pHealth 2020
B. Blobel et al. (Eds.)

doi:10.3233/SHTI200627

Representation and Interpretation of Genetic Analysis: A Strategy of Development for the Personal Genetics Card Information System

Michal HUPTYCH[a1] and Lenka LHOTSKÁ[a,b]
[a] *Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Czech Republic*
[b] *Faculty of Biomedical Engineering, Czech Technical University in Prague, Czech Republic*

Abstract. In this paper, we describe a strategy for the development of a genetic analysis comprehensive representation. The primary intention is to ensure the available utilization of genetic analysis results in clinical practice. The system is called Personnel Genetic Card (PGC), and it is developed in cooperation of CIIRC CTU in Prague and the Mediware company. Nowadays, genetic information is more and more part of medicine and life quality services (e.g. nutritional consulting). Therefore, there is necessary to bind genetic information with the clinical phenotype, such as drug metabolism or intolerance to various substances. We proposed a structured form of the record, where we utilize the LOINC® standard to identify genetic test parameters, and several terminology databases for representing specific genetic information (e.g. HGNC, NCBI RefSeq, NCBI dbNSP, HGVS). Further, there are also several knowledge databases (PharmGKB, SNPedia, ClinVar) that collect interpretation for genetic analysis results. In the results of this paper, we describe our idea in the structure and process perspective. The structural perspective includes the representation of the analysis record and its binding with the interpretations. The process perspective describes roles and activities within the PGC system use.

Keywords. terminology, standards, eHealth, information system, genetic analysis

Introduction

In the last decade, a genetic analysis becomes a more common process within standard health care as well as in nutrition consultancy and sports medicine. From the clinical perspective, it is important the connection between the genetic profile of a person and reaction on the medication, intolerance of some substances, and predisposition for diseases. There are a couple of specific conditions for the genetics area. First, the interpretations are based only on empirical evidence. Thus, there is a chance that the interpretation will be changed in future. Second, the issue of security and privacy is extremely important, more than other medical data cases. This means that the system

[1] Corresponding Author. Michal Huptych, Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Czech Republic; Email: Michal.Huptych@cvut.cz

must use not only pseudo anonymization but full encrypted data (according to the best practice).

Thus, from our perspective, there are two crucial restrictive aspects for the use of genetic information: (1) the analysis must have a substantial benefit for healthcare or life quality; (2) there must be structured and powerfully secured storage and manipulation with the genetic information. Therefore, the development strategy of our project Personal Genetic Card (PGC) is based on determination of existing terminologies and standards with the unambiguous connection of the personal genetic information to the valid clinical interpretation with the highest possible level of evidence.

We partially use the recommendation of the Clinical Pharmacogenetics Implementation Consortium (CPIC®, https://cpicpgx.org/) recommendations [1].

1. Terminology for the Genetics Analysis Representation

First, we determined terminologies for the representation of the genetic analysis. For parameters identification, we decided to use Logical Observation Identifiers Names and Codes (LOINC®) [2,3]. There is also a Systematized Nomenclature of Medicine (SNOMED, www.snomed.org) [4]. However, the LOINC's primary aim is the representation of laboratory and clinical results, and it is satisfactorily comprehensive and structured system for our purpose. Each LOINC component is defined by unique code, component's name and a set of five component's attributes (property, time aspect, sample/system for test, the scale of the result representation, and used method of the analysis). Moreover, a lot of components include inner defined value-set (called answers) or the recommendation of the value-set format or terminology database. Thus, the LOINC helps to determine terminology for genetic analysis values representation. We determined as the most important terminology following databases:

- The HUGO Gene Nomenclature Committee (HGNC) [5]. The database contains the definition of codes and names for particular genes – e.g. HGNC:5 – A1BG or HGNC:30005 – A3GALT2. HGNC has the web page https://www.genenames.org, and there is possible to obtain the data through REST API web-service, which offers the data in JSON or XML formats or there is the download of the file with the data in TXT or JSON formats.
- The NCBI Reference Sequence (RefSeq) Database [6] is a non-redundant, comprehensive, commented database that contains files of sequences, included DNA transcriptions and proteins. On the contrary to the HGNC (that identify gene), the NCBI RefSeq represents coding of the more areas of the specific gene. The RefSeq data is available at the FTP ftp://ftp.ncbi.nlm.nih.gov/refseq.
- The NCBI Single Nucleotide Polymorphism Database (NCBI dbSNP) [7] represents the polymorphism by a code with “rs” prefix and a various number of digits. The database is available on the web https://www.ncbi.nlm.nih.gov/snp/. This representation of genetic information presents an essential key for search and integration in the various parallel and follow-up systems. Together with HGNC code for gene, the dbSNP code creates the most important representation of the genetic information that is commonly used in interpretation databases. The NCBI dbSNP data is accessible via defined REST API.

- Human Genome Variation Society (HGVS) [8] is the next form of the sequence variations description. The HGVS database is available on web https://varnomen.hgvs.org/. HGVS codes have an inner logic structure, and it is possible to determine information about changes in the given sequence. The HGVS coding is often stated together with NCBI RefSeq code.

2. Interpretation of the Genetic Analysis Results

If the genetics information is appropriately used at the clinical level, there is necessary to extend the record of the genetic analysis by interpretations of the genetic analysis results at the clinical level. The primary resources of interpretations are experts and studies. However, there are also several databases designated to collect and distribute the known connections of the genetic and clinical information. There is possible to use this information directly (following the license), or it is possible to refer to these resources.

- The PharmGKB [9] (available at https://www.pharmgkb.org) is essential interpretation databases in its focus as well as its scale. The database focused on pharmacogenomics information contains crucial clinical information of genetic and chemical substances associations in the field of efficacy, dosage, toxicity or metabolism and also contains associated diseases. The narrative interpretation (phenotype) is related to the allele combination (genotype) for given polymorphism (mainly represented by dbSNP rs code). Each interpretation is supported by the level of evidence parameter, where 1A is the best evidence, and 4 is the worst evidence.
- The SNPedia [10] is an overview database focused on SNPs and is available at https://www.snpedia.com/index.php/SNPedia. The dbSNP rs codes are the primary key for searching in the database. The database contains selected relevance PubMed reference to journal papers linked with the given polymorphism as evidence of presented interpretation conclusions. In this case, the risk factor of disorder or disease represents the interpretation of genetic information (e.g. obesity or diabetes mellitus 2nd type). The SNPedia has its evidence power parameter (called Magnitude) defined from 0 (not information) up to 10 (significant information).
- The ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) [11,12] is a database within the NCBI portfolio. ClinVar is directly connected with the dbSNP, where there are interactive links to the clinical information in the ClinVar on the contrary the previous two databases, the ClinVar uses the HGVS as the primary key for the binding. The interpretation conclusions are supported by published studies as in previous databases, again by reference to the PubMed database.

Finally, some authorities create guidelines for a drug prescription dependent on the genetic information, e.g. U.S. Food and Drug Administration (FDA), Dutch Pharmacogenetics Working Group or Clinical Pharmacogenetics Implementation Consortium (CPIC). These guidelines are fundamental because they present summarized view (recommendation) of the genetic impact on the drug administration.

3. Data Exchange

Although the analysis results are primarily in the PGC, we suppose that possible communication between PGC and Electronic Health Record (EHR) will be covered by an appropriate communication standard. For exchange of the medical data, it is very relevant to use some HL7 standards. Standards for the exchange of genetics information are developed in the HL7 Clinical Genomics Work Group [13], which describes the set of documents in the genetic information exchange. We just mention a selection of important materials for each type of HL7 communication standard:

- HL7 version 2 contains two types of messages with genetic information: messages from the cytogenetics area [14] and messages for genetic variation [15,16]. The documents [14,15,16] specify the content of messages in quite a detail and are different in the LOINC panels that the messages use. Generally, message type Genetic Test Result Reporting [14,15] is defined hierarchically (parent-child) connected panels of LOINC.
- HL7 FHIR [17] utilizes class Observation, with several extensions that are explicitly defined for genetics. From the perspective of our purpose, the most interest materials are Observation-genetics Profile [18] and DiagnosticReport-genetics Profile [18]. Both profiles declare extensions for the genetic information exchange.
- The Standard Clinical Document Architecture (CDA) [19] is part of HL7 standards version 3, derived from the Reference Information Model (RIM). In the genetic analysis area, the use of CDA is same as in the other laboratory tests. There is the implementation guide for the message from genetic area [20].

4. Results – a Strategy of a Personal Genetic Card

The PGC system development strategy includes two parts: the structure of the system and the definition of processes. Figure 1A) represents the schema of resources integration. Parameters of the test parts are defined by LOINC, the values are either simple data types (numbers, string) or defined by terminology databases. The connection between genetic analysis and interpretations is shown in Figure 1B). The interpretation can be divided into three possible forms. First, each part of the test (each polymorphism) has partial interpretations (the number of interpretations depends on the number of polymorphisms and the number of factors (e.g. the number of drugs bound to the given polymorphism). Second, the test has a summary interpretation for each factor. This summary represents the overall recommendation for the factor and cannot be related to all partial interpretations. These two interpretation types are predefined in the genetic card by the provider and automatically associated with the analysis results. The third type is the user extension of the interpretations. The extensions allow adding another information directly to the report. It means that they are bonded with a specific patient and not represented in the knowledge database. The geneticist who adds the extension must sign them, and he/she bears the responsibility. All interpretations must refer to valid scientific studies.

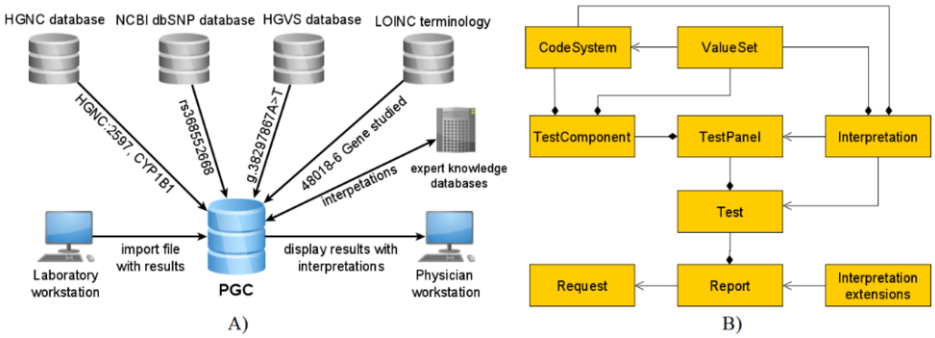

Figure 1. A) Schema of Personal Genetic Card elements, B) Simplified class diagram of the PGC system

The processes within the PGC system are depicted in the use case diagram in Figure 2.

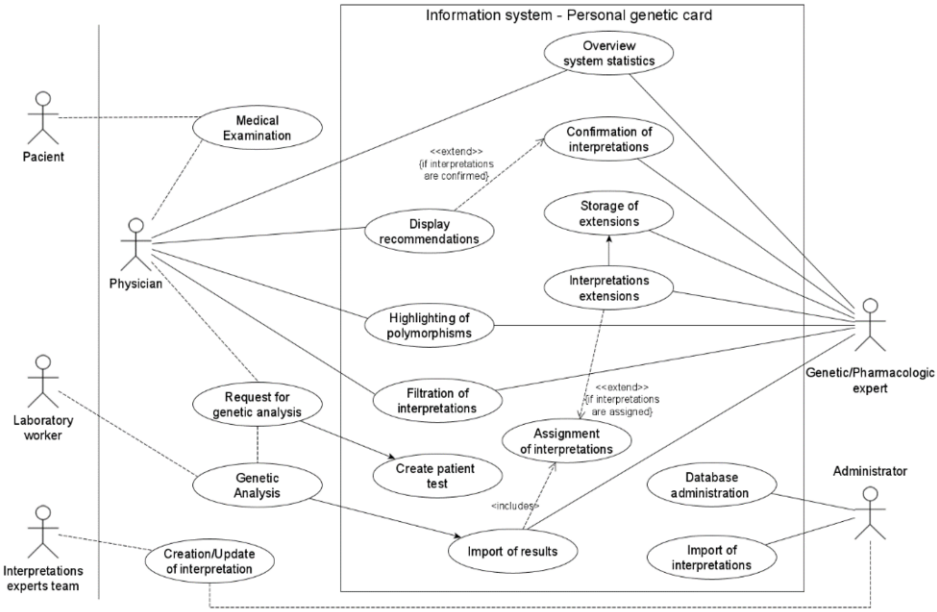

Figure 2. Use case diagram of the proposed PGC system

The initial process is the request of the analysis from a physician within a medical examination (with drug administration). The second step is the genetic analysis performed by the laboratory worker. The result of the analysis (file with data) is uploaded by the expert to the PGC system for specified patient and type of test (defines polymorphisms for interpretations assignment). After results upload, the predefined interpretations are automatically assigned to the results. The predefined interpretations are created by expert team on the side of the system provider and imported into the PGC system. At this moment, the expert can browse interpretations, filter them by a factor (e.g. substances) and highlighted polymorphisms. The genetic/pharmacologic expert can add an extension of interpretations. The last step on the laboratory side is confirmation

of the interpretations (in the case that the record contains interpretation extension, the record must be authorized by the electronic signature of the expert). All the data of analysis bound with interpretations are symmetrically encrypted before storing. Finally, the physician can display the interpretations and recommendations (with the functionality of filtering) and use them for the clinical result of the examination and drug administration.

5. Conclusion

In this contribution, we described our development strategy of the information system called Personal genetic card for representation and interpretation of the genetic analysis. The main purpose of the proposed system is a connection between genetic analysis and clinical issues like drug administration. The clinical use of genetic analysis relies on clinical interpretation because physicians will use the interpretations and even more the recommendations, not the raw data of the analysis. We described terminology systems that are appropriate for the representation of genetic analysis in the structured form. The structured representation of the analysis is completed by a structured form of the interpretations. The use of terminology standards ensures clean and interoperable representation of the information. Further, we defined particular roles and determined processes that are defined within the PGC system use. By these steps, the system PGC allows to transfer the personal genetic analysis results into the clinical interpretation and recommendation and potentially increase the quality of personalized healthcare.

Acknowledgment

This project is supported by the grant of the Ministry of Industry and Trade of the Czech Republic No. 2018 FV30421 GENOMKIT – Progressive technology for the rationalization of personalized pharmacogenomics, nutrigenomics and sports medicine.

References

[1] Hoffman JM, Dunnenberger HM, Kevin Hicks J, et al. Developing knowledge resources to support precision medicine: principles from the Clinical Pharmacogenetics Implementation Consortium (CPIC). J Am Med Inform Assoc. 2016; 23: 796-801. doi:10.1093/jamia/ocw027.

[2] Regenstrief Institute, Inc. Logical Observation Identifiers Names and Codes - LOINC®. www.loinc.org (accessed July 18, 2020).

[3] Deckard J, McDonald CJ, and Vreeman DJ. Supporting interoperability of genetic data with LOINC, Journal of the American Medical Informatics Assoc. 2015; 22: 621–627. doi: 10.1093/jamia/ocu012.

[4] International Health Terminology Standards Development Organization. SNOMED CT & Other Terminologies, Classifications & Code Systems. https://www.snomed.org/snomed-ct/sct-worldwide (accessed July 18, 2020).

[5] Yates B, Braschi B, Gray K, Seal R, Tweedie S, Bruford E. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2017; 45: 619-625. doi: 10.1093/nar/gkw1033.

[6] Leary NA, et. al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44: 733-45. doi: 10.1093/nar/gkv1189.

[7] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29: 308-11. doi: 10.1093/nar/29.1.308.

[8] den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update, HUMAN MUTATION, 2016; 37: 564–569. doi: 10.1002/humu.22981.

[9] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB and Klein TE. Pharmacogenomics Knowledge for Personalized Medicine. Clinical Pharmacology & Therapeutics, 2012; 92: 414-417. doi: 10.1038/clpt.2012.96.
[10] Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Research 2011; 40 (Database issue): D1308-12. DOI: 10.1093/nar/gkr798.
[11] Landrum MJ, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018; 46: 1062–1067. doi: 10.1093/nar/gkx1153.
[12] Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016; 44: 862–868. doi: 10.1093/nar/gkv1222
[13] HL7 Clinical Genomics Work Group. http://www.hl7.org/Special/committees/clingenomics/ (accessed July 18, 2020).
[14] HL7 International Inc. HL7 version 2 Implementation Guide: Clinical Genomics; Fully Loinc-Qualified HL7 International Inc. Cytogenetics Model, Release 1, (2014).
[15] HL7 International Inc. HL7 version 2 Implementation Guide: Clinical Genomics; Fully Loinc-Qualified Genetic Variation Model, Release 1, ORU^R01, HL7 Version 2.5.1. (2009)
[16] HL7 International Inc. HL7 Version 2.5.1 Implementation Guide: Orders and Observations: Interoperable Laboratory Result Reporting to her, Release 1, ORU^R01, HL7 Version 2.5.1, (2007)
[17] HL7 International Inc. HL7 FHIR Release 4. https://www.hl7.org/fhir (accessed July 18, 2020).
[18] HL7 International Inc. HL7 FHIR Genomics Implementation Guidance. https://www.hl7.org/fhir/genomics.html (accessed July 18, 2020).
[19] Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo A. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc. 2006 Jan-Feb; 13,1: 30–39. doi: 10.1197/jamia.M1888.
[20] HL7 International Inc. Implementation Guide for CDA: Genetic Testing Report (GTR), (2013).