Fix What's Broke: A Use Case-Driven Framework for Vocabulary Update and Maintenance

Asieh Golozar[1,2], Dan Smith[3], Patrick Alba[4], Andrew Nute[5], John Methot[6], Vlad Korsik[7], Polina Talapova[8], Stelios Theophanous[9], Espen Enerly[10], Georgina Kennedy[11], Henry Morgan Stewart[1], Kimmo Porkka[12], Qi Yang[13], Annelies Verbiest[14], Christian Reich[1,2]

[1]Nemesis Health, [2]OHDSI Oncology WG, [3]Emory University Hospital, Atlanta, USA, [4] VINCI, Veteran Affairs, USA, [5]Providence St. Joseph Health, USA, [6]Dana Farber Cancer Institute, USA,[7] EPAM Systems, Belarus, [8]SciForce, Ukraine, [9]Leeds Teaching Hospitals NHS Trust, UK, [10]Cancer Registry Norway, [11]University of New South Wales,[12]Helsinki University Hospital Comprehensive Cancer Center, Finland, [13]IQVIA, USA, [14]Antwerp University Hospital, University of Antwerp, Belgium

Background

Reliable evidence generation in oncology relies on a network of data partners equipped with comprehensive, longitudinal datasets encompassing cancer diagnoses, treatments, and outcomes. Maintaining and updating vocabulary is crucial for ensuring consistency in these datasets and enabling meaningful analysis. In the absence of direct integration of Vocab Workgroup into the vocabulary team, the responsibility for oncology vocabulary updates and maintenance falls to the broader community and the community contribution mechanism. Due to the extensive scope of this task, prioritization becomes essential to avoid addressing less critical issues or failing to address vital ones. By anchoring vocabulary work to real-world use cases, efforts can be strategically directed toward areas with the greatest impact, ensuring resources are deployed effectively and efficiently.

Methods

To address the challenges of oncology vocabulary updates and maintenance, a use case-driven process was developed to identify and prioritize vocabularies needing refinement. This process involves maintaining a dynamic list of oncology use cases, which is systematically updated twice a year. These use cases guide the prioritization of vocabulary updates, ensuring alignment with real-world clinical and research needs.

	Base Dx	Metastasis	Stage	Grade	Lymph nodes	Others (specify)	-Omics	Regimens	Radiation	Surgery	Extent	Dynamic	Episode of care	Death
Use case requirement	0.93	0.57	0.66	0.13	0	0	0.38	0.46	0.16	0.08	0.11	0.39	0.1	0.56
Vocab readiness	1	1	1	1	0.5	0.5	1	1	0.3	0.5	0.9	0.9	1	1
Model readiness	1	1	1	1	1	1	1	1	0.1	1	1	1	1	1
Available data/algorithm	0.77	0.65	0.79	0.69	0.48	0.58	0.40	0.69	0.50	0.62	0.46	0.35	0.31	0.69

Additionally, a standardized, privacy-preserving method was established to assess data readiness across the network, which also influences the priorities of vocabulary changes. This methodology incorporates three targeted queries focusing on general oncology concepts, genomic biomarkers, and episode-level data, such as progression and treatment

lines. These queries are executed locally, producing summary statistics while safeguarding patient-level information.

Vocabulary issues are resolved through two approaches: standard vocabulary updates through the community contribution mechanism or, when necessary, short-term patches of the data tailored to address urgent issues necessary to participate in an ongoing study.

Results

The approach is successfully implemented (Oncology use case repo and https://oncology.ohdsi.org/). The first community-driven vocabulary update based on this process will be released on July 8 and will be available on oncology.ohdsi.org/vocabularies. iCAN NSCLC studyathon requirements derived the prioritization of the work. This approach proved successful during the iCAN NSCLC Studyathon, where readiness improved significantly and sites were able to consistently implement phenotype and treatment line logic.

Conclusion

This process establishes a framework for regular vocabulary updates and maintenance within oncology. It provides a scalable approach to improving data quality based on real-world use cases. By aligning vocabulary updates with analytic needs, this method ensures efforts remain focused and effectively address clinical and research priorities.