*Digital Personalized Health and Medicine*
*L.B. Pape-Haugaard et al. (Eds.)*

*doi:10.3233/SHTI200199*

# The openEHR Genomics Project

*To the memory of Gianluigi Zanetti*

Cecilia MASCIA [a,1], Francesca FREXIA [a], Paolo UVA [a], Gianluigi ZANETTI [a], Luca PIREDDU [a], Gideon GIACOMELLI [b], Christina JAEGER-SCHMIDT [b], Aurelie TOMCZAK [b,c], Simon SCHUMACHER [b], Florian KRAECHER [b], Roland EILS [b], Silje LJOSLAND BAKKE [d], Heather LESLIE [d]

[a] *CRS4: Center for Advanced Studies, Research and Development in Sardinia*
[b] *HiGHmed Consortium*
[c] *Institute of Pathology, University Hospital Heidelberg*
[d] *Clinical Model program, openEHR Foundation*

**Abstract.** Current high-throughput sequencing technologies allow us to acquire entire genomes in a very short time and at a relatively sustainable cost, thus resulting in an increasing diffusion of genetic test capabilities, in specialized clinical laboratories and research centers. In contrast, it is still limited the impact of genomic information on clinical decisions, as an effective interpretation is a challenging task. From the technological point of view, genomic data are big in size, have a complex granular nature and strongly depend on the computational steps of the generation and processing workflows. This article introduces our work to create the openEHR Genomic Project and the set of genomic information models we developed to catch such complex structure and to preserve data provenance efficiently in a machine-readable format. The models support clinical actionability of data, by improving their quality, fostering interoperability and laying the basis for re-usability.

**Keywords.** genomic models, openEHR, mutations, variations, structured data

## 1. Introduction

Within the healthcare domain, Next Generation Sequencing (NGS) data sets are a potential asset of clinically valuable information that, if properly read and deciphered, can pave the way for life-enhancing applications such as personalised therapeutic treatments or early detection of certain hereditary diseases, just to name a few. However, handling genomic data is a very complex task from both clinical and technological perspectives. Several initiatives are addressing the technological challenges [1,2], but they are mainly focused on the exchange aspects, while many facets of the subject still remain unaddressed. In the first place, genomic information is composed of many levels, derived from each calculation step, but usually the intermediate results are expressed in a wide variety of text file formats (e.g., SAM/BAM, VCF, etc.), thus hindering the efficient reuse, sharing and possible inclusion of data in an electronic health record (EHR). In addi-

[1] Corresponding Author: Cecilia Mascia, C/O CRS4, Loc. Piscina Manna - Edificio 1, 09050 Pula (CA), Italia; E-mail: cecilia.mascia@crs4.it.

tion, bioinformatic workflows typically rely on many external resources (e.g., reference genome sequences, biological annotation databases, analysis tools) that evolve rapidly and have a significant impact on the final results. Hence, it is essential to capture the details of each workflow execution to ensure, for instance, accurate data auditability and reproducibility or to make results comparable when a different pipeline is used.

The availability of NGS data in a structured and machine-readable format would therefore facilitate their integration into effective clinical actions, thanks to the possibility of capturing granularity and provenance. This need has motivated the creation of genomic models in the form of openEHR archetypes [3,4], fostering the foundation of the Genomics openEHR modelling group – consisting of members from the CRS4 (Italy), the HiGHmed Consortium (Germany), the EUCANCan project (EU, Canada), the CINTESIS group (Portugal), the BigMed project (Norway) and the international openEHR Clinical Modelling Program [5]. This article describes how our collaborative work led to the launch of the openEHR Genomics Project, the results obtained so far and the future directions.

## 2. Method

### *2.1. The OpenEHR Approach and the OpenEHR Clinical Modelling Program*

OpenEHR is set of open specifications for structured data, designed to represent medical knowledge separating domain semantics from any specific technology [6]. OpenEHR archetypes are the maximum set of attributes describing a clinical concept in a machine-readable format. They enable to represent concepts of different scale of granularity, to add specific constraints on elements and to embed their semantics. Conceptual models are formalised in computable entities through the Archetype Definition Language (ADL), which can be easily converted to a number of other popular formats like XML. The OpenEHR framework also includes a mechanism to aggregate archetypes in use case-oriented structures (i.e., templates), a query language optimised for archetypes (i.e., AQL-Archetype Query Language), a set of open APIs and a conformance test approach to validate real implementations.

The openEHR Clinical Modelling Program gathers clinicians, researchers and implementers in an international collaborative action for the development of multilingual archetypes and templates. The direct involvement of domain experts in the design process is one of the key strategies adopted to obtain high quality models, whose re-use is encouraged. Archetypes are collaboratively reviewed and made available to the community through the international Clinical Knowledge Manager (CKM) [7] – the openEHR artefacts repository – which has also some national instances developed to meet specific demands.

### *2.2. The Initial Genomic OpenEHR Archetypes Collection*

The starting point was the set of 11 archetypes proposed in [3] that should be used within the *Laboratory test result* archetype to represent genetic test results, developed on the basis of the Variant Calling Format (VCF) file specification [8]. VCF files consist of a set of rows each describing a variation that occurred in a sample sequence together with customisable functional annotations like, for instance, *'transcript features'* and the *'predicted impact'*. Within the archetypes, mutations are described according to the Sequence Variant Nomenclature[2] of the Human Genome Variation Society (HGVS) [9]. Further, the models included elements like the *'conservation score'* and the *'allele frequency'* that are usually not included in the medical report itself, but are potentially relevant on the bioinformatics research side. Finally, each tool or external database used in the analysis can be tracked by explicitly recording their name, version and, where appropriate, URL, thus integrating provenance within the data.

### *The Design Process*

Two use cases drove the creation of a common semantic data model for the representation of variant information derived from whole genome and exome sequencing (WGS, WES) data:

- Rare Diseases: to identify pathogenic variants responsible for rare diseases through WES tests of patients and their parents.
- Oncology: to enable an evidence-based and better informed diagnostic and treatment design, based on molecular stratification.

The whole design process has been guided by the co-chairs of the openEHR Clinical Modelling Editorial Group (CMEG), responsible for the modeling governance within the community.

The first step was the rethinking of content and structure, when the whole set of nodes constituting the archetypes have been assessed from two opposing perspectives. Starting from the overall picture, we prepared a mind map (using the collaborative tool XMind[3]) to depict all the necessary pieces of information describing the use cases, to detect possible repetitions or overlapping with existent archetypes and to simplify the structure. Starting from the single nodes, we checked the adequacy of the concept name, its adherence with standard nomenclatures and the consistence with the element's description, adding examples to clarify the intended usage.

After reaching a good level of agreement, working with the Archetype Designer[4], we converted the mind maps into archetypes, taking further decisions on data types, node occurrences and constraints and adding, when possible, bounds to external standard ontologies.

## 3. Results

On April 2019, the Genomics Project[5] was officially created in the openEHR CKM. The project collects and makes publicly accessible the 13 archetypes developed in this work.

The *Genetic variant* is the core archetype designed to be nested within the *Laboratory test result* to report observations and annotations related to mutations found in the genome as the result of a sequencing test. The morphological description of the muta-

[2] Guidelines available at: `http://varnomen.hgvs.org/`

[3] XMind 8: `https://www.xmind.net/xmind8-pro/`

[4] Archetype Designer: `https://ehrscape.marand.si/designerv2/#/`

[5] CKM Genomic project: `https://ckm.openehr.org/ckm/#showProject=1013.30.50`

tions – i.e, nucleotide changes and genomic coordinates – is given at the DNA level with reference to a genomic reference sequence. The model supports all the simple change types identified in the HGVS guidelines [9] plus translocations and copy number variations – two structural variations that are often relevant in cancer studies. In addition, further descriptions of the mutations are possible at the transcript level in terms of changes in the coding sequence ('DNA changes') and in the protein ('amino acid changes'). The reference sequence used to annotate the variations is explicitly recorded too, through a dedicated archetype (*Reference sequence*). The archetype design allows the recording of annotations for each transcript affected by that specific mutation, pointing out the one with the highest predicted impact, considered relevant for the clinical evaluation. In this regards, the functional interpretation of the variants can be specified following the ACMG classification [10] or by citing relevant scientific literature. All the external resources used by the data analysis pipeline, including biological databases and software, can be explicitly recorded by giving their IDs, version and, possibly, a unique resource locator (URL); an archetype is dedicated to capturing this information in a structured manner (*Knowledge base)*. In order to clearly state the concepts' semantic and improve data quality and interoperability, most of the model's nodes are codified using LOINC[6], an international standard ontology for identifying health measurements, observations, and documents.

OpenEHR's flexibility allows to project the archetype data onto a variety of different use cases, enabling to archive all the elements automatically captured during pipeline execution, but presenting to the final user only those relevant to the task at hand. To demonstrate how the archetypes can be combined and integrated, an example template for a generic sequencing test report has been included in the Project.

The initial review round of the *Genetic variant* archetype started on April 26 2019, and, afterwards, we commenced in staggered steps the review process of each archetype, inviting all the interested domain experts in the project to give their opinions and comments. The process is managed by the project editors and performed completely via the CKM, which also allows the tracking of archetypes changes. At the moment of writing, one archetype has been published after one review round while the others are still in the consolidation phase. In the meantime, the Norwegian translations of the archetypes are currently being refined.

## 4. Discussion

The focus of our work was the creation of information models to capture genomic data for clinical and research applications. The openEHR community has overall expressed a good consensus on the structure of the models, together with a series of useful feedbacks. For instance, a series of suggestions has been provided to improve the coverage of the genomic domain, particularly with respect to the representation of complex mutations, such as splicing variants affecting RNA sequences, which are not yet supported by model. The review process has also provided important hints on how to improve and extend the usability of the models in real contexts, as a set of suggestions emerged to refine the requirements and to generalise the models towards different use cases than those

[6] Logical Observation Identifiers Names and Codes (LOINC): `https://loinc.org/`

initially considered. The feedback will be integrated in the next versions of the models while trying to balance the different perspectives.

More generally, the number of reviewers involved in this specific CKM domain tend to be fewer than those involved in more long-standing ones. We believe this is mainly due to the novelty and the specialised nature of the project, and that the situation can be improved with targeted actions to disseminate the potential of the obtained results.

## 5. Conclusion and Future Work

In this work we presented the openEHR Genomics Project and its first results: the models to represent clinical genomic data and its provenance with a set of openEHR archetypes. These results have been developed through a collaborative process supported by the openEHR Clinical Modelling Program. Capturing the complexity of genetic information remains a challenging task, but the results presented in this work are the promising starting point for an ongoing development and improvement process. Our next steps include to extend the models to cover additional concepts, like the complex genetic variants, and to continue the ongoing review process in the international CKM. Further, additional model reviews at the Norwegian and German repository instances are already planned, and more will be organized in due time.

## 6. Acknowledgements

The authors from CRS4 dedicate this work to the memory of Gianluigi Zanetti, whose legacy continues to inspire. This work has been partially supported by the European Joint Programme on Rare Disease (grant agreement N. 825575).

## References

[1] Genomics - FHIR v4.0.0, https://www.hl7.org/fhir/genomics.html, (Accessed on 09/29/2019).

[2] Phenopackets - Concepts and Technology, http://phenopackets.org/, (Accessed on 09/29/2019).

[3] C. Mascia, et al., OpenEHR modeling for genomics in clinical practice, International Journal of Medical Informatics 120 (2018) 147–156 (Dec. 2018). doi:10.1016/j.ijmedinf.2018.10.007.

[4] P. A. Maranhão, et al., Challenges in design and creation of genetic openEHR-archetype, Studies in Health Technology and Informatics 247 (2015) (2018) 835–839 (2018). doi:10.3233/978-1-61499-852-5-835.

[5] Genomics information - openEHR Clinical - openEHR wiki, https://openehr.atlassian.net/wiki/spaces/healthmod/pages/50987020/Genomics+information, (Accessed on 10/03/2019).

[6] What is openEHR?, https://www.openehr.org/about/what_is_openehr, (Accessed on 09/10/2019).

[7] Clinical knowledge manager, https://ckm.openehr.org/ckm/, (Accessed on 09/10/2019).

[8] The Variant Call Format (VCF) Version 4.2 Specification, https://samtools.github.io/hts-specs/VCFv4.2.pdf, (Accessed on 09/27/2019) (2019).

[9] J. T. Den Dunnen, et al., HGVS Recommendations for the Description of Sequence Variants: 2016 Update (2016). doi:10.1002/humu.22981.

[10] S. Richards, et.al, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology, Genetics in Medicine 17 (5) (2015) 405–423 (2015). arXiv:15334406, doi:10.1038/gim.2015.30.