Genomic Data Harmonization through the OMOP Standardized Vocabularies

Denys Kaduk[2], Violetta Komar[2], Asieh Golozar[6], Peter Robinson[9], Alex H. Wagner[10], Michael Gurley[4], Seng Chan You[3], Shin Seojeong[3], Shaadi Mehr[8], Andrew Williams[4], Rimma Belenkar[7], Meera Patel[7], Shilpa Ratwani[1], Ron Miller[1] and Christian Reich[1],

[1]IQVIA, Plymouth Meeting, PA. USA, [2]Odysseus Data Services Inc., Cambridge, MA, USA, [3]AJOU University, Suwon-si, Gyeonggi-do, South Korea, [4]TUFTS, Medford, MA, USA, [5]Northwestern University, Evanston, IL, USA, [6]Regeneron Pharmaceuticals, Inc, Tarrytown, New York USA, [7]Memorial Sloan Kettering Cancer Center, New York, NY, USA, [8]MMRF Norwalk, Connecticut, USA, [9]The Jackson Laboratory for Genomic Medicine, Hartford, Connecticut Area, USA, [10]Washington University School of Medicine, St. Louis, MO, USA

Abstract

In a typical observational study, cohorts, exposures, and outcomes can be sufficiently defined through the presence or absence of clinical events encoded by a defined set of concepts. Precision Oncology however is more challenging than that of most other conditions. Accurately integrating EHRs with genomic data for the discovery of clinically actionable variants can generate new insights into disease mechanisms and provide better tools for identification of effective treatments. Currently, the OMOP Standardized Vocabularies lacks a canonical, comprehensive and non-redundant representation of genomic variants. We developed a preliminary simple heuristic for canonization and evaluated six prominent somatic cancer variant knowledgebases to serve as a source of clinically relevant variants to include. We found these to have a low level of overlap and recommend adopting a solution that combines these repositories and incorporates a deduplicated set for the purpose of Standard Concepts.

Introduction

Clinical research in precision oncology requires concise, standardized and searchable interpretations of detected variants. Many institutions have created collections of clinically relevant variants and formalized interpretations such as disease implication and targeting by pharmaceutical compounds through curation from the biomedical literature. These efforts have resulted in disparate knowledge representation of the genomic information. Currently, there is no comprehensive genomic terminology available in the public domain that would support harmonization of genomic oncology data, which is necessary for standardized analytics in a research network. For OHDSI, representing the various knowledgebases in a standardized manner requires an open, interoperable sharing of variant interpretation and an automated methodology so a comprehensive approach to cancer precision medicine can be developed.

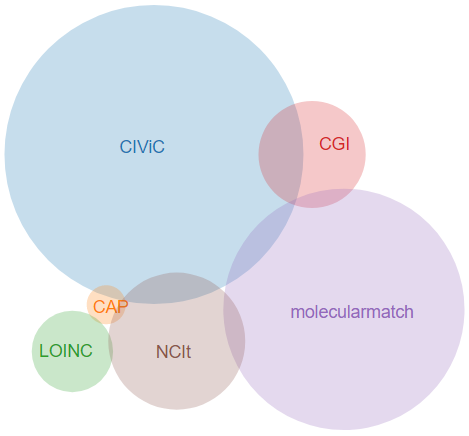
To illustrate the challenges of harmonizing across multiple variant representations, we looked at all interpretations describing the variant AKT1 E17K, a change in the amino acid residue at position 17 in the RAC-alpha serine/threonine-protein kinase protein where glutamic acid has been replaced by lysine, which is implicated in lung and breast cancer. This variant is present in four of the six databases (see below), but each knowledgebase represents this variant differently. The differences are due to representations at the genomic, complementary DNA (cDNA) (which could be in antisense direction) and protein levels, each of which has different reference sequences in different versions from different databases (NCBI, EMBL, LRG):

Dimension	Number of representations
Genomic reference sequences	4
Versions of genomic reference sequences	8
cDNA reference sequences	3
Versions of genomic reference sequences	3
Protein reference sequences	17
Versions of protein reference sequences	17

To address these challenges and to develop interoperability between genomic data and clinical care, a collaboration between the OHDSI Oncology Workgroup and the VICC consortium (1) is seeking to develop a canonical representation of equivalent variants to serve as standard concepts in the OMOP Standardized Vocabularies.

We developed a simple consolidation approach of duplications, insertions, deletions and duplications on the basis of gene symbol, sequence type (g, c, p), reference sequences, versions and locations. As a starting point, we incorporated six such source vocabularies: National Cancer Institute Thesaurus (NCIt), College of American Pathologists Cancer Checklists (CAP), Clinical Interpretation of Variants in Cancer (CIViC), Cancer Genome Interpreter (CGI), MolecularMatch and Logical Observation Identifiers Names and Codes (LOINC) (figure 1).

Figure 1. Consolidation of six genomic databases of variants relevant in cancer.



Overlapping Vocabularies	Count
CAP	10
CAP, CIViC	9
CAP, CIViC, LOINC, MolecularMatch, NCIt	1
CAP, CIViC, LOINC, NCIt	1
CAP, CIViC, MolecularMatch	1
CAP, CIViC, MolecularMatch, NCIt	2
CAP, CIViC, NCIt	12
CAP, LOINC	1
CAP, LOINC, NCIt	1
CAP, NCIt	5
CGI	154
CGI, CIViC	116
CGI, CIViC, MolecularMatch	30
CGI, MolecularMatch	23
CIViC	2165
CIViC, LOINC, MolecularMatch, NCIt	2
CIViC, LOINC, NCIt	4
CIViC, MolecularMatch	67
CIViC, MolecularMatch, NCIt	29
CIViC, NCIt	73
LOINC	173
LOINC, NCIt	3
MolecularMatch	1444
MolecularMatch, NCIt	36
NCIt	358

Conclusion

The limited overlap between cancer variant databases suggests a union of these variants and a heuristic to deduplicate the overlap. The resulting collection can be instantiated as Standard Concepts and evaluated for utility in real-world genomic data research.

References

1. Wagner, A.H., Walsh, B., Mayfield, G. et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat Genet 52, 448–457 (2020)