OMOP Genomic mapping capacities in conversion of comprehensive genomic profiling results

Maria Rogozhkina[1], Vlad Korsik[1], Varvara Savitskaya[1], Alexander Davydov[1]
[1] - Odysseus Data Services

Introduction

Omics data (genomics, proteomics, metabolomics and etc) tends to be the most important data currently because of its possibility to influence decisions in a variety of medical fields. The primary aim of this study was to evaluate the efficacy of representation of genetics data by OMOP Genomic vocabulary. We utilized a real world data depersonalized database (so that it can be applied to any enriched genomic data format such as FMI (FoundationOne), VCF (Variant Call Format), GFF (General feature format) and others) to test the conversion capacity of the vocabulary.
The secondary aim was to define the best conversion strategy for real world database test results (it is one of the most commonly used comprehensive genomic profiling systems worldwide).

Methods

We processed 3 source genomic data repositories from a real world databases including comprehensive genomic profiling systems.

- Source_data_type_1 = copy number = CN (i.e. amplification or deletion),
- Source_data_type_2 = short variant = SV (a single nucleotide variations or small insertion/deletion).
- Source_data_type_3 = multiple myeloma genetic data = MMGD.

The conversion was performed on CDM 5.4 version with OMOP Vocabulary version: v5.0 09-APR-22.

Among The CN table we analyzed 6 033 de-identified records with name of gene, synonym name, type of mutation (amplification or deletion) and copy number:

source_table_name	gene_name	gene_syn	amplificationordeletion	copynumber
CN	STK11	STK11	deletion	0
CN	CDK8	CDK8	amplification	8
CN	CUL4A	CUL4A	amplification	8

For mapping automation we applied the full-match approach with subsequent manual curation. Every distinct source name of the gene was a full name counterpart in postcoordination approach. In precoordination approach distinct source name of the gene and type of mutation were a full name counterpart. The list of targets was aggregated by concept_id list of preferred names and synonyms.

Single Nucleotide Variation table includes 233 793 records with information about the gene, substitution in DNA, RNA and protein with position, number of chromosome, type of alteration, sequencing coverage and many other (22 columns). Here are shown only the most important ones:

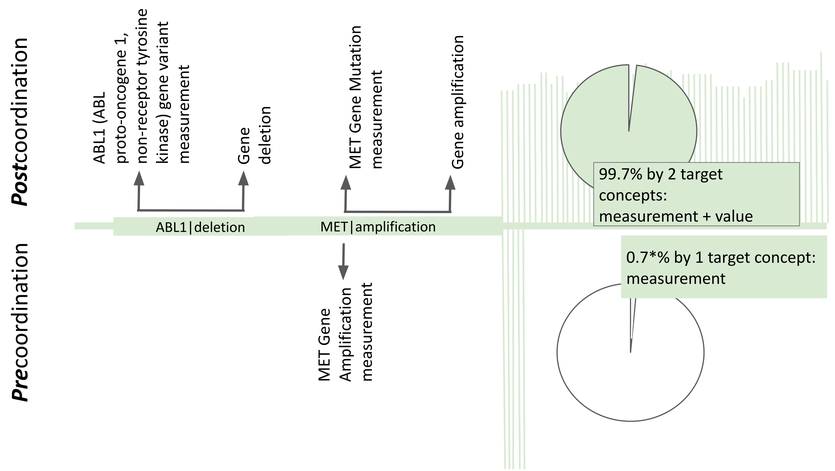
source_table_name	gene	RNA	protein	chromosome	codingtype	sequencingcoverage
SV_extended	SMO	2081delC	P694fs*82	chr7	frameshift	1622
SV_extended	APC	2078A>C	K693T	chr5	missense	4394
SV_extended	TP53	473G>T	R158L	chr17	missense	6742

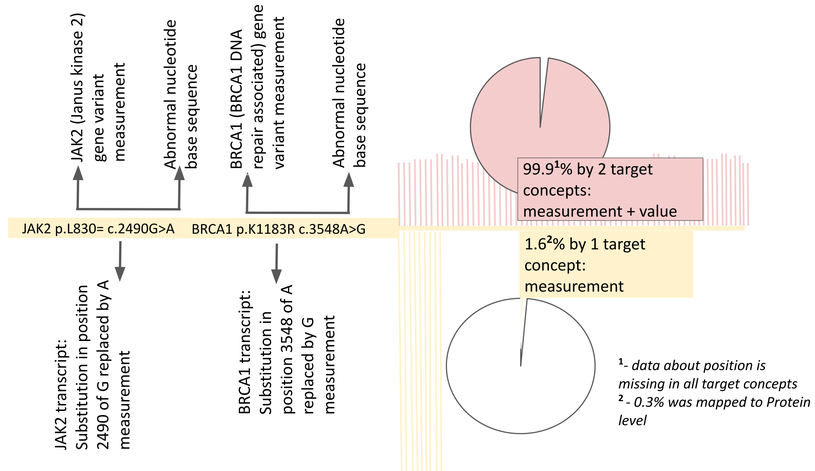
Both protein and RNA columns are filled well so source_concept_name was compiled as gene, RNA, protein. At first we made a conversion on the RNA column, if nothing was found, we did matching on the protein column, and if nothing was found again, we tried to do uphill mapping on the Genetic Variation class.

MMGD is a little table containing only 33 rows with information about specimen, method, gene and type of abnormality with all listed fields used to define source_code.

source_table_name	specimen	method	gene
MMGD	Bone Marrow Aspirate	FISH	T(14;16)
MMGD	Bone Marrow Biopsy	FISH	AMP1Q21

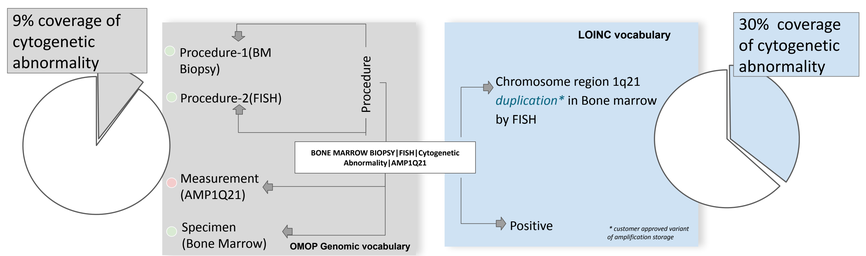
Results





For copy number table the coverage of source data was performed by postcoordination as: meaningful coverage of both genetic variation and related copy number type may be achieved in 99,7% of cases with *post*coordiantion with semantics equivocal distribution between event and it's value.

In contrast, *pre*coordination resulted in coverage of 0,7% of data, with absolute (one-to-one) semantics match coverage. I.e. *postcoordination* seems to be optimal strategy to define the changes in number of genes.



Explicit, 1-to-1 representation request is a MMGD scenario. Impossibility of OMOP Genomic to provide the way to reflect the cytogenetic abnormality , the need to store specimen and method attributes as separate facts resulted in targeting to non-genomic terminology.

Lack of appropriate target concepts results in need to uphill mapping in most cases of nucleotide variation tables.
Despite almost perfect (99,9%) SV table's records mapping to OMOP Genomic concepts the semantic coverage remains improper.

Full semantic match is attributed to 1,3% of codes were targeted to RNA Variant and 0,3% of codes were mapped to Protein Variant. Leftover concepts (97,7%) were targeted to Genetic Variation I.e. uphill mapping is the major storing strategy.

Table name	Modeling Approach	Mapping Rate	Coverage rate		
			Gene coverage	Alteration type coverage	Alteration coordinate coverage
SNV	Pre	~2%	→100%	→100%	→100%
	Post	~100%	→100%	0%	0%
CN	Pre	~1%	→100%	~1%	NA
	Post	~100%	→100%	→100%	NA

Rationale for appropriate modeling approach selection for specific alteration types

Conclusion

OMOP Genomic vocabulary covers a large number of needs well, but some improvements are also needed. It is required to:

- perform deduplication and expand the list of synonyms for a better search using Clingen database
- increase the number of concepts to cover a larger number of cases: transfer information from DoCM and Cancer Hotspots, take Clinically relevant variation from ICGC to the OMOP Genomic vocabulary.
- make the genomic LOINC/SNOMED etc concepts non-standard, and then map them to the standard OMOP Genomic concepts.
- prevent "combinatorial explosion" by allowing postcoordiantion at least for Copy Number changes
- ratify the logic for storage of method and specimen

OHDSI EUROPE OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS
ODYSSEUS DATA SERVICES