Mapping variants of known significance to the OMOP Genomic Vocabulary

Michael Gurley, Asieh Golozar

Background

To support observational research on somatic variants of known significance, OHDSI has released the OMOP Genomic vocabulary. The OMOP Genomic vocabulary was constructed by aggerating multiple knowledge bases of assertions of significance about somatic variants: CGI, CIViC, ClinVar (subset), JAX, NCIt and OncoKB. The OHDSI vocabulary team took the approach of restricting variants to those present in knowledge bases to reduce the number of variants housed in the OMOP vocabulary. For context, over 500,000 variants have been submitted to Clinvar.[1] Housing over 500,000 variants within the OMOP relational model would degrade performance. The OHDSI vocabulary team's approach assumed that the knowledge bases provide reasonably complete coverage of variants called significant in real world data. To test the coverage of the OMOP Genomic vocabulary, we mapped 5466 "short variants" from Foundation One called 'significant' on patients from Northwestern Medicine. A short variant is defined by Foundation Medicine as a variant observation (base substitution or short insertion/deletion).

Methods

We imported 5466 unique short variants (including all fields provided by the Foundation One XML format) into a PostgreSQL database housing a OMOP 5.4 CDM schema and the 'OMOP Genomic 20210727' version of the OMOP Genomic vocabulary. We mapped the 'high-level' name of each variant to the OMOP Genomic vocabulary and each knowledge base vocabulary within the OMOP vocabulary. We defined 'high-level' name as how the variant appears in a Foundation One report: the concatenation of gene name, a colon and protein effect. For example: 'TP53:C242Y'. We performed a case-insensitive match of the 'high-level' name against CONCEPT.concept_name and CONCEPT_SYNONYM.concept_synonym_name. Next, we mapped each 'high-level' name to the CIViC API[2] to isolate any side-effect of the importation of CIViC into the OMOP vocabulary. The CIViC API mapping involved making a call to the CIViC API 'genes' resource with the 'gene' component of the 'high-level' name and exact matching the 'protein effect' component of the 'high-level' name with the 'genes' resources response's 'variants' array. Then we constructed a HGVS representation of each variant from available fields within the Foundation One XML format. The only possible HGVS representation possible to construct was a coding DNA refence sequence or a '.c' format. We took the 'transcript' field and normalized it to the current version by screen-scraping NCBI Nuccore database[3]. For example: 'NM_000546' normalized to 'NM_000546.6'. Then we concatenated the normalized 'transcript' to a colon, a 'c.' prefix and the coding sequence effect field from the Foundation One XML format. For example: "NM_000546.6:c.725G>A". We performed a case-insensitive match of the HGVS representation against CONCEPT.concept_name and CONCEPT_SYNONYM.concept_synonym_name. Finally, we also submitted the HGVS representation to the Clingen Allele Registry API[4] 'allele' resource to enable assessment of the coverage provided by Clinvar not filtered by knowledge bases.

Results

Figure 1 provides the results detailing the coverage of mapping 'high-level' names and '.c' HGVS representations of 5466 "short variants" from Foundation One called 'significant' on patients from

Northwestern Medicine

Representation	OMOP Genomic Vocabulary	CIViC API	Clingen Allele Registry API
High-level name	15%	7%	Not applicable
HGVS ‘.c’	5%	Not applicable	74%

Figure 1. Mapping results for mapping ‘high-level’ names and ‘.c’ HGVS to OMOP Genomic Vocabulary, CIViC API and Clingen Allele Registry API.

Conclusion

The coverage of variants declared ‘significant’ within real-world reports capable of being mapped to the OMOP Genomic vocabulary is low. Using lower-level HGVS representation did not improved the coverage of the variants. The suboptimal performance of mapping directly to the CIViC API points to the conclusion that building a list of variants based on a single publicly available knowledge base may not provide sufficient coverage of variants declared significant in real world data. To better understand the cumulative coverage of the OMOP Genomic Vocabulary and the contributing knowledge bases, the coverage of the reported variants across other knowledge bases such as CGI, ClinVar (subset), JAX, NCIt and OncoKB should also be evaluated. An alternative approach to enhance the coverage of the OMOP Genomic vocabulary can be through mapping significant variants to Clingen Allele Registry and incorporation of those variants to the vocabulary.

References/Citations

1. Xiang, J., Yang, J., Chen, L. et al. Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. Sci Rep 10, 331 (2020).
2. https://civic.readthedocs.io/en/latest/api.html
3. https://www.ncbi.nlm.nih.gov/nuccore/
4. https://allele-registry.tech-docs.io/