УДК 681.518.3 **doi: 10.26906/SUNZ.2022.4.066**

G. Golovko, D. Isai

National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine

USAGE OF IT TECHNOLOGIES IN MEDICINE AND GENOMICS

Abstract. In this article, we will consider what IT technologies are most used in medicine and by genomics methods in particular, also we will take a look at the use of big data in this matter. Additionally, we will learn what a connectome is, analyze 4M and 3V frameworks in genomics. Statistics in medicine is one of the analysis tools experimental data and clinical observations, as well as the language by means of which the obtained mathematical results are reported. However, this is not the only task of statistics in medicine. Mathematical apparatus widely used for diagnostic purposes, solving classification problems and search for new patterns, for setting new scientific hypotheses. The use of statistical programs presupposes knowledge of the basic methods and stages of statistical analysis: their sequence, necessity and sufficiency. In the proposed presentation, the main emphasis is not on detailed presentation of the formulas that make up the statistical methods, and on their essence and application rules. Finally, we talk through genome-wide association studies, methods of statistical processing of medical data and their relevance. In this article, we analyzed the basic concepts of statistics, statistical methods in medicine and data science, considered several areas in which large amounts of data are used that require modern IT technologies, including genomics, genome-wide association studies, visualization and connectome data collection.

Keywords: genomics, connectome, medicine, statistics, GWAS, data science.

Introduction

In the modern world, more and more industries are using IT tools to increase efficiency in the study of certain issues, including medicine, and since there is a lot of work with data, the most common methods are statistics and data science.

Data science is a branch of computer science that studies the problems of analyzing, processing and presenting data in digital form. Combines methods for processing data in a large volume and high level of parallelism, statistical methods, methods of data mining and artificial intelligence applications for working with data, as well as methods for designing and developing databases.

The main practical goal of professional activity in data science is the discovery of patterns in data, the extraction of knowledge from data in a generalized form. To explain the skills required for activities in this area, a Venn diagram is often used, in which the skills required by a specialist are reflected at the intersection of areas of general subject experience, practical experience in information technology (hacking skills) and knowledge mathematical statistics.

As an epistemological feature of the discipline, the priority of the practical applicability of the results, that is, the success of predictions, over their causality is indicated, while in traditional research areas it is essential to explain the nature of the phenomenon. In comparison with classical statistics, on the methods of which data science is largely based, it implies the study of super-large heterogeneous arrays of digital information and an inextricable connection with information technologies that provide their processing. In comparison with activities in the field of design and work with databases, where it is assumed that a preliminary design of a data model reflecting the relationship of the subject area and the subsequent study of the loaded data using relatively simple (arithmetic) methods, data science is supposed to rely on the apparatus of mathematical statistics, artificial intelligence, machine learning, often without first loading the data into the model. Compared to the profession of an analyst, whose main goal is to describe phenomena based on accumulated data with relatively simple user tools (like spreadsheets or Business Intelligence class tools).

Relevance of statistics in medicine. Statistics in medicine is one of the analysis tools experimental data and clinical observations, as well as the language by means of which the obtained mathematical results are reported. However, this is not the only task of statistics in medicine. Mathematical apparatus widely used for diagnostic purposes, solving classification problems and search for new patterns, for setting new scientific hypotheses. The use of statistical programs presupposes knowledge of the basic methods and stages of statistical analysis: their sequence, necessity and sufficiency. In the proposed presentation, the main emphasis is not on detailed presentation of the formulas that make up the statistical methods, and on their essence and application rules.

Statistical processing of medical research is based on the principle that what is true for a random sample is also true for a population from which this sample was obtained. However, choose or select a truly random sample from the population very difficult. Therefore, one should strive to ensure that the sample is representative of the population under study, i.e. adequately enough reflecting all possible aspects of the condition or disease under study in population, which is facilitated by a clear formulation of the goal and strict adherence to criteria for inclusion and exclusion both in the study and in the statistical analysis.

The main part of the article

Most often, the well-known "3Vs" (Volume, Velocity and Variety), which were introduced by Gartner analyst Doug Laney in 2001, are considered the main definition of big data. More precisely about the "3Vs" abbreviation:

Volume – the most obvious is where we'll start. Big data is volume. Volumes of data that can actually reach unprecedented heights. It is estimated that 2.5 quintillion bytes of data are created each day, resulting in 40 zettabytes of data being created by 2020, indicating a 300-fold increase over 2005. companies to have terabytes and even petabytes of data in storage devices and servers. This data helps shape the future of the company and its actions by tracking progress.

Velocity – the growth of data and its consequent importance has changed the way we see data. We once didn't understand the importance of data in the corporate world, but as the way we collect it has changed, we've come to rely on it day in and day out. Velocity essentially measures how fast the data arrives. Some data will come in real time, while other data will come in bursts, sent to us in bursts. And because not all platforms will process incoming data at the same speed, it's important not to generalize, discount, or jump to conclusions without having all the facts and figures.

Variety – Once upon a time, data was collected from one place and delivered in one format. Once taking the form of database files such as Excel, csv, and access, they are now presented in non-traditional forms such as video, text, PDF, and social media graphics, and through technology such as wearables. While this data is extremely useful to us, it creates more work and requires more analytical skills to decipher this incoming data, make it manageable, and let it work. Big data is much more than just "a lot of data". It's a way to empower new and existing data and open up new ways to collect future data to really make a difference for business operators and make it more agile [1].

Types of statistical data in medicine. Statistical data can be presented as quantitative (numerical continuous or discrete), and qualitative (categorical ordinal or nominal) variables. Necessary clearly indicate the type (kind) of the variable when filling the database and accurately adhere to the selected data type, as this may affect further processing of variables in many currently used statistical programs. For example, you cannot simultaneously enter into a column variable and numeric and textual, even similar in meaning, data: if filling "yes / no" in the form of 1 or 0, then do not enter alphabetic abbreviations and vice versa.

Quantitative (numerical) data suggest that the variable takes some numeric value. They are **discrete** data which can take strictly defined values, while **continuous** can be represented by any value. Unique An example of quantitative data is the representation of age by two types: in the form of a continuous variable - the exact age of the patient is indicated, and in the form discrete variable - only the number of completed years is indicated (50.3 years and 50 years; 50.9 years and 51 years).

Categoriality is the basis for the semantic understanding of qualitative variables. Categorical data is used to describe the state of an object by assigning it a number corresponding to the category to which this object belongs. An important condition for applying categorical data is belonging of one research object to only one possible category for one criterion.

Qualitative nominal data are used if the categories are not ordered. The numbers in this case are only a designation for state of an object and do not order that state. For example, by gender: 1 - male, 2 - female.

Qualitative ordinal (rank, ordinary) data - data for which categories can be ordered. For example, from feeling unwell to good: 1 - good, 2 - fair, 3 - poor. In practice often uses the translation of quantitative data into a qualitative categorical ordered presentation, especially when calculating threshold values (cut-off) for subsequent calculations of risk characteristics or predictive value with using a contingency table. For example, 1 is the concentration of the total cholesterol is less than or equal to 5.2 mmol / l (the risk ratio for developing coronary artery disease is less than 1, the predictive value of a positive result is more than 80%), 2 - concentration of total cholesterol more than 5.2 mmol/l (risk ratio of development IHD over 1, positive predictive value over 80%).

Types of statistical data analysis. In the practice of processing the results of research two types of statistical data analysis are used – primary (scheduled) and secondary (unscheduled).

Primary data analysis - used to study and describe regularities, the existence of which is assumed by the researcher, and which are the actual hypothesis of the study. In this case analyzes the features, the study of which is taken into account when planning research, and pre-formulated hypotheses are tested.

Secondary data analysis - used to form prospects for the study, search, exploration of potential patterns and hypotheses. In this case, "sifting" is performed unplanned data in a particular job, which often happens it is advisable already at the first stage of acquaintance with the data.

Genomics and Big Data. The genome of even one simple organism consists of thousands or even tens of thousands of base pairs. The manual analysis of just one DNA strand from one chromosome takes years, if not decades. Add to this the fact that sequencing often involves cutting DNA into small pieces, and we have another task - to collect the deciphered pieces in the right order. This task, called genetic mapping, is truly titanic. And although it is still impossible to do without human mental work when finalizing the data and writing the conclusion, a significant part of the analytical work is performed by the computer (Fig. 1) [4].

Computer analysis also helps with annotation - gene labeling. This process needs to be automated because most genomes are too large for manual annotation, not to mention the need to annotate as many genomes as possible since sequencing speed is no longer an issue. The annotations are made possible by the fact that genes have recognizable start and end regions (promoters and terminators that often have similar or identical composition in different groups of organisms), although the exact sequence found in these regions may vary between genes [3].

Within the framework of computational genomics, there is also such a thing as Interactome or networks of molecular interaction. In a nutshell, this is a set of localizations and interactions of a particular molecule within one particular cell. Such a model can also describe sets of indirect interactions between genes. Molecular interactions can occur between molecules belonging to different biochemical families (proteins, nucleic acids, lipids, carbohydrates, etc.), as well as within each family.

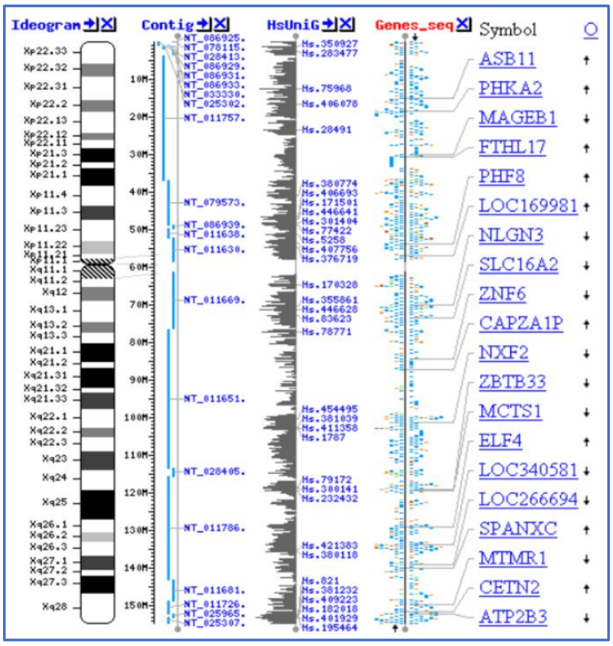

Fig. 1. Map of the human X-chromosome

When such molecules are linked by physical interactions, they form networks of molecular interactions that are usually classified according to the nature of the compounds involved. Most commonly, interactome refers to the protein-to-protein interaction network (PPI) (PIN) or variations thereof [2].

This is an extremely complex circuit, the implementation of which would not be possible without computer simulation. It is so technically complex and practically important that it has recently been singled out as an independent field of bioinformatics (Fig. 2).

Genome-wide association studies, in which hundreds of thousands of single nucleotide polymorphisms (SNPs) are strained for association with complaint in hundreds or thousands of people, have revolutionized the hunt for inheritable influences on complex traits. similar circumscriptions, unlike monogenic diseases, are caused by numerous inheritable and environmental components acting consecutively, each of which has a fairly small effect and just a many of which are absolutely critical for the onset of the complaint.

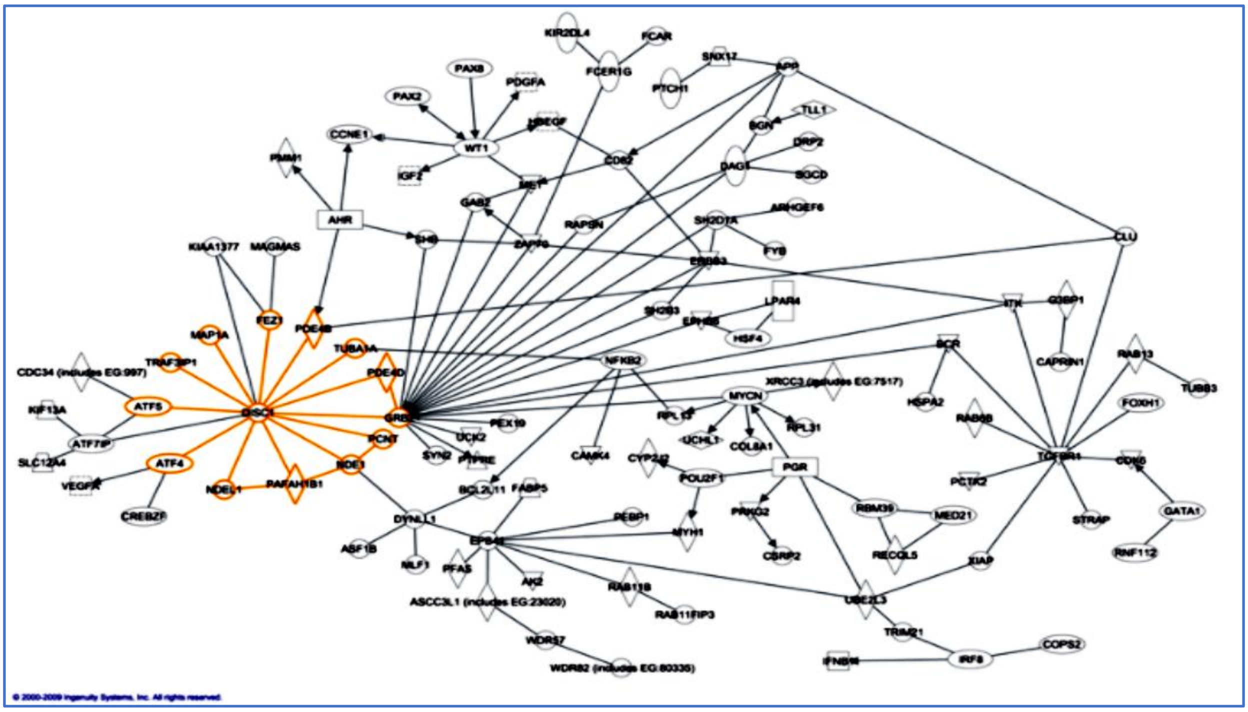

Fig. 2. Part of the interaction of DISC1 (the gene whose mutations are responsible for schizophrenia) ith genes is represented by text in boxes, and the interactions are marked by lines between genes.

Although complex conditions have been characterized as a geneticist's agony, over the past 5 years, genome-wide association studies have linked SNPs associated with hundreds of reliably replicating loci for common traits (Fig. 3) [6].

The quantum of data in these studies is four to five orders of magnitude lesser than that in the former generation of case – control studies, which tested only a sprinkle of variants, frequently in a specific seeker gene. This unknown volume poses unusual statistical cases for the analysis, demonstration, and interpretation of the data [7].

One constantly used approach to managing size is the tiered design, in which a subset of SNPs set up to be considerable in the genome-wide association study(occasionally called the discovery set) is genotyped in a alternate league(a replication set), delivering a lower subset of significantly banded SNPs that are also tested in a third league (an alternate replication set), and so on. This process helps to identify false positive associations. Carrying ahead a big number of SNPs linked through a genome-wide association study into a test of replication additionally minimizes false negative solutions while raising the bar for the establishment of true positive results.

The pooling of answers attained in genome-wide association studies (Fig. 4) under the auspices of big colleges is frequently needed for the discovery of variants with small goods on the threat of complaint. similar pooled advisements, like all genetic association studies, must be delved and administrated for alterations in allele frequency between categories that can lead to spurious (false positive) associations. The most dependable substantiation of a true genetic association, short of prescribing the unproductive variant functionally, is replication of the association, notably if it appears in multiple populations [8].

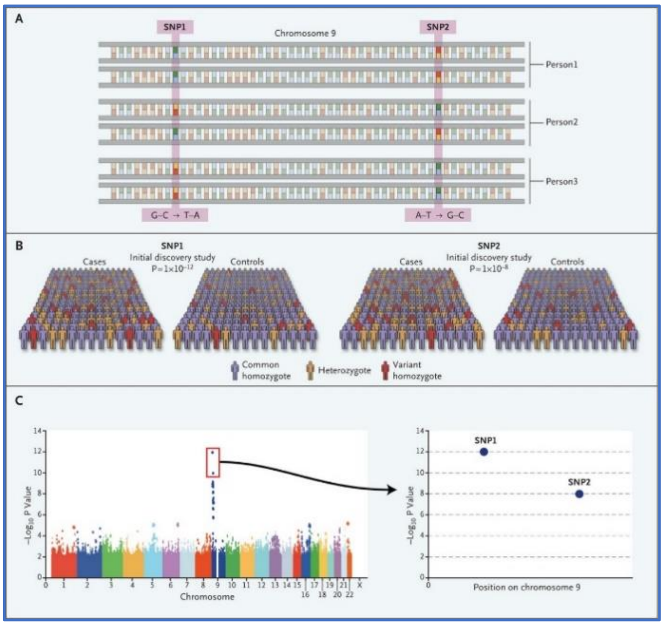

Fig. 3. The genome-wide association study

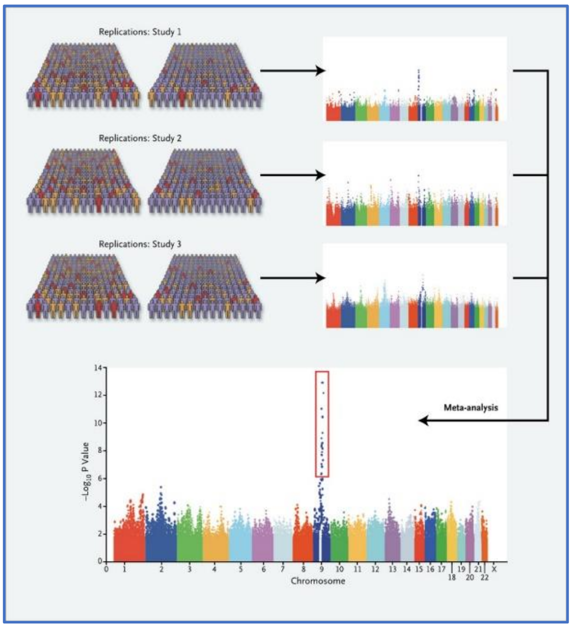

Fig. 4. Meta-analysis of genome-wide association studies

Connectome. Networks of brain connections can be emblematized at individual scales, which accord to the situations of spatial resolution in brain imaging. These situations can be roughly classified as microscale, mesoscale, and macroscale. Eventually, it'll be possible to join the results attained at different situations into a single hierarchical chart of neuronal association, which will be suitable to show a single neuron in a population of neurons up to similar large systems as cortical regions. Due to the fact that different individualities will have differences in connectomes, any unified chart is likely to give probabilistic information about the connectivity of neurons [5].

Datasets and Network Measures. Network commentaries were acquitted out for high- resolution connection matrices (n = 998 regions of interest (ROIs) with an average size of1.5 cm2), as well as for indigenous connection matrices (n = 66 anatomical subregions). All networks defended the complete cortices of both components but barred subcortical bumps and appointments. When not alluded else, the data shown in this article are grounded on the analysis of individual high- resolution connection matrices, observed by comprising across five human actors [9].

Network Visualizations. A representative illustration of a high- resolution structural connection matrix of an individual human brain is shown in Fig. 5, A. Entries of the matrix emblematize fiber consistence between dyads of single ROIs. The matrix shown in the illustration displays an aggregate of,865 symmetric connections (connection viscosity3.0). To visualize structural arrangements within this connection matrix, we uprooted the connectivity backbone, which is displayed in Fig. 5,B with a layout deduced from the Kamada-Kawai force-spring algorithm enforced in Pajek. The algorithm generates a spatial arrangement of ROIs along easily defined anterior-posterior and medium-side axes and reveals clusters of thick connectivity within posterior, carnal, and anterior cortex. Fig. 5,C shows the connectivity backbone colluded in anatomical equals. The rearward view shows groupings of largely connected clusters of ROIs arranged along the medium cortical face, extending from the precuneus via posterior and anterior cingulate cortex to the medium orbitofrontal cortex. Rearward and side views also show clusters of temporal and anterior ROIs in both components [9].

Some of the major challenges in erecting the human connectome at the micro position
moment include:

- data collection would take times with current technology,
- machine vision tools are presently in their immaturity,
- there is no proposition or algorithms to dissect the incoming data.

Conclusion

In this article, we analyzed the basic concepts of statistics, statistical methods in medicine and data science, considered several areas in which large amounts of data are used that require modern IT technologies, including genomics, genome-wide association studies, visualization and connectome data collection.

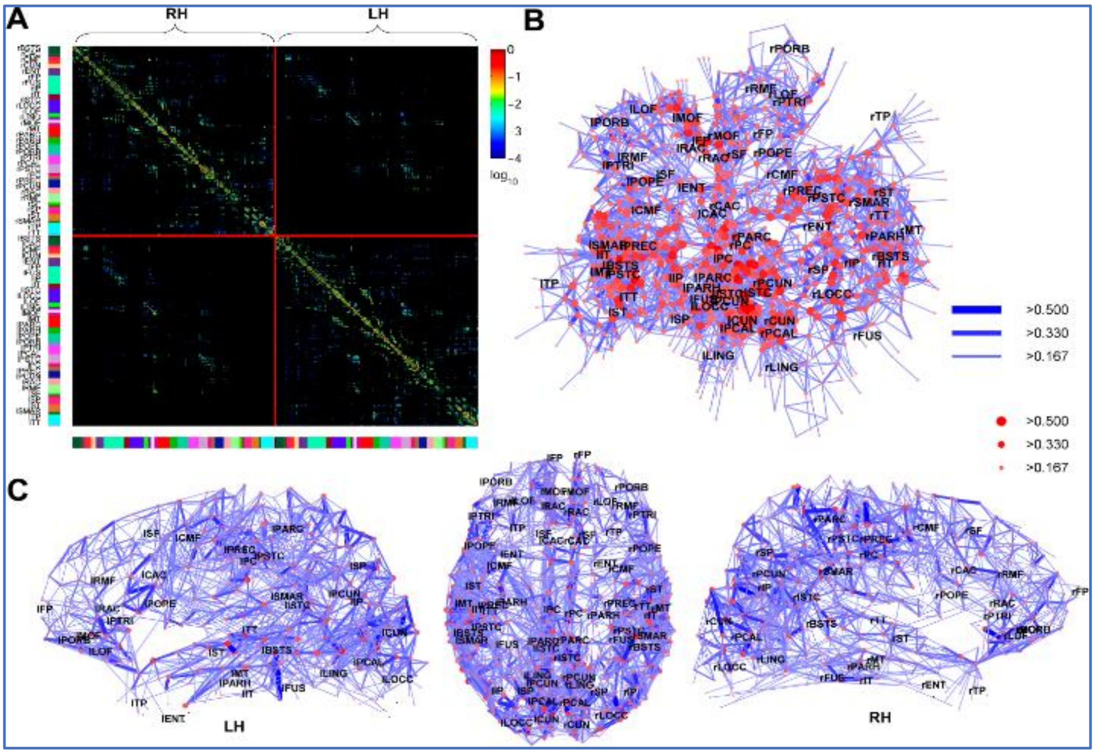

Fig. 5. High-Resolution Connection Matrix, Network Layout and Connectivity Backbone (Participant A, scan 2)

References

1. Naur, Peter. A Basic Principle of Data Science // Concise Survey of Computer Methods. - Lund, 1974. - 397 p. - (Studentlitteratur). — ISBN 91-44-07881-1.
2. William S. Cleveland. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics // International Statistical Review: Journal. - Willey & Sons, 2001. - Vol. 69, issue 1. - P. 21-26. — ISSN 1751-5823.
3. Hey T, Trefethen A. The data deluge: an e-science perspective. In: Berman F, Fox G, Hey T, editors. Grid computing: making the global infrastructure a reality. Chichester: Wiley-Blackwell; 2003. p. 809–24.
4. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M. Interrelating different types of genomic data, from proteome to secretome: ‘oming in on function. Genome Res.2001;11:1463–8.
5. Sebastian Seung, Connectome: How the Brain’s Wiring Makes Us Who We Are, 2012 — ISBN 978-0547508184
6. Manolio TA (July 2010). "Genomewide association studies and assessment of the risk of disease". The New England Journal of Medicine. 363 (2): 166–76.
7. Smith SM, Douaud G, Chen W, Hanayik T, Alfaro-Almagro F, Sharp K, Elliott LT (2021). "An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank"
8. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M (May 2010). "Genome-wide association studies in diverse populations". Nature Reviews Genetics. 11 (5): 356–66.
9. Hagmann P., Cammoun L., Gigandet X., Meuli R., Honey C. J., Wedeen V. J., Sporns O. Mapping the structural core of human cerebral cortex // PLoS Biol.: journal. — 2008. — July (vol. 6, no. 7). — P. e159.

Received (Надійшла) 12.09.2022
Accepted for publication (Прийнята до друку) 16.11.2022

Використання IT-технологій в медицині та геноміці

Г. В. Головко, Д. А. Ісай

Анотація. У цій статті ми розглянемо, які IT-технології найбільше використовуються в медицині і методами геноміки зокрема, а також розглянемо використання big data в цьому питанні. Додатково ми дізнаємося, що таке коннектом, розберемо фреймворки 4M і 3V в геноміці. Статистика в медицині є одним із інструментів аналізу експериментальних даних і клінічних спостережень, а також мовою, за допомогою якої повідомляються отримані математичні результати. Однак це не єдине завдання статистики в медицині. Математичний апарат широко використовується для діагностичних цілей, вирішення задач класифікації та пошуку нових закономірностей, для постановки нових наукових гіпотез. Використання статистичних програм передбачає знання основних методів і етапів статистичного аналізу: їх послідовності, необхідності та достатності. У запропонованій презентації основний акцент робиться не на детальному викладі формул, з яких складаються статистичні методи, а на їх сутності та правилах застосування. Нарешті, ми говоримо про дослідження асоціацій у всьому геномі, методи статистичної обробки медичних даних та їх актуальність. У цій статті ми проаналізували основні поняття статистики, статистичні методи в медицині та науку про дані, розглянули кілька сфер, у яких використовуються великі обсяги даних, які вимагають сучасних IT-технологій, включаючи геноміку, дослідження загальногеномних асоціацій, візуалізацію та дані коннектомів.

Ключові слова: геноміка, коннектом, медицина, статистика, GWAS, наука про дані.