

Nowoczesne Systemy Zarządzania
Zeszyt 14 (2019), nr 3 (lipiec-wrzesień)
ISSN 1896-9380, s. 99-111

Modern Management Systems
Volume 14 (2019), No. 3 (July-September)
ISSN 1896-9380, pp. 99-111



Instytut Organizacji i Zarządzania
Wydział Bezpieczeństwa, Logistyki i Zarządzania
Wojskowa Akademia Techniczna
w Warszawie

Institute of Organization and Management
Faculty of Security, Logistics and Management
Military University of Technology

Healthcare data management, regarding the growing popularity of precision medicine

Zarządzanie danymi w systemie ochrony zdrowia uwzględniające popularność medycyny precyzyjnej

Katarzyna Maciejewska

Military University of Technology
Institute of Computer and Information System

Abstract. Precision medicine that is studying DNA-based diseases is gaining more and more popularity in the world thanks to the dynamic development of sequencing technology. In 2003, the Human Genome Project (HGP, Human Genome Project) was completed, lasting 15 years and costing 2.7 billion dollars. Currently, the cost of running a genome test is around USD 1,000 and lasts only a few hours to several days, depending on the technology chosen. This document contains the most important concepts related to precision medicine and healthcare data management. Due to the significant reduction in costs associated with carrying out genetic tests, they have become more available and can be taken into account in patient studies. The techniques and methods discussed generate a large amount of medical data that should be managed and also used in preventive healthcare.

Keywords: DNA, EHR, ICD, SNOMED CT.

Abstrakt. Medycyna precyzyjna, która zajmuje się badaniem genetyki chorób na podstawie DNA, zyskuje coraz większą popularność na świecie za sprawą dynamicznego rozwoju technologii sekwencjonowania materiału genetycznego. W roku 2003 został ukończony projekt poznania ludzkiego genomu (ang. Human Genome Project, HGP), który trwał 15 lat i który kosztował 2,7 miliarda dolarów. Obecnie koszt przeprowadzenia badania genomu wynosi około 1000 dolarów i trwa zaledwie od kilku godzin do kilku dni w zależności od wybranej technologii. W niniejszej pracy zostały opisane wybrane pojęcia związane z medycyną precyzyjną i zarządzaniem danymi w systemach ochrony zdrowia. Ze względu na znaczne obniżenie kosztów związanych z przeprowadzaniem testów genetycznych stały się one bardziej dostępne i mogą być brane pod uwagę w procesie diagnozowania i leczenia pacjentów. Omawiane techniki i metody generują dużą ilość danych medycznych, które powinny być zarządzane i wykorzystywane również w profilaktycznej opiece zdrowotnej.

Słowa kluczowe: DNA, EHR, ICD, SNOMED CT.

Introduction

In 2015, President Barack Obama created a research project called the Precision Medicine initiative, the goal was to make advances in tailoring medical care to the individual (Whitehouse. gov, 2015). Precision medicine in particular consists in improving the treatment of patients through genetic tests and molecular profiling. It focused on identifying which approaches will be effective for which patients based on genetic, environmental, and lifestyle factors (U.S. National Library of Medicine, 2018). Modern medicine does not have effective treatment methods that focus on the true etiology of the disease and not its consequences (Loscalzo, Antman, 2016).

Worldwide can be a rise of interest in genomics, bioinformatics and precision medicine observed. Continuous improvement of DNA sequencing technology, which contributed to the rapid reduction of costs shown in figure 1, allows for intensive research in this field, which also greatly increases the amount of data produced and processed. This leads to many challenges in terms of volumetric, data quality, safety, interoperability, terminology and knowledge extraction.

So far current electronic health record (EHR) systems allow to store a range of data, but they weren't designed to store the information about the patients genome and the identified mutations in it. A solution is needed to support doctors in treating patients, that will combine medical historical data with information detected in DNA and that could be used in preventive healthcare.

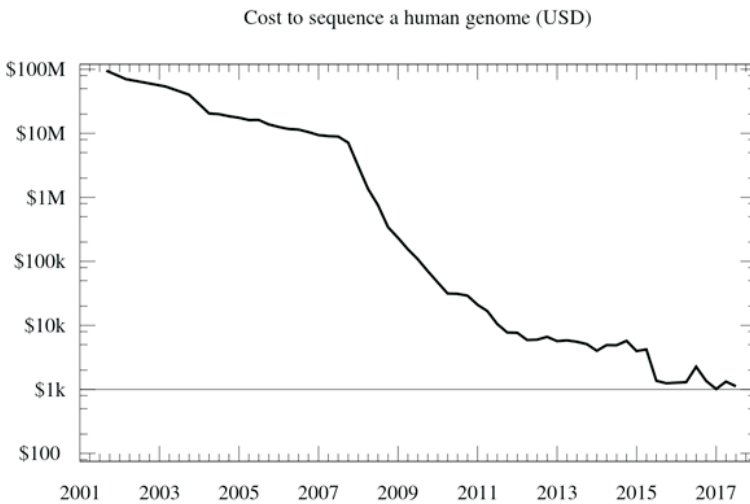


Fig. 1. DNA sequencing cost per genome
Source: elaboration based on (Wikipedia 1)

Objective of research of this article is to present the practical recommendations of using SNOMED CT as a linkage between EHR data and DNA data.

Tasks:

- To analyze the available medical data formats and standardizations in Poland and worldwide;
- To examine the possibilities of combining several data format in the healthcare system.

Research methods – analysis and synthesis of documents and scientific literature, comparative analysis, analysis of data systemization, logical analysis and synthesis, grouping.

1. Information in the genome

A genome is the genetic material of an organism, by humans it is the complete DNA sequence encoded from the 23 chromosome pairs retrieved from the cells as shown on figure 2. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Adenine always pairs with thymine, and guanine always pairs with cytosine. A human DNA sequence is estimated to be about 3.2 billion bases long.

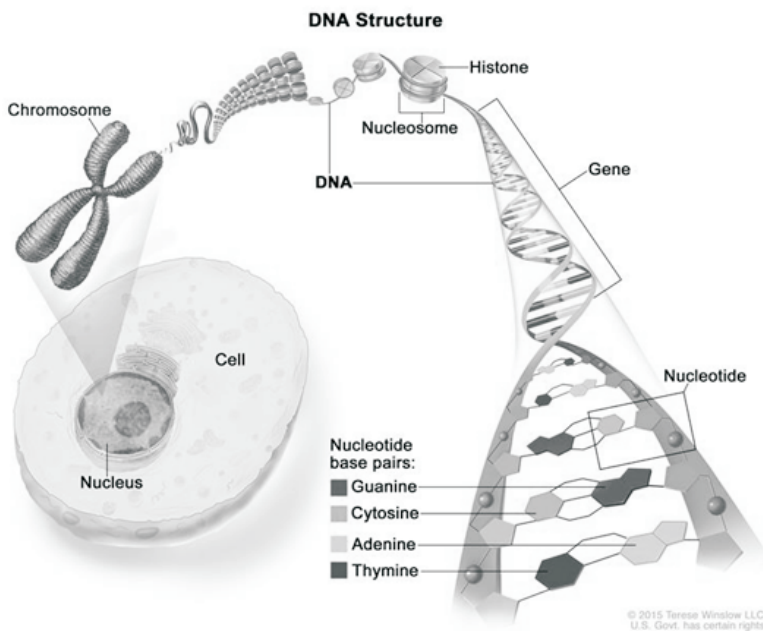


Fig. 2. DNA Structure

Source: elaboration based on (NCI Dictionaries, 2019)

Two percent of DNA consists of genes – coding sections that carry the instructions to make proteins and 98% non-coding sections. Noncoding DNA contains sequences that act as regulatory elements, determining when and where genes are turned on and off. A mutation in noncoding DNA can cause a protein to be expressed in the wrong place or at the wrong time or can reduce or eliminate expression of an important protein when it is needed (U.S. National Library of Medicine, 2018). Not all mutations in DNA have an impact on health, but those that are responsible for a change in the expression of a protein can cause a health problem. Genetic disorders can be divided in two groups – Mendelian and complex disorders.

Mendelian disorders may be caused by changes in a single gene, as in the case of cystic fibrosis. These disorders have recognizable inheritance patterns and are rare in the population. In complex disorders, genetic variation in a single gene is seen as a risk of developing the disease, but it is not treated as a sufficient condition and other factors are taken into account (Medicine Encyclopedia, 2002).

Thanks to the Human Genome Project's effort, a reference sequence has been published, which anyone can use for free to compare with own analyzed DNA sequences and find the differences.

Observable features or traits such as its morphology, development, biochemical or physiological properties are called phenotypes (Miko, LeJeune, 2014). The phenotype results from the expression of the genetic code of the organism.

These studies are therefore interesting because they can be used to diagnose patients more quickly, to better select drugs and their doses, which leads to a reduction in treatment costs.

2. Electronic health record – EHR

The healthcare industry has been constantly changing for several years, one of which is the transition from paper-based documentation to digital. The first step was to build IT systems that stored electronic medical record (EMR). An EMR is a digital version of the paper chart in the clinician's office. It contains the medical and treatment history of the patients in one practice.

The second generation of IT solutions are systems that allows to store electronic health records. EHRs are built to share information with other health care providers, such as laboratories and specialists, so they contain information from all the clinicians involved in the patient's care. The systems store demographics data, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information (Top Mobile Trends, 2014).

To achieve interoperability, data must be standardized so it can be effectively shared between various systems. Classification systems and reference terminologies, such as ICD-10 or SNOMED CT, provide a uniform way of collecting and maintaining patient data.

3. Terminology and standards

In Poland for the implementation of standards and classification is the National Centre for Health Information Systems (CSIOZ) responsible, this is a unit of the Ministry of Health. For improving the exchange of data the unit released the Polish National Implementation of HL7 CDA (Clinical Document Architecture). CDA is an XML-based markup standard intended to specify the encoding, structure and semantics of clinical documents. Templates, the CDA XSD schema and clinical document visualization using XSLT can be downloaded on the unit's website <https://www.csioz.gov.pl/HL7ENG/pl-cda-html-en-US/index.html>. At the moment the current edition uses the ICD 10 classification. In the Polish healthcare system, the ICD classification is obligatory, it allows grouping of diseases and is useful for statistical purposes, but has also limitations in the use in EHR systems, in addition, Genomics has its own standards and formats that do not correspond to standard EHR systems.

3.1. ICD-10

ICD stands for international classification system for surgical, diagnostic and therapeutic procedures. It is a standardization for the names of diseases and their basic systematization in a hierarchical structure. For the term myocardial infarct are 21 codes in the classification present for example:

- I21 – Acute myocardial infarction;
- I21.9 – Acute myocardial infarction, unspecified;
- I22 – Subsequent myocardial infarction.

3.2. Human Phenotype Ontology

HPO is a structured and controlled vocabulary for the phenotypic features encountered in human hereditary and other disease. This is a flagship product of the Monarch Initiative, an NIH-supported international consortium dedicated to semantic integration of biomedical and model organism data with the ultimate goal of improving biomedical research. An example for the entry – myocardial infarction is shown at figure 3.

Myocardial infarction

http://purl.obolibrary.org/obo/HP_0001658

Necrosis of the myocardium caused by an obstruction of the blood supply to the heart and often associated with chest pain, shortness of breath, palpitations, and anxiety as well as characteristic EKG findings and elevation of serum markers including creatine kinase-MB fraction and troponin. [HPO:probinson]

Synonyms: MI, Heart attack

Tree view | Term history

- All
 - Phenotypic abnormality
 - Abnormality of the cardiovascular system
 - Abnormality of cardiovascular system physiology
 - Myocardial infarction**
 - abnormal phenotype
 - Abnormality of the cardiovascular system
 - Abnormality of cardiovascular system physiology
 - Myocardial infarction**
- Phenotype
 - Phenotypic abnormality
 - Abnormality of the cardiovascular system
 - Abnormality of cardiovascular system physiology
 - Myocardial infarction**
 - abnormal phenotype
 - Abnormality of the cardiovascular system
 - Abnormality of cardiovascular system physiology
 - Myocardial infarction**

Graph view

Reset tree

Show all siblings

Term info

database cross reference

- MSH:D009203
- SNOMEDCT_US:22298006
- UMLS:C0027051

layperson term

Heart attack [ORCID:0000-0001-5208-3432]

abbreviation

MI

has obo namespace

human_phenotype

id

HP:0001658

Term relations

Subclass of:

- + Abnormality of cardiovascular system physiology

Fig. 3. HPO example for Myocardial infarction

Source: elaboration based on (OLS Ontology, 2019)

3.3. Snomed CT

Snomed – Systematized Nomenclature of Medicine, provides a set of concepts and relationships that offers a common reference point for comparison and aggregation of data about the health care process. It is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT can be used to directly record clinical details of individuals in electronic patient records. It also provides the user with a number of linkages to clinical care pathways, shared care plans and other knowledge resources, in order to facilitate informed decision-making, and to support long-term patient care. It can be cross-mapped to other international standards and classifications like ICD, HPO etc. SNOMED CT provides a compositional syntax that can be used to create expressions that represent clinical ideas which are not explicitly represented by SNOMED CT concepts.

An example for the possible term representation in SNOMED CT for myocardial infarction:

- **Pre-coordinated Expression:**
22298006 |Myocardial infarction (disorder)

- **Expression from Stated Concept Definition:**

57809008 |Myocardial disease (disorder)| :
 { 116676008 |Associated morphology (attribute)| = 55641003 |Infarct (morphologic abnormality)|, 363698007 |Finding site (attribute)| = 74281007 |Myocardium structure (body structure)| }

- **Expression from Inferred Concept Definition:**

57809008 |Myocardial disease (disorder)| +
 – 414545008 |Ischemic heart disease (disorder)| +
 – 251061000 |Myocardial necrosis (finding)| +
 – 609410002 |Necrosis of anatomical site (disorder)| :
 { 116676008 |Associated morphology (attribute)| = 55641003 |Infarct (morphologic abnormality)|,
 363698007 |Finding site (attribute)| = 74281007 |Myocardium structure (body structure)| }

- **Expression diagram** shown on figure 4.

SNOMED CT as a flexible nomenclature, which could be adapted to any medical specialty, also allows storing DNA and clinical data of patients in the EHR in a format useful for decision support or research. At SNOMED EXPO 2016, the example shown in figure 5 illustrates how the DNA information could be represented.

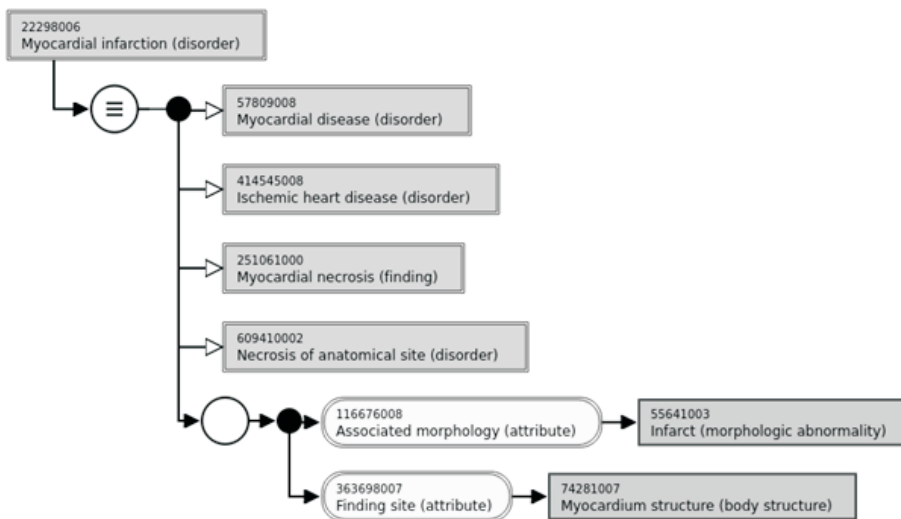


Fig. 4. Snomed CT diagram for myocardial infarction

Source: elaboration based on (IHTSDO Browser, 2019)

➤ Body structures>>Cell structures>>Nucleotide sequences and Genes

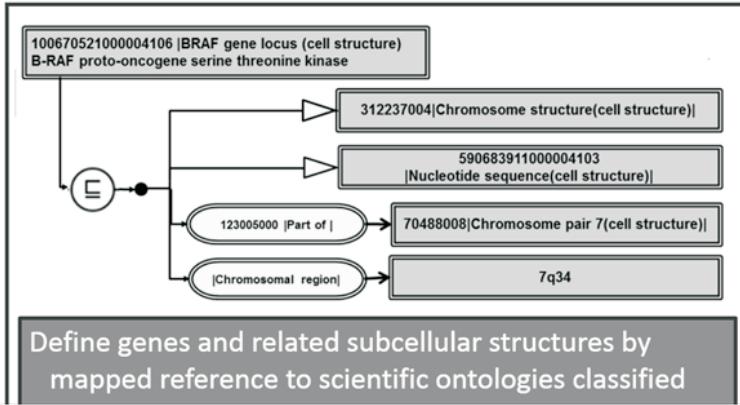


Fig. 5 DNA representation example

Source: elaboration based on (Campbell, Campbell, Pedersen, 2016)

4. Sequencing technology

Whole genome sequencing (WGS) and whole exome sequencing (WES) are high-throughput sequencing techniques also called next generation sequencing (NGS). As the names themselves indicate, the WGS method is used for sequencing the whole genome of the patient, and the WES method for sequencing only the coding sections in the DNA. In 1960, the term “junk DNA” was used to describe non-coding DNA regions (Starling, 2013), but scientists also discovered that some sequences in junk DNA act as switches that determine where and when genes are expressed what is important in studying complex diseases.

Humans have about 180,000 exons, which is about 1-2% of the human genome. Until now, the WES technique was much cheaper than WGS, therefore it was preferred for research. Exome sequencing is useful to discover the gene for rare, Mendelian disorder of unknown cause (Ng, Buckingham, Lee et al., 2010).

Next generation sequencers produce DNA reads as an output file called FASTQ (Cock, Fields, Goto et al., 2010).

4.1. FASTQ

FASTQ is a common file format for sharing sequencing read data combining both the sequence and an associated per base quality score. Both the sequence letter

and quality score are each encoded with a single ASCII character. A FASTQ file uses four lines per sequence:

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
- Line 2 raw sequence letters.
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

A FASTQ file example:

```
@SEQ_ID
GATTTTGGGGCCCCAAAGCAGTATCGATCAAATAGTAAATCC
ATTTGTTCAA
+
!*(((***+))%%%+)(%%%)1***-+**))*55CCF>>>>>>CCCCCCC
```

The character '!' represents the lowest quality while '~' is the highest. The quality value characters in left-to-right increasing order of quality.

4.2. SAM/BAM (Samtools 1, 2018)

SAM stands for Sequence Alignment/Map format. It is a TAB delimited text format consisting of a header section, which is optional, and an alignment section. It was developed by Heng Li and Bob Handsake (Li, Handsaker, Wysoker et al., 2009) for storing biological sequences aligned to a reference sequence. The format supports short and long reads (up to 128 Mbp). NGS is a sequencing technique that produce many reads for the same regions, because of that after retrieving a FASTQ file an alignment and mapping need to be done. An example for a SAM file is shown on figure 6.

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1   TTAGATAAAGGATA*CTG
+r002     aaaAGATAA*GGATA
+r003     gcctaAGCTAA
+r004     ATAGCT.....TCAGC
-r003     ttagctTAGGC
-r001/2   CAGCGGCAT
```

Fig. 6. SAM file example
Source: SAM specifications

4.3. VCF

VCF (Samtools) stands for Variant Call Format it is a text file format (most likely stored in a compressed manner). It contains meta-information lines (prefixed with “##”), a header line (prefixed with “#”), and data lines each containing information about a position in the genome and genotype information on samples for each position. Variant calling is the process by which we identify variants from sequence data. An example from the VCF documentation is shown at figure 7. VCF is a preferred format. Many millions of variants can be stored in a single VCF file.

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f3794618ff66beb2da,species="Homo sapiens",taxonomy=>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 C A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0/0:48:1:51,51 1/0:48:8:51,51 1/1:43:5:
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0/0:49:3:58,50 0/1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1/2:21:6:23,27 2/1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0/0:54:7:56,60 0/0:48:4:51,51 0/0:61:2
20 1234567 microsati1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

Fig. 7. VCF file example
Source: VCF specification

5. Methodology – genome wide association studies (GWAS)

Genome wide association studies (GWAS) are hypothesis free methods to identify associations between genetic regions and traits (including diseases). GWAS study collects data to find out the common variants in a number of individuals, both with and without a common trait. Variants associated with the disease, or within the same haplotype as a variant associated with a disease, will be found at a higher frequency in cases than controls. Statistical analysis is carried out to indicate how likely a variant is to be associated with a trait. Results are subject to statistical analyses to assign a p-value to each variant. A p-value indicates the significance of the difference in frequency of the allele tested between cases and controls – the probability that the allele is likely to be associated with the trait.

For whole genome or exome variant calling process three step should be made:

- Retrieve the whole genome or whole exome sequencing to create FASTQ files.
- Align the sequences to a reference genome, creating SAM/BAM files.

- Identify where the aligned reads differ from the reference genome and write to a VCF file.

At this point, topics such as precision medicine, bioinformatics and genetics are very popular. Research centers and companies outdo each other in publishing new discoveries.

In 2015, an associative analysis was conducted for variants occurring in the DNA of patients who have had a myocardial infarct. They studied 406 patients hospitalized in the research institute for complex issues of cardiovascular disease in Russia. The average of the patients age was 60 years. The control group without the disease was 183. The frequencies of genotype between the two groups were compared by using Pearson X^2 test (Golubenko, Slakhov, Makeeva et al., 2015). This study revealed similarities in the DNA by patients with a myocardial infarct history and found also potential risk patients but the method for this study wasn't based on WGS reads.

A study from 2017 of genomic variants for Alzheimer's disease (Butkiewicz, Blue, Leung et al., 2018) were based on the same methods selected for this study. Patients were selected by phenotypes for the project A total number of 578 individuals from 111 families were whole genome sequenced and 10913 individuals were whole exome sequenced. The WGS method discovered 5 million novel variants the WES 1 586703 variants.

Another study from 2017 with WGS method was performed for precision medicine screening to identify disease risk in adults (Perkins, Caskey, Brar, 2018). The research showed that by using only the WGS method possible age related chronic diseases were identified by 19% patients from the study. By adding to this results also advanced imaging and other medical testing methods this number increased to 26,27%. These types of research in the future will allow rapid identification of many diseases using appropriate screening tests. but to make this possible, we have to develop systems that will allow to combine and manage all these data.

6. Genomic IT solutions

The largest companies and research centers on the market are constantly working on software related to DNA data analysis. For example, there may be mentioned:

- EHR-Phenolyzer (Son, Xie, Yuan et al., 2018) – is a high-throughput EHR framework for extracting and analyzing phenotypes. EHR-Phenolyzer extracts and normalizes Human Phenotype Ontology (HPO) concepts from EHR and then prioritizes genes with causal variants on the basis of the HPO-coded phenotype manifestations.
- Google released a tool called DeepVariant that uses the latest AI techniques to build a more accurate picture of a person's genome from sequencing data.
- Intels Genome Analysis Tool Kit (GATK) is a software package developed at the Broad Institute to analyze next-generation sequencing data.

- Philips and Illumina announced in 2017 a collaboration. Illumina’s DNA sequencing technology and BaseSpace® Sequence Hub with Philips’ innovative cloud-based genomics platform for the acquisition, analysis and interpretation of genomics data in cancer research.

Companies put a lot of effort to improve the quality of DNA data readings and analyses. They introduce standardization and appropriate classifications and nomenclatures to improve the knowledge retrieval of the generated data. They are working on new algorithms to compare large sets of DNA-specific data, some based on artificial intelligence.

Conclusion and practical recommendation

Integration and analysis of all these data is not easy, however, to achieve the best results for the patient and the health care system, solutions must be developed. A possible solution would be to use the SNOMED nomenclature as a data link from EHR to genomic shown on figure 8.

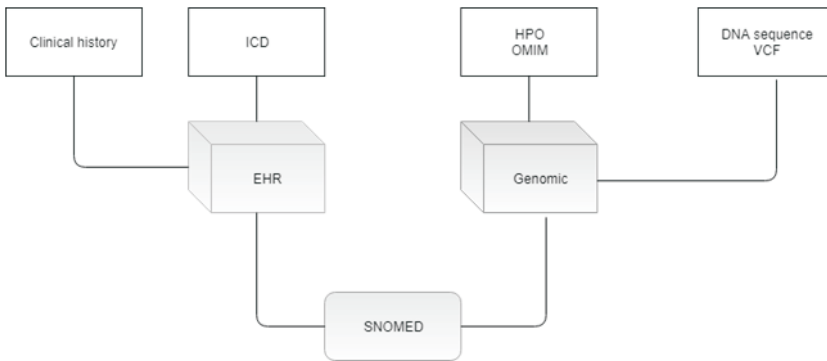


Fig. 8. Prototype integration model
Source: own elaboration

The next step would be to check. whether using the nomenclature would allow the use of AI tools to build advanced prediction systems.

BIBLIOGRAPHY

[1] BUTKIEWICZ M., BLUE E.E., LEUNG Y.Y. et al, 2018, *Functional Annotation of Genomic Variants in Studies of Late-Onset Alzheimer’s Disease*, “Bioinformatics”, Vol. 34, No. 16.

[2] CAMPBELL J.R., CAMPBELL W.S., PEDERSEN J., 2016, *SNOMED CT and Clinical Genomics: Use Case for Precision Medicine*, *SNOMED CT Expo 2016*, Departments of Internal Medicine and Pathology University of Nebraska Medical Center, Omaha.

- [3] CHIAL H., 2008, *DNA Sequencing Technologies Key to the Human Genome Project*, “Nature Education”, Vol. 1, No. 1.
- [4] COCK P.J.A., FIELDS CH.J., GOTO N. et al., 2010, *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants*, “Nucleic Acids Research”, Vol. 38, No. 6.
- [5] Golubenko M.V., Slakhov R.R., Makeeva O.A. et al., 2015, *Association of Mitochondrial DNA Polymorphism with Myocardial Infarction and Prognostic Signs for Atherosclerosis*, “Molecular Biology”, Vol. 49, No. 6.
- [6] LI H., HANDSAKER B., WYSOKER A. et al., 2009, *The Sequence Alignment/Map Format and SAM-tools*, “Bioinformatics”, Vol. 25, No. 16.
- [7] LOSCALZO J., ANTMAN E.M., 2016, *Precision Medicine in Cardiology*, “Nature Reviews Cardiology”, Vol. 13, No. 10.
- [8] MIKO I., LEJEUNE L. (eds.), 2014, *Essentials of Genetics*, NPG Education, Cambridge.
- [9] NG S.B., BUCKINGHAM K.J., LEE CH. et al., 2010, *Exome Sequencing Identifies the Cause of a Mendelian Disorder*, “Nature Genetics”, Vol. 42, No. 1.
- [10] PERKINS B.A., CASKEY C.T., BRAR P. et al., 2018, *Precision Medicine Screening Using Whole-Genome Sequencing and Advanced Imaging to Identify Disease Risk in Adults*, “PNAS”, Vol. 115, No. 14.
- [11] SON J.H., XIE G., YUAN CH. et al., 2018, *Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes*, “American Journal of Human Genetics”, Vol. 105, No. 1.

INTERNET SOURCES

- [1] IHTSDO Browser, 2019 (access: 30.03.2019), <https://browser.ihtsdotools.org/>.
- [2] MEDICINE ENCYCLOPEDIA, 2002, *Genetics of Disease: Mendelian and Complex Disorders* (access: 30.03.2019), <http://medicine.jrank.org/pages/2134/Disease-Genetics-Mendelian-Complex-Disorders>.
- [3] NCI Dictionaries, 2019, *Nucleotide* (access: 30.03.2019), <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/nucleotide>.
- [4] OLS Ontology, 2019, *Myocardial infarction* (access: 30.03.2019), <https://www.ebi.ac.uk/ols/index>.
- [5] SAMTOOLS 1, 2018, *Sequence Alignment/Map Format Specification* (access: 30.03.2019), <https://samtools.github.io/hts-specs/SAMv1.pdf>.
- [6] SAMTOOLS 2, 2018, *The Variant Call Format Specification* (access: 25.03.2019), <https://samtools.github.io/hts-specs/VCFv4.3.pdf>.
- [7] STARLING J., 2013, *The Origin of the Term “Junk DNA”: A Historical Whodunnit* (access: 25.03.2019), <http://judgestarling.tumblr.com/post/64504735261/the-origin-of-the-term-junk-dna-a-historical>.
- [8] TOP MOBILE TRENDS, 2014, *Mobile Tech Contributions to Healthcare and Patient Experience* (access: 25.03.2019), <https://web.archive.org/web/20140530024928/http://topmobiletrends.com/mobile-technology-contributions-patient-experience-parmar/>.
- [9] U.S. NATIONAL LIBRARY OF MEDICINE, 2018 (access: 25.03.2019), <https://ghr.nlm.nih.gov/primer/precisionmedicine/precisionvpersonalized>.
- [10] WHITEHOUSE.GOV, 2015, *FACT SHEET: President Obama’s Precision Medicine Initiative*. (access: 25.03.2019), <https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative>.
- [11] Wikipedia 1, *Whole Genome Sequencing* (access: 25.03.2019), https://en.wikipedia.org/wiki/Whole_genome_sequencing.