of the state of the object from the normal condition (on the basis of this, it is assigned to the corresponding category);
- *monitoring* expert systems focused on continuous interpretation of real-time data and signaling that certain parameters have exceeded acceptable boundaries, in particular, expert medical monitoring systems in intensive care units;
- *forecasting* expert systems logically build probabilistic conclusions about the future course of events, based on current situations, taking into account all circumstances. In medicine, using these systems, the course of the disease is predicted with different treatment protocols, determining the best protocol in each case;
- *training* expert systems determine deficiencies in studying a particular discipline, collecting and analyzing data on "weak points", and then give the necessary explanations and recommendations according to which the necessary exercises are selected to improve training of future doctors;
- *planning* expert systems determine the optimal action plans of objects capable of performing certain functions;
- *design* expert systems prepare documentation for the creation of objects with predetermined properties, containing even ready-made drawings and an appropriate description.

To build an ES, formal and informal (neural network) logic approaches can be used. The specifics of each of these approaches are given in the table.

ES type / Property	Formal logic ES	Informal logic ES (neural networks)
Knowledge source	Formalized experience of an expert, expressed in terms of logical statements, rules and facts, unconditionally accepted by the system	Cumulative experience of an expert teacher who selects examples for training + own experience of the neural network learning from these examples
Nature of knowledge	Formal-logical "left-hemispheric" knowledge in the form of rules	Associative "right-hemispheric" knowledge in the form of connections between network neurons
Knowledge development	Expanding the set of rules and facts (knowledge base)	Additional training on an extra sequence of examples, with specification of the boundaries of categories and formation of new categories
Role of an expert	Defines, on the basis of rules, the full amount of knowledge of the expert system	Selects characteristic examples without specifically formulating the argumentation for his choice
Role of the artificial system	Search for a chain of facts and rules for proving a judgment	Formation of individual experience in the form of categories derived from examples, and categorization of images

9.5. Expert system structure

1. **Knowledge base** is the whole set of all available information about the problem area for which the given expert system is designed, recorded using certain formal structures for representing knowledge (a set of rules, frames, semantic networks, etc.).

2. **Knowledge acquisition subsystem** automates the process of filling and replenishing the expert system with expert knowledge, that is, it provides the knowledge base with all the necessary information from the particular subject area.

3. **Logical inference machine** is a formal logical system in the form of a software module, which, using the rules and methods of the knowledge base, converts specific information about the object to a form corresponding to the purpose of the expert system (diagnosis, action plan, etc.).

4. **Working memory** (**database**) is intended for storing the initial and intermediate facts of the problem currently being solved.

5. **Dispatcher** determines the functioning of the expert system, plans the procedure for setting and achieving goals.

6. **Interface** provides user communication with the expert system in a form convenient for him; allows the user to transfer information that makes up the contents of the database, to contact the system with a question or for an explanation.

7. **Explanation module** explains how the system received a solution to this problem (or why it did not receive this solution) and what knowledge it used doing so. In other words, the explanation module creates a report on the work done.

TOPIC 10. Decision making in medicine

10.1. Basic concepts

In various medical tasks (collecting information about the patient, diagnosis, choice of treatment tactics), the doctor faces a common problem – the decision-making problem. At the same time, requirements to the accuracy of the diagnosis and its reliability, i.e., truth, are increasing every year.

Decision making implies a special process of human activity aimed at choosing the most acceptable solution to the problem. An example is the process of deciding on the type (form) of a disease according to known source information (test results, external manifestations of the disease).

The following **main stages** of the decision-making procedure are distinguished:

1. Definition of purpose.
2. Formation of the set of alternatives (determination of the set of possible solutions).
3. Formation of an assessment method that allows to compare the alternatives.
4. Selecting the best solution from the set of possible solutions (optimization problem).

Decision making is essentially nothing more than a **choice**. To make a decision is to choose a specific option from a set of options that are commonly called **alternatives**. The set of alternatives depends on the existing knowledge base and on the problem itself.

A **solution** is an alternative (alternatives) that satisfies the rules contained in the preference system.

The **preference system** is a set of rules, criteria by which alternatives are compared and decisions are made. These criteria, expressed mathematically, are called **objective functions**.

The **consequence** of making a decision is an event (*outcome*), the possibility of which is determined by this decision.

The general decision-making problem (the choice problem) can be formulated as follows.

Let X be the set of alternatives (solutions), Y the set of possible consequences (outcomes, results). It is assumed that there is a causal relationship between the choice of an alternative x_i and the onset of the appropriate outcome y_i. In addition, it is assumed that there is a mechanism to assess the quality of choice, usually by evaluating the quality of the outcome. It is required to choose the best alternative for which the corresponding outcome has the best quality score.

10.2. Classification of decision-making tasks

Based on the connections between decisions and outcomes, the following classification of decision-making problems is adopted.

1. **Deterministic** problems. Each of the alternatives leads to a *unique* result.

2. **Non-deterministic** problems. The opposite case, when each alternative is associated with more than one outcome (non-deterministic task) breaks down into two types:

- decision-making problems in **risk** conditions (probabilistic **certainty**): each alternative x_i corresponds to a probability density function on the set of outcomes Y;

• decision-making problems in the conditions of **stochastic** (probabilistic) **uncertainty**, when the specified probability density is unknown.

In conditions of probabilistic uncertainty of alternative-outcome relationships, in turn, there are two types of tasks:

• decision-making tasks in the conditions of **passive interaction** of the decision maker (DM) and the external environment, that is, the external environment behaves passively with respect to the DM;

• decision-making tasks in a **conflict (game)** situation. In this case, the external environment behaves *actively* with respect to the decision-maker, which is manifested by the actions of another person.

Decision making tasks are also divided into **static** and **dynamic**. If during the decision-making process the decision maker does not receive or lose information, then the decision-making can be considered as an instantaneous act. Corresponding tasks are called **static**. On the contrary, if a decision maker receives or loses information during the decision-making process, then such a task is called **dynamic**.

10.3. Decision making algorithm

All the requirements formulated in real problems and written in the form of mathematical expressions make up the so-called **mathematical formulation** of the problem. The process of mathematical formulation of the problem and its subsequent solution can be represented in the form of a series of stages.

1. **Studying the object**: analysis of the features of the object functioning. At this stage, factors affecting the object are identified, and the degree of their influence is determined; the characteristics of the object are studied under various conditions; optimizing criteria (*objective functions*) are selected.

2. **Descriptive modeling**: definition and recording the main relationships and dependencies between the characteristics of a process or phenomenon.

3. **Mathematical modeling**.

4. **Choice or creation of a solution method**. At this stage, a multitude of *possible solutions* is created, i.e., possible sets of the desired variables that satisfy the constraints of the problem.

5. **Solving the problem**. The solution to the problem is the set of values from the multitude of possible solutions for which *the objective function reaches its maximum or minimum value*. Problems that describe the behavior of real objects, as a rule, have many variables and many dependencies between them. Therefore, in a reasonable time they can only be solved using a computer.

6. **Analysis of the decision**. Analysis of the decision can be **formal** and **substantial**. In the *formal* (mathematical) analysis, the correspondence of the obtained solution to the constructed *mathematical model* is checked (are the input data correct, are the computer programs functioning correctly, etc.). In a *substantial* analysis, the correspondence of the obtained solution to the *real object* that was modeled is checked. As a result of substantial analysis, changes can be made to the model, and the whole process is repeated.

7. **Analysis of solution stability**. To check the stability of the solution, changes are made to the initial data within the limits of possible errors or intervals of existence of variables, and then the behavior of the solution is studied by analytical or numerical methods.

10.4. Sensitivity and specificity of the diagnostic test

When developing and applying diagnostic methods and tools, one always poses a question on **reliability** of these methods. Reliability of a test used to separate healthy people from sick ones can be characterized using such properties of the test as sensitivity and specificity.

The problem of the reliability of the diagnostic method is formulated as follows. An arbitrary collection of patients is given that can be in one of two conditions with respect to a certain disease – *normal* or *pathology*. Each of these conditions has its own distribution function for results of the diagnostic test. For each patient, you need to make the best choice between these two conditions, that is, in fact, make a diagnosis of "normal" or "pathology" based on this test.

According to the relation between the result of the checked diagnostic method and the true state of the patient, 4 different outcomes are possible:

Test results	True state	
	Pathology	Normal
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

1) TP – a *true positive* diagnosis (determined by the diagnostic method of the "gold standard");

2) FP – a *false positive* diagnosis (presence of a positive test if there is no disease), also called a *type 2 error*;

3) TN – a *true negative* diagnosis (determined by the diagnostic method of the "gold standard");

4) FN – a *false negative* diagnosis (a negative test in the presence of a diagnosed disease), called a *type 1 error*.

Sensitivity is the proportion of positive results that are correctly identified as positive (that is, the probability of recognizing a truly ill patient as ill according to the test results). Sensitivity reflects the property of the test to *accept the true hypothesis* with high accuracy, that is, to avoid type 1 errors.

$$\text{Sensitivity} = \frac{N(TP)}{N(TP) + N(FN)}.$$

In other words, sensitivity is the ratio of the number of true positive test results to the total number of patients with this disease.

Specificity of the test is the proportion of negative results that are correctly identified as negative (that is, the probability of recognizing a truly healthy person as healthy according to the test results). The specificity reflects the property of the test to *reject the false hypothesis* with high accuracy, i.e., to avoid type 2 errors.

$$\text{Specificity} = \frac{N(TN)}{N(TN) + N(FP)}.$$

Specificity is the ratio of the number of true negative results to the total number of patients that do not have this disease.

Both characteristics can take values from 0 to 1 and are often expressed as a percentage (from 0% to 100%, respectively).

It should be noted that the sensitivity and specificity indicators reflect *different properties* of the diagnostic test and are *independent of each other*, for example, a test may have a sensitivity close to 100 % and low specificity, or vice versa.

10.5. Calculation of the probability of having a disease with a positive test

Using the characteristics of sensitivity and specificity of the test and the frequency of cases of the disease, it is possible to calculate the probability of presence of the disease if the test is positive.

The calculation is made according to the **Bayes formula**. The Bayes theorem (or Bayes formula) is one of the basic theorems of elementary probability theory which allows to determine the probability of an event, provided that another statistically interdependent event has occurred.

Let the patient suspect a certain disease, i.e. there is a hypothesis H about the presence of this particular disease. For this disease, the initial (*a priori*) probability $P(H)$ of its occurrence in the patient is known. The patient undergoes a diagnostic test, the positive result of which is denoted as D. It is required to find the probability of illness with a positive test result $P(H \mid D)$.

Then, according to the Bayes formula, the probability $P(H \mid D)$ that a person is really sick with a positive test result (*a posteriori* probability) is

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D \mid H)P(H) + P(D \mid \bar{H})(1 - P(H))},$$

where $P(H)$ is the a priori probability of the disease;

$1 - P(H)$ is the probability of absence of the disease;

$P(D \mid H)$ is the probability of a true positive test for this disease, i.e., the sensitivity of the test;

$P(D \mid \bar{H})$ is the probability of a false positive diagnosis (the probability of a positive test result in a healthy person, i.e., the probability of a type 2 error).

10.6. The problem of differential diagnosis of diseases

Bayes' theorem allows one of several diagnostic hypotheses to be selected based on the calculation of the probability of disease from the probability of symptoms found in patients. The most common problem of this type is the problem of differential diagnosis between two diseases with one common symptom.

In clinical practice, symptoms are known whose presence unambiguously determines the disease. On the other hand, there are symptoms that exclude a particular diagnosis. However, most often the main symptoms determining the clinic can occur with a certain frequency in various diseases. In such cases, decision making technology uses the concept of **odds** associated with Bayes' theorem.

Consider the problem of differential diagnosis between diseases H_1 and H_2 having a common symptom D. Let the average probability of the disease H_1 be $P(H_1)$, the probability of the disease H_2 be $P(H_2)$. The probability ratio $\frac{P(H_1)}{P(H_2)}$

shows the **initial odds** in favor of the diagnosis H_1 without taking into account additional conditions:

$$C_0 = \frac{P(H_1)}{P(H_2)}.$$

The task is to find the **ultimate odds** C_1 of diagnosis H_1 in the presence of symptom *D*, i.e., the ratio

$$C_1 = \frac{P(H_1 \mid D)}{P(H_2 \mid D)}.$$

Bayes' theorem states that the final probability $P(H \mid D)$ of a hypothesis is proportional to its initial probability $P(H)$, multiplied by its **likelihood** $P(D \mid H)$:

$$P(H \mid D) \sim P(H)P(D \mid H).$$

Accordingly, to calculate the ultimate odds, it is necessary to find the ratio of two likelihoods, or the **likelihood ratio** (*LR*), i.e., the ratio of the probabilities $P(D \mid H_1)$, $P(D \mid H_2)$ of the symptom *D* in diseases H_1 and H_2.

$$LR = \frac{P(D \mid H_1)}{P(D \mid H_2)}.$$

The ultimate odds C_1 of diagnosis H_1 in the presence of the characteristic symptom *D* are calculated as the initial odds C_0 multiplied by the likelihood ratio *LR*:

$$C_1 = C_0 \cdot LR = \frac{P(H_1)}{P(H_2)} \cdot \frac{P(D \mid H_1)}{P(D \mid H_2)}.$$

TOPIC 11. Mathematical modeling of biomedical processes

11.1. Basic concepts

One of the most important research methods in various fields of science, including medicine, is the modeling method.

A **model** is an object of any nature artificially created by a person that replaces or recreates the object under study in such a way that studying the model can provide *new information* about the object. In other words, a model is a *new* object that reflects the *essential features* of the studied object, phenomenon or process.

A model is not only a reflection of our knowledge about the object under study, but also *a source of new knowledge*. Studying the model allows to evaluate the behavior of the simulated object at new conditions or under various influences that cannot be verified on a real object (human research) or are difficult to verify (expensive objects or negative consequences of experiments).

The object of research in biology and medicine is a living organism, which is a fairly complex system. Therefore, the researcher inevitably chooses a *simplified point of view* that is suitable for solving a specific task. The choice of model is determined by the objectives of the study.

Models are divided into three classes:

- material;
- energy;
- information.

Material models are models that reproduce the *structure of an object* and the *relationships of its parts*. Examples of such models in medicine are various prostheses that look similar to the real body parts that they replace.

Energy models are used to model *functional relationships* in the studied objects. These models do not look like modeled objects in appearance, but their goal is *to perform the functions* of these objects. Examples of such models in medicine are the artificial kidney or artificial respiration apparatus. The properties of material and energy models can be combined. Such models include biocontrolled prostheses, an artificial eye lens, and the latest developments in the field of artificial hearts.

Information models are *descriptions* of an object. Until recently, in biomedical research, mainly verbal models were used to describe the operation of biological systems. However, using verbal models, it is difficult to state the laws of operation of the studied object clearly. Therefore, **mathematical models** become increasingly popular. They use *quantitative relationships* between the parameters of the biosystem under study. These models are the most important models in biomedical research.

Mathematical models are divided into **deterministic** and **probabilistic**. In deterministic models, variables and parameters are assumed to be described by deterministic functions. In probabilistic models, variables and parameters are random functions or random variables. Deterministic mathematical models most often are represented as a *system of algebraic or differential equations*. Probabilistic models are based on the results of experimental determination of dynamic characteristics of objects based on methods of *mathematical statistics*.

Let us consider several examples of mathematical models used in medicine and biology.

10.2. "Predator-prey" mathematical model

For the first time in biology, the Italian mathematician Vito Volterra and colleagues proposed a mathematical model of the periodic change in the number of antagonistic animal species. This model was a development of the idea outlined in 1924 by Alfred Lotka in the book Elements of Physical Biology. Therefore, this classic model is known as the **Lotka-Volterra model**.

The simulation problem is formulated as follows. In some ecologically closed area, there are animals of two species (for example, lynx and hare). Hares (**preys**) feed on plant foods available in sufficient quantities (this model does not take into account the limited resources of plant foods). Lynxes (**predators**) can eat only hares. It is necessary to determine how the number of preys and predators will change over time.

We denote the number of preys by *N*, and the number of predators by *M*. The values of *N* and *M* are functions of time *t*. In this model, we take into account such factors:

1) natural reproduction of preys;
2) natural extinction of preys;
3) reduction in the number of preys due to consumption by predators;
4) natural extinction of predators;
5) increase in the number of predators due to reproduction in the presence of food.

It is necessary to derive *equations* that include all of the indicated factors and that describe the dynamics, i.e. change in the number of preys and predators over time.

Let during a time interval Δt the number of preys changes by ΔN, and the number of predators changes by ΔM.

The change in the number of **preys** ΔN over time is determined by the first three factors from the list above.

The increase in the number of preys due to their **natural reproduction** (ΔN_1) is proportional to the number of preys *N* currently existing, with a proportionality coefficient A:

$$\Delta N_1 = AN\Delta t .$$

The decrease in the number of preys due to **natural extinction** (ΔN_2) is also proportional to their number N at the moment, with the coefficient of proportionality B (note that $B < A$):

$$\Delta N_2 = -BN\Delta t .$$

The minus sign indicates a decrease in population.

The basis of the equation that describes a decrease in the number of preys due to **consumption by predators** (ΔN_3), is the idea that the more often they meet, the faster the number of preys decreases. The frequency of encounters between a predator and a prey is proportional to both the number of preys *N* and the number of predators *M*. Denoting the proportionality coefficient by *C*, we write:

$$\Delta N_3 = -CMN\Delta t .$$

Given all three factors, we obtain the equation for the change of prey population

$$\Delta N = AN\Delta t - BN\Delta t - CMN\Delta t ,$$

or, in differential form,

$$\frac{dN}{dt} = AN - BN - CMN .$$

The change in the number of **predators** ΔM is determined by the remaining two factors from the list.

The increase in the number of predators due to **natural reproduction** (ΔM_1) depends on the number of predators M and the amount of food (preys) N with a proportionality coefficient Q:

$$\Delta M_1 = QMN\Delta t\,.$$

A decrease in the number of predators due to **natural extinction** (ΔM_2) is proportional to the number of predators M at the moment with a coefficient P:

$$\Delta M_2 = -PM\Delta t\,.$$

Adding these two factors, we obtain

$$\Delta M = QMN\Delta t - PM\Delta t\,,$$

or, in the differential form,

$$\frac{dM}{dt} = QMN - PM\,.$$

As a result, we get a system of two differential equations:

$$\begin{cases} \dfrac{dN}{dt} = AN - BN - CMN \\ \dfrac{dM}{dt} = QMN - PM \end{cases}$$

The solution of this system gives a periodic dependence of the number of predators and preys on time, as shown in the graph.

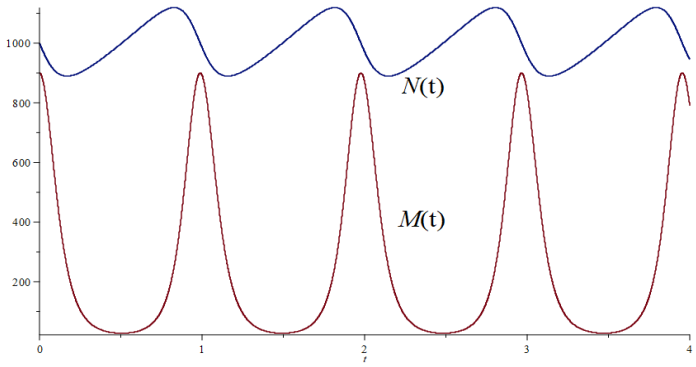

Fig. 5. Behavior of the predator-prey model

11.3. Mathematical model of immune response

Immunity is a complex set of body reactions to invasion of *antigens*, i.e. foreign objects: molecules, viruses, cells, tissues, etc. A specific immune response at the molecular level begins with the fact that specialized *plasma cells* produce a large number of protein molecules – *antibodies* that neutralize antigens.

Antibodies have a conformation complementary to a specific surface area of the antigen. Therefore, the antibody combines with the antigen, like a key with a lock, and the complex formed in this process is lysed by enzymes.

Consider a model of the immune system during a prolonged infectious disease. This model is used in clinical practice in the treatment of viral hepatitis and acute pneumonia.

The interaction of antigens and the body's immune forces in this mathematical model has a nature similar to the behavior of the “predator-prey” system. The “prey”

here is a foreign agent, which in the model will be described by the amount of the corresponding antigen *X*, and the "predator" is the antibody in quantity *Y*, formed by the plasma cells in quantity *Z*. All three variables change over time *t*.

In this model, the following processes and factors are taken into account.

1. Reproduction of antigens with rate AX.

2. The natural decay of antigens with rate $-CX$.

3. The natural breakdown of antibodies with rate $-LY$.

4. Natural death of plasma cells with rate $-NZ$.

5. Antigen-antibody interaction in the agglutination reaction, the frequency of which is proportional to the probability of the corresponding antibody meeting with the antigen. This interaction leads to a decrease in the number of antigens and antibodies with rates $-BXY$, $-KXY$, respectively.

6. Production of antibodies by plasma cells and their entry into the blood with rate DZ.

7. The formation of plasma cells at a rate depending on the concentration of antigens *X* in a complex way: $MF(X)$.

Here *A*, *B*, *C*, *D*, *K*, *L*, *M*, *N* are proportionality coefficients.

In this model, the coefficients and are considered to be temperature dependent, and the function $F(X)$ has the form $F(X)=\frac{X}{Q+X}$, where *Q* is a certain coefficient.

Finding the rate of change over time of the number of antigens $\frac{dX}{dt}$, antibodies $\frac{dY}{dt}$ and plasma cells $\frac{dZ}{dt}$, taking into account all of the above factors, we obtain a system of three differential equations:

$$\begin{cases} \frac{dX}{dt} = AX - BXY - CX; \\ \frac{dY}{dt} = DZ - KXY - LY; \\ \frac{dZ}{dt} = \frac{MX}{Q+X} - NZ. \end{cases}$$

The study of the mathematical model consists in solving this system of differential equations with known coefficients *A*, *B*, *C*, *D*, *K*, *L*, *M*, *N*, *Q* and under the initial conditions (at time $t=0$) $X(0)$, $Y(0)$, $Z(0)$. It is important to note that the same model under different initial conditions or coefficients gives completely different dynamics of the process. In this case, four main forms of the course of an infectious disease are possible:

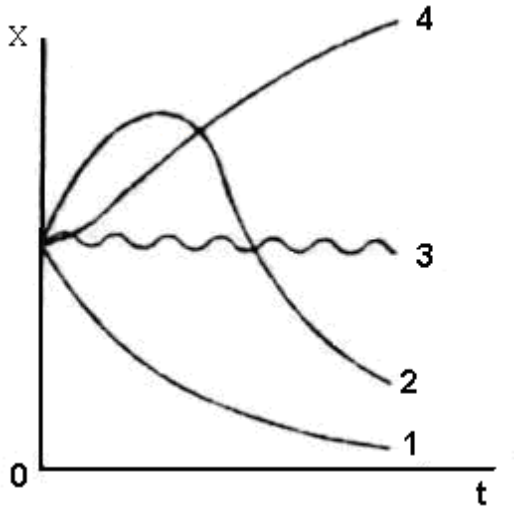

Fig. 6. Possible forms of the infectious disease course

1. Subclinical form: passes without physiological disorders in the body and without external manifestations. Immune defenses easily destroy antigens, preventing them from proliferating up to a dangerous limit.

2. Acute form: in this case, the body is attacked by an unknown antigen in large quantities. First, its enhanced reproduction occurs. When the immune system produces enough antibodies against it, the number of antigens will sharply decrease.

3. Chronic form: dynamic equilibrium is established in the amounts of antigens and antibodies (similar to the predator-prey model). The disease condition becomes persistent.

4. Lethal form: the immune response is too late, and a large number of antigens leads to irreversible changes in the organism.

11.4. Mathematical model of infectious disease spread in a city

The process of spread of an infectious disease in the simplest case can be described by a system of three 1st order differential equations:

$$\begin{cases} \dfrac{dX}{dt} = -QAXY, \\ \dfrac{dY}{dt} = QAXY - \dfrac{1}{R}Y, \\ \dfrac{dZ}{dt} = \dfrac{1}{R}Y, \end{cases}$$

where Q is the number of inhabitants of the settlement; A is the average number of inhabitants that every patient infects every day; R is the average duration of the disease, days; X is the number of healthy people who have not suffered this disease and do not have immunity to it; Y is the number of sick people; Z is the number of those who have been ill and acquired immunity.

This model makes it possible to calculate the rate of spread of the disease among the population, the maximum number of people simultaneously sick, the total number of people who have had the disease, and the duration of the epidemic.

11.5.Pharmacokinetic Models

Pharmacokinetics studies the distribution of a biologically active substance in the body and the change in its concentration over time. Biologically active substances include, in particular, drugs. The laws of change of the drug concentration in the body are different depending on methods and parameters of its administration and excretion.

To simplify the modeling in pharmacokinetics, the body is conventionally divided into **chambers**, i.e., into parts in which the studied drug is distributed evenly.

The simplest pharmacokinetic model is a **linear single-chamber model**. It is adequate to describe the behavior of drugs injected into blood.

In a **linear** pharmacokinetic model, the rate of drug elimination from the body $\frac{dM}{dt}$ is proportional to the first degree of the drug mass M in the chamber. We suppose that the drug is evenly distributed throughout the body volume V (this model is called a **single-chamber** model).

The differential equation of the single-chamber linear pharmacokinetic model is:

$$\frac{dM}{dt} = -kM\ ,$$

where k is the **elimination constant**, i.e., the coefficient of removal of the drug from the body.

The solution to this differential equation has the form

$$M = M_0 e^{-kt},$$

where M_0 is the mass of the drug at $t = 0$, i.e., immediately after administration.

A similar time dependence is true for the drug concentration $C = \frac{M}{V}$:

$$C = C_0 e^{-kt}\ ,$$

where C_0 is the concentration of the drug at the administration time.

As you can see, the concentration of the drug in the blood is continuously decreasing according to a decreasing exponential law, like the law of radioactive decay. Similar to the radioactive half-life, in pharmacokinetics there is a concept of **half-life** ($T_{1/2}$), which equals

$$T_{1/2} = \frac{\ln 2}{k}$$

and means the period of time during which one half of the initial mass of the drug will be eliminated from the body. This characteristic is often indicated in the instructions for medical usage of drugs.

Thus, the model allows to calculate the time dependence of drug concentration. In real medical practice, it enables to plan the intake of certain doses of the drug so that its concentration remains within the specified limits.

TOPIC 12. Multimedia technologies

12.1. Basic concepts

Multimedia technologies are technologies that allow presenting information to the user simultaneously in various forms (text, graphics, animation, sound, video) in an interactive mode. Literally, the word "multimedia" means "a lot of information carriers".

Multimedia products can be divided into several categories depending on which consumer groups they are focused on:

- computer games;
- business applications;
- educational programs;
- programs designed for creating multimedia products.

Multimedia technology includes special hardware and software. Multimedia computers should have a configuration that allows you to get a high-quality image of graphic objects, demonstrate high-resolution video, listen to sound reproduced with quality not worse than the sound quality of audio devices. Such a computer can connect many different devices:

- printer and plotter – devices for outputting text and other graphic information on paper;
- 3D printer – a device using the method of layer-by-layer creation of a physical object using a digital three-dimensional model;
- scanner – a device for inputting images from paper or slides;
- projector – to display an image (video or presentation) on a large screen;
- devices for connecting to local or global computer networks (network adapters, modems, wireless devices);
- speakers and other audio systems – for sound reproduction;
- TV tuner – for receiving television or radio broadcasts;
- joystick – a manipulator in the form of a handle with buttons fixed on hinges, which is used to control objects in computer games; other types of manipulators;
- camcorder;
- technical means for creating virtual reality;
- external storage drives.

Multimedia technologies can be used both on a separate computer (locally), for example, playing a movie from a hard disk or from a DVD-ROM, or by downloading or directly playing a stream from the Internet.

Streaming technologies present two key requirements for a computing system:

1) high data transfer rates;

2) the ability to play in real time.

High data rates are due to the nature of visual and acoustic information. The organs of vision and hearing of a person are capable of processing huge amounts of data per second; therefore, they need to deliver information at a speed that will provide an acceptable level of perception quality. For example, the data transfer speed when playing music is 128 ... 320 kbit/s, high-resolution video – 10 ... 25 Mbit/s and higher.

The second requirement imposed by multimedia applications on the system is the need for real-time data delivery. The graphic component of the video consists of a sequence of frames transmitted at a certain frequency (usually 25, 30 or 60 frames per second). Frames should be delivered at exact time intervals so that the image does not "twitch". In the case of an audio signal, the requirements are even more stringent. The sensitivity of the human ear exceeds the sensitivity of the eye, so when playing sound, the deviation in delivery time even in a few milliseconds will be noticeable.

The unevenness of the data delivery time, i.e., the scatter of the minimum and maximum data packet transit times from the average packet transit time, is called **jitter**. For example, 100 packets are sent. The minimum packet transit time is 395 ms, the average is 400 ms, and the maximum is 405 ms. In this case (405 – 400 = 5; 400 – 395 = 5), the jitter is 5 ms. To ensure high quality playback, jitter must be kept within certain limits.

The transmission medium parameters required for acceptable real-time multimedia playback are called **quality of service parameters** (QoS). They include:

1) bandwidth (data transfer rate);
2) transmission delay;
3) jitter;
4) probability of packet or bit loss.

For example, a network operator can offer a service that guarantees an average throughput of 4 Mbps, 99% of transmission delays in the interval from 105 to 110 ms (jitter in this case will be 2.5 ms) and a bit loss probability of 10^{-10}, which are acceptable parameters for transmitting a movie in MPEG-2 format.

12.2. Multimedia files

Any multimedia product includes the following main components:

- text;
- graphics;
- sound;
- video and animation;
- other types, for example, interactive three-dimensional presentations.

These components correspond to certain types and file formats. A typical multimedia product consists, as a rule, not of one file, but of a **large number of files** of various types having a certain structure.

12.3. Sound encoding and compression

Sound waves, which are mechanical waves by nature, are received by a microphone and converted into an electrical signal. Then this signal is digitized using an **analog-digital converter** (ADC), i.e., the ADC receives an electrical voltage at the input and generates a binary number at the output. The ADC samples (measures) the received electrical signal not continuously, but *discretely*. The frequency with which the ADC reads is called the **sampling rate**.

The minimum ADC sampling rate can be found using the **Nyquist-Shannon theorem** (**Kotelnikov theorem**): if the sound wave is not purely sinusoidal, but is the sum of several sinusoidal waves with frequencies from 0 to *f*, then for the subsequent

complete restoration of the signal, it is sufficient to measure the signal with the sampling frequency 2*f*. This statement was first proposed by Nyquist in 1928, and the frequency 2*f* was called the **Nyquist frequency**.

For example, to transmit sound in the entire audible range (from 16 Hz to 20000 Hz), you must use a sampling frequency of at least 40000 Hz. The most commonly used sampling rates in multimedia systems are 44100 Hz and 48000 Hz.

Digitized readings are never accurate, because the ADC can measure the input signal not exactly, but with a certain step in the signal level named the quantization step. The error resulting from inaccurate correspondence of the digitized (quantized) signal to the original signal is called **quantization noise**. With an insufficient number of bits that represent each sample of the signal, this noise can be so great that it will be heard as a distortion of the original signal or as extraneous noise.

Audio CD discs contain an audio signal digitized with a sampling frequency of 44100 Hz, as a result, they can store sounds with frequencies up to 22 kHz. Each sample is a 16-bit (2-byte) binary integer proportional to the signal amplitude.

The system **bandwidth** required for sound reproduction is calculated as the product of the sampling frequency and the number of bits in one sample. In this example, it is 705.6 kbit/s for a mono signal and 1.411 Mbit/s for a stereo one (two channels).

There are various digital sound **compression algorithms** that allow to use a significantly lower data rate (bandwidth) to transmit sound of the same quality. The most popular sound compression algorithm is the MP3 algorithm, which allows you to compress digitized sound by about 10 times.

Digital audio is easy to process on a computer. There are many applications for personal computers that allow users to record, play, edit, mix and store sound. Today, all professional sound recording and sound editing is carried out digitally.

12.4. Image encoding

The human retina has **inertial properties**, that is, a bright image that quickly appears on the retina remains on it for several milliseconds before fading away. If a sequence of identical or close images appear and disappear with a sufficiently high frequency, the human eye will not notice that it is looking at discrete images. The frequency at which the eye does not notice blinking of light (image) should be at least 50 Hz. All video systems use this principle to create moving images.

To understand how video systems function, it is best to start with a simple old-fashioned black and white television. To convert a two-dimensional image into a one-dimensional dependence of voltage on time, the camera quickly scans the image with an electron beam, breaking it into horizontal lines (scan lines) and recording the light intensity as it moves. Scanning the frame is similar to reading a book, i.e. from left to right in each line and from top to bottom in the frame. After scanning, the beam returns to its original point, and the process repeats. Different countries used different standards that describe the scanning parameters (number of scan lines, frame rate, etc.).

Modern video systems are matrix liquid crystal monitors and digital cameras with CCD sensors (charge-coupled device). These devices do not have a cathode ray tube or a scanning beam, but the principle of video signal transmission remains the same.

In order to reproduce a smooth movement of an object, the frequency of 25 frames per second is sufficient. However, at this frame rate, viewers perceive the image as blinking. This is due to the fact that the retina has time to recover before a new frame appears. To correct this drawback, it is necessary to increase the frame rate, which would require an increase in the volume of stored and transmitted information. Instead, another solution was chosen. Scan lines are displayed on the screen not in series, but first all odd lines, and then all even ones. This technique is called **interlacing**. Using this method, the image on the screen is drawn with twice the original frame rate, i.e. 50 times per second. Experiments have shown that with this sequence of displaying lines on the screen, people do not notice flickering of the image even when transmitting 25 frames per second.

In addition to interlaced, nowadays there is also **progressive scan**, i.e. the scan lines are displayed in a series (in order), and the smoothness of the image is achieved by a high frame rate. Progressive scan is used in the popular HDTV (High Definition Television) image transmission standard.

In color video, the same scanning principle is used as in black and white one, with the difference that instead of a single color, the image is represented by **three colors** – red, green and blue (RGB). The combination of these three colors is enough to reproduce any color due to the features of the human eye.

So far, we have considered analog video. Now let us discuss digital video. The simplest form of digital video presentation is a sequence of frames consisting of a rectangular grid of picture elements called pixels. A **pixel** is the smallest logical element of a two-dimensional digital image or a physical element of a display matrix forming an image. In color television, it is enough to use 8 bits for each of the three RGB colors, which gives a total of 24 bits per pixel. Although the number consisting of 24 bits can indicate about 16 million color shades, the human eye is not even able to distinguish such a huge number of color shades, not to mention larger quantities.

For smooth motion transmission, as in analog video, at least 50 frames per second must be displayed in digital video.

12.5. Image and video compression

Let us estimate the bandwidth of the communication channel necessary for transmitting an uncompressed video signal. For example, take a digital video signal containing 50 frames per second with a resolution of 1920×1080 pixels, in which each pixel contains 24 bits. Multiplying all these values, we obtain a throughput of about 2.3 Gbit/s, which is many times higher than the throughput of typical computer networks. Obviously, transmission of multimedia information in an uncompressed form is out of the question. Over the past few decades, many **compression methods** have been developed that make it possible to transmit multimedia information.

All compression methods require two algorithms: one to compress data at the source of information, and the other to recover data from its recipient. In the literature, these algorithms are called, respectively, encoding and decoding algorithms.

The **JPEG** standard is used to compress **still images** with continuously changing colors (such as photographs).

MPEG standards (MPEG-1, MPEG-2, MPEG-4) are used to compress the **video signal**. The first stage of compression is that the frames of the video signal are encoded according to the JPEG standard. Additional compression can be achieved by taking advantage of the fact that consecutive frames are often nearly identical.

In scenes in which the camera and background are stationary and a small number of objects move slowly, almost all pixels in adjacent frames will be identical. In this case, simply subtracting each frame from the previous one and processing the difference with the JPEG algorithm will give a fairly good result. However, this method is not suitable for scenes where the camera rotates or zooms into an object. For such scenes, various methods are used to compensate for this camera movement. After compression, digital video consists of frames of various types containing both full images and differential information.

12.6. Multimedia presentations

One of the most popular applications of multimedia technologies is creation of multimedia presentations.

A **multimedia presentation** is a way of presenting information using multimedia technologies. A multimedia presentation may include text materials, photographs, drawings, slide shows, sound and narration, video clips and animation, three-dimensional graphics. Multimedia presentations are created using special programs, for example, Microsoft PowerPoint, LibreOffice Impress.

Currently, multimedia presentations are most often used for presenting reports, advertising, etc., which is associated with their following advantages.

Compactness. Multimedia presentations are stored on digital media: disks, USB-drives. Such media are light in weight and small in size. At the same time, several dozens of different presentations can be stored on one medium.

Mobility. The presentation recorded on the media can be used wherever there is a computer. Another option is a laptop, with which you can make a presentation in any environment.

Volume. In one multimedia presentation, you can put a large amount of various information: audio, video, graphics, text. In a convenient form, a client, a listener can obtain a lot of information.

Visibility. Modern information technology allows to work with graphics and video at the highest level.

Effect on emotions. The combination of video and sound in a multimedia presentation allows for a comprehensive effect on emotions. A well-chosen palette, video images and musical accompaniment can affect the subconscious, affecting deep levels that are difficult to reach with words.

Cost effectiveness. Storing multimedia presentations on digital media is very inexpensive. In fact, it costs less than printing a decent booklet or book. In addition, multimedia presentations can be used repeatedly; there is no need to create new presentations for every occasion.

Self-study topics

1. Coding, classification, standardization and algorithmization of medical problems.
2. Visualization of biomedical data. Processing and analysis of medical images and biological signals.
3. Evidence-based medicine.
4. Medical hardware and software systems. Devices and systems for replacing lost human functions.
5. Nanotechnology in medicine.
6. System analysis in medical research.
7. Biological, medical and physiological cybernetics.
8. Ethical and legal principles of information management in the healthcare system.

Literature

Basic

1. Handbook of Biomedical Informatics. Електронний ресурс:
https://en.wikipedia.org/wiki/Book:Handbook_of_Biomedical_Informatics
2. David J. Lubliner. Biomedical Informatics: An Introduction to Information Systems and Software in Medicine and Health / David J. Lubliner // Auerbach Publications. – 2015. – 434 p.
3. Nanette B. Health Information Management Technology: An Applied Approach / B. Nanette // American Helath Information Management Association. – 5th ed. – 2016 – 686 p.
4. Mervat Abdelhak. Health Information: Management of a Strategic Resource / Mervat Abdelhak, Mary Alice Hanken // Saunders. – 5 edition. – 2015. – 800 p.
5. . Editors J.H. Handbook of Medical Informatics / J.H. Editors, V. Bemmel, M.A. Musen // Electronic resource:
http://www.mieur.nl/mihandbook;
http://www.mihandbook.stanford.edu
6. Mark A. Musen B. Handbook of Medical Informatics /Mark A. Musen B. // Електронний ресурс:
ftp://46.101.84.92/pdf12/handbook-of-medical-informatics.pdf
7. Edward H. Biomedical Informatics / H. Edward, J. Shortliffe, J. Cimino. – 2014 // Electronic resource:
http://www.rhc.ac.ir/Files/Download/pdf/nursingbooks/Biomedical%20Informatics%20Computer%20Applications%20in%20Health%20Care%20and%20Biomedicine-2014%20-%20CD.pdf

Additional

8. Hebda T. L. Handbook of Informatics for Nurses & Healthcare Professionals (5th Edition) / T. L. Hebda, P. Czar // Kindle Edition. – 2012. – 624 p.

9. Medical Informatics: Computer Applications in Health Care and Biomedicine. – 2011 // Електронний ресурс: https://books.google.com.ua/books?id=WYvaBwAAQBAJ&pg=PA321&lpg=PA321&dq=book++medical+informatics&source=bl&ots=VjPvStLtIk&sig=b39YVoBltS31QSJkUf4bnAjTqfY&hl=uk&sa=X&ved=0ahUKEwiqkeTdpIzQAhUGWSwKHTyIBfw4ChDoAQhHMAc#v=onepage&q=book%20%20medical%20informatics&f=false

10. Medical Informatics=Медична інформатика: підручник / І.Є. Булах, Ю.Є. Лях, В.П. Марценюк, І.Й. Хаимзон. – Київ : ВСИ "Медицина", 2012. – 368 с.

Information resources

1. http://repo.knmu.edu.ua/handle/123456789/162 (KhNMU Repository)

2. Handbook of Biomedical Informatics https://en.wikipedia.org/wiki/Book:Handbook_of_Biomedical_Informatics

3. Societies: www.amia.org, http://imia-medinfo.org/wp/, www.himss.org, www.tmi.or.th

4. U.S. Office of the National Coordinator for Health IT: http://www.healthcareitnews.com

5. Healthcare Informatics www.healthcare-informatics.com

6. Journal of the American Medical Informatics Association: www.jamia.org

7. http://www.ecdl.org/ (ECDL Foundation official site)

8. https://support.office.com/ (References and tutorials in Microsoft Office)

CONTENTS

Навчальне видання

Книгавко Володимир Гілярійович
Зайцева Ольга Василівна
Бондаренко Марина Анатоліївна
Батюк Лілія Василівна
Рукін Олексій Сергійович

МЕДИЧНА ІНФОРМАТИКА

Навчальний посібник
для іноземних англомовних студентів медичних університетів
(*англійською мовою*)

Відповідальний за випуск М. А. Бондаренко

Комп'ютерний набір О. С. Рукін
Комп'ютерна верстка О. Ю. Лавриненко

Формат А4. Ум.-друк. арк. 2,72. Зам. № 19-33856.

Редакційно-видавничий відділ
ХНМУ, пр. Науки, 4, м. Харків, 61022
izdatknmurio@gmail.com

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру видавництв, виготівників і розповсюджувачів видавничої продукції серії ДК № 3242 від 18.07.2008 р.