Harmonization of OMOP vaccine-related vocabularies through the Vaccine Ontology

[1]Yuanyi Pan[1,*], Warren Manuel[2,*], Rashmie Abeysinghe[3,*], Xubing Hao[2], Alexander Davydov[4], Qi Yang[5],

Asiyah Yu Lin[6,#], Licong Cui[2,#], Yongqun Oliver He[1,#]

[1] University of Michigan Medical School, Ann Arbor, MI, USA; [2] McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA; [3] Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX, USA; [4] Odysseus Data Services, Inc., Cambridge, MA; [5] IQVIA, Inc., King of Prussia, PA, USA; [6] National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA.

* These authors share first authorship; # Co-corresponding authors.

Background

Vaccines have played an important role in fighting against infectious diseases such as COVID-19. OHDSI/OMOP CDM and its associated vocabularies (e,g, CDC Vaccine Administered CVX, RxNORM, and SNOMED-CT) include a variety of vaccine-related terms. However, these vaccine vocabularies have different coverages and use different design patterns and representation styles[1]. As a result, the vaccine terms in these vocabularies could not be easily mapped and integrated.

To address the above challenge, we have formed an OMOP Vaccine Vocabulary Working Group (Vaccine Vocab WG) to map and integrate different vaccine vocabularies. Our basic strategy is to use the Vaccine Ontology (VO)[2,3], a community-based biomedical ontology in the vaccine domain, as the platform to systematically represent the mapped results for vaccine terms in individual vocabularies. We started with the mapping of CVX vaccine terms to the VO using manual and semi-automatic strategies.

Methods

Manual mapping and VO updating: The CVX vaccine terms were extracted from the CDC website[4]. The VO source was obtained from VO GitHub[5]. Ontobee[6] was used to query vaccine terms from the VO and related ontologies. A manual evaluation was performed for VO-CVX term mapping. The "rdfs:seealso" annotation property was used to add CVX codes in VO if such information is not available. For those CVX terms with no corresponding VO mapping, we generated new VO terms based on our standard VO design pattern[3]. The Protege OWL editor[7] was used for manual editing, and the Ontorat tool[8] was used for automatic addition of many VO terms based on pre-defined ontology design patterns. An Excel file that lists individual CVX terms and associated information is accessible on VO GitHub[9].

Semi-automated mapping approach: In order to identify candidate VO concepts for each CVX term, three similarity-based approaches were investigated: (1) Word-level similarity; (2) Embedding-based similarity; and (3) Hybrid approach. All these methods generated a similarity score for a CVX and VO term-pair. A threshold was set by experimentation where all the CVX and VO term-pairs above the threshold were considered as “matched”. Furthermore, each method generated up to 10 candidate sets

of matching VO terms for a CVX term which could be further reviewed by domain experts for confirmation.

Prior to similarity calculation across all the methods, normalization of lexical information was performed via lowercase and ASCII conversion, and expansion of common vaccine-related abbreviations and trade names sourced from the CDC [10,11]. The resulting text was tokenized, and used in the different matching approaches.

The first approach was to identify candidate pairs purely based on their word-level similarity. The lexical similarity was computed via the Jaccard similarity[12] wherein the coefficient is calculated as the intersection between tokens of the two concepts divided by the union of the two.

The second approach was to use pre-trained sentence encoders to obtain embeddings for concepts that are representative of their meanings. For this purpose, BioSentVec[13], a sentence encoder trained on PubMed and MIMIC-III documents, was utilized. The embeddings for each VO concept were then compared with each CVX term embeddings to identify the most similar terms. The similarities of these 700-dimensional embeddings were calculated using cosine similarity. The final method was a hybrid of the initial methods. Here, for each CVX term, the top-25 candidate VO matches were obtained from the embedding method. These candidates were then scored using word-level similarity to identify up to 10 final matches.

Results

Manual CVX-VO mapping and VO updating: Out of 254 vaccine terms in CVX, four terms, such as "99: RESERVED - do not use," were considered as irrelevant and excluded from our study, resulting in 250 vaccine terms for further analysis. A total of 88 CVX-VO mapping pairs were identified, which were mostly one-to-one, with a few exceptions exhibiting a one-to-many relationship. Additionally, we identified 69 CVX terms that have corresponding terms in VO but have no direct mapping annotation; to address this, we added the mapping annotation using "rdfs:seeAlso" in VO.

Our study found 134 CVX terms not initially present in the VO, which were then added to VO accordingly. Notably, CVX includes 16 immune globulin vaccine terms, which were all missed in VO. After group discussion, we formed a consensus that these should be classified as passive vaccines. All these 16 immune globulin terms were then added as subclasses under 'immune globulin passive vaccine', a subclass of 'passive vaccine' in the VO (Figure 1). Proper hierarchies and annotations were also added.

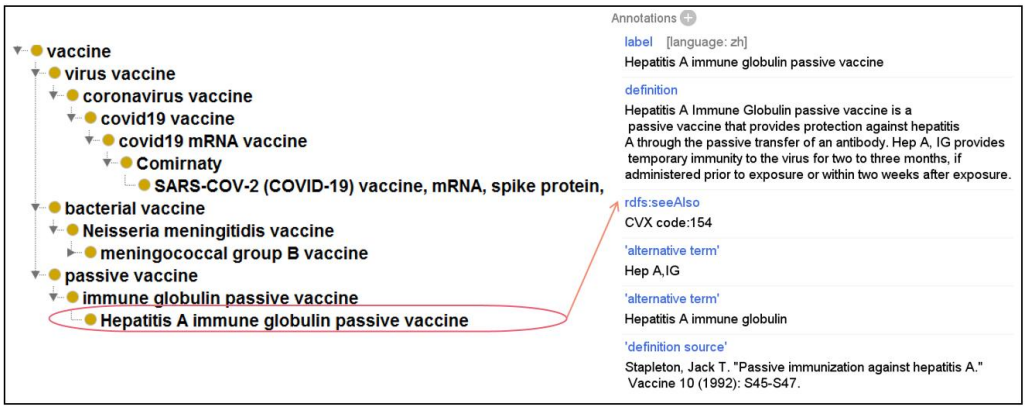

Figure 1. VO hierarchy and annotation.

Semi-automated mapping approach: The 4,102 vaccine terms under the VO concept *'vaccine'* (VO:0000001) and all CVX terms were considered here. The word-level, embedding-based, and hybrid approaches identified candidate mappings for 59, 65, and 59 CVX terms respectively. The results of the semi-automated method were compared with the manually annotated mappings. With the manual annotation gold standard, we evaluated the performance of the approaches in terms of precision, recall, and F-1 score. The results are given in Table 1.

Table 1: Performance of each model considering manual evaluation as a gold standard.

	Precision	**Recall**	**F-1 score**
Word-level similarity	0.6705	0.4758	0.5566
Embedding similarity	0.4851	**0.5242**	0.5039
Hybrid	**0.6782**	0.4758	**0.5592**

Overall, in terms of F-1 score, the hybrid method was found to be the best out of the three methods. Table 2 shows 5 examples for valid mappings obtained with the hybrid method.

Table 2: Five valid CVX to VO mappings identified by the hybrid method.

CVX term	**VO term**
CVX_130: DTaP-IPV	VO_0000067: Kinrix
CVX_20: DTaP	VO_0000064: Infanrix
CVX_75: vaccinia (smallpox)	VO_0000003: ACAM2000
CVX_187: zoster recombinant	VO_0003317: Shingrix
CVX_160: Influenza A monovalent (H5N1), ADJUVANTED-2013	VO_0003083: Influenza A (H5N1) Virus Monovalent Vaccine, Adjuvanted by GSK

Furthermore, we applied the hybrid approach to map RxNorm terms to VO where 765 RxNorm terms generated candidate matching VO terms. The manual review of the results are in prospect.

Conclusion

Overall, we applied both manual and semi-automatic methods to map CVX and VO vaccine terms and updated VO correspondingly. Out of three semi-automated methods, the hybrid method was shown to outperform the other two methods. The semi-automated methods can be promising as they require significantly less human effort than purely manual approaches. We are currently in the process of improving and extending our semi-automated mapping approach to other vaccine vocabularies and

correspondingly update the mapping method as needed. With expanded coverage and interoperability, the updated VO will further be used for systematic and integrative analysis of vaccine-related clinical data available in the OHDSI/OMOP compliant systems.

References

1. Abeysinghe R, Black A, Kaduk D, et al. Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method. J Biomed Inform. 2022;134:104162.
2. He, Y., Cowell, L., Diehl, A. et al. VO: Vaccine Ontology. Nat Prec (2009). https://doi.org/10.1038/npre.2009.3552.1
3. Lin Y, He Y. Ontology representation and analysis of vaccine formulation and administration and their effects on vaccine immune responses. J Biomed Semantics. 2012 Dec 20;3(1):17.
4. IIS: Current HL7 Standard Code Set CVX -- Vaccines Administered [cited 2023 June 7]. Available from: https://www2a.cdc.gov/vaccines/iis/iisstandards/vaccines.asp?rpt=cvx
5. https://github.com/vaccineontology
6. Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, Mungall C, Courtot M, Ruttenberg A, He Y. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res. 2017 Jan 4;45(D1):D347-D352.
7. Musen MA; Protégé Team. The Protégé Project: A Look Back and a Look Forward. AI Matters. 2015 Jun;1(4):4-12.
8. Xiang Z, Zheng J, Lin Y, He Y. Ontorat: automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns. J Biomed Semantics. 2015 Jan 9;6:4.
9. https://github.com/vaccineontology/VO/tree/master/docs
10. Vaccine Abbreviations | CDC [Internet]. [cited 2023 June 7]. Available from: https://www.cdc.gov/vaccines/terms/vacc-abbrev.html
11. U.S. Vaccine Names | CDC [Internet]. [cited 2023 June 7]. Available from: https://www.cdc.gov/vaccines/terms/usvaccines.html
12. Jaccard P. Nouvelles Recherches Sur la Distribution Florale. Bull la Soc Vaudoise des Sci Nat. 1908;44:223–70.
13. Chen Q, Peng Y, on ZL-2019 IIC, 2019 undefined. BioSentVec: creating sentence embeddings for biomedical texts. ieeexplore.ieee.org [Internet]. [cited 2023 June 7]; Available from: