

From repair to roadmap: history and future direction of the OHDSI Vaccine Vocabulary Workgroup

Alexander Davydov, MD¹, Jie Zheng, PhD², Qi Yang, MD, PhD³, Aisyah Yu Lin, PhD⁴, Lixia Yao, PhD⁵,
Yongqun He, PhD², OHDSI Vaccine Vocabulary WG

¹Thesaurus Health, Inc., New York, NY, USA; ²University of Michigan Medical School, Ann Arbor, MI, USA; ³IQVIA, Inc., King of Prussia, PA, USA; ⁴OntoData Research and Solutions, LLC, Bethesda, MD USA; ⁵Polygon Health Analytics LLC, Philadelphia, PA, USA

Background

Currently, vaccine studies across data sources are time- and resource-consuming as vaccine representation is very heterogeneous across vocabularies and vaccine capture is very diverse across data sources. There is no single vocabulary to meaningfully serve as a reference standard that reconciles these differences. CVX is pragmatic and largely flat; RxNorm/RxNorm Extension are ingredient- and product-centered and often too granular for simple vaccine cohorts; Anatomical Therapeutic Chemical (ATC) Classification is incomplete for vaccine combinations and evolving products; procedure vocabularies may carry partial vaccine information; and mapping to Standard can lose or add clinically relevant attributes. The consequence is not only usability difficulty in ATLAS or Athena, but also a research problem: concept sets for "all HPV vaccines," "cell-based influenza vaccines," "mRNA COVID-19 vaccines," or vaccine-season strata are hard to construct reproducibly.

The Vaccine Vocabulary Workgroup (WG) was formed in 2021 to make vaccine exposure computable for observational studies of uptake, comparative effectiveness, and safety.^{1,2} This report synthesizes the WG path and proposes a 2026 roadmap for moving from problem repair to a use-case-driven, ontology-guided classification approach.³⁻⁵

Methods

We reviewed WG proposals, meeting notes, decision matrices, Objectives and Key Results (OKRs), Symposia abstracts, published studies, submission materials, lessons learned after Vocabulary WG review, Vocabulathon notes, Onto-OMOP plans, and IQVIA/N3C use case materials. We abstracted each activity by the decision it changed rather than by every deliverable: problem definition, candidate design pattern, implementation evidence, feedback, and next action. Published outputs anchored the timeline, while OKRs and use-case materials informed the roadmap and open questions.

Results

The history falls into two eras (**Table 1**). In the pre-VO (VO: Vaccine Ontology) era, the WG compared several ways to force coherence onto an inconsistent source landscape: integrating CVX and RxNorm/RxNorm Extension into one hierarchy; adding generic vaccine ingredients; decomposing vaccine concepts into attributes and using formal concept analysis; manually curating a new hierarchy; and adjusting RxNorm ingredient design. These explorations made the design tension explicit. A general hierarchy was essential for users, but generic ingredients lost detail; decomposition improved transparency but required difficult attribute choices; manual top-down curation produced a research-friendly tree but raised governance and adoption concerns. In parallel, a semi-automated mapping-quality study showed that mapping defects were not hypothetical: among 2,087 potential vaccine mapping inconsistencies, expert review of 200 candidate cases confirmed 95 valid issues.³

The VO era began when the WG converged on Vaccine Ontology as an existing, community-based semantic backbone rather than a brand-new OHDSI-only vocabulary.^{4,5,8} The 2023 CVX-to-VO work established a repeatable pattern: automatic candidate matching, manual verification, annotation, and addition of missing VO terms. The 2024 RxNorm/RxNorm Extension work extended that pattern to thousands of terms and represented attributes such as pathogen or disease, vaccine platform, dose form, brand, and flu season as queryable ontology structure.^{6,7} The first 2024 VO-to-OHDSI submission was a turning point. Feedback showed that duplicating existing CVX/RxNorm/RxNorm Extension concepts was less valuable than contributing simple, curated classification terms and relationships that OHDSI tools can use. The WG therefore shifted from "import VO" to "use VO to generate and

govern a practical classification layer."

By 2025-2026, the hybrid approach was central: bottom-up VO enrichment from all OHDSI vaccine concepts, constrained by top-down, community-prioritized use cases. The Onto-OMOP pipeline⁹ provides the technical path for extracting classification terms and submission-ready relationships, while the 2026 OKRs define the social path: present the refined model to the Vocabulary WG, agree on two to three pilots, test a pre-release, evaluate results in IQVIA and N3C OMOP data, execute a community contribution, and prepare a manuscript.⁹⁻¹¹ Current open questions are deliberately practical: which classifications should be first, should some of the concepts be Standard, how much hierarchy is enough, when should attributes and their combinations become instantiated classes, how will QA/QC and lifecycle governance work, and which use cases prove clinical value?

Table 1. Vaccine Vocabulary Workgroup milestones and 2026 implications

Year/phase	Main question	Decision or achievement	2026 implication
2021 problem formation	What blocks vaccine real-world-data studies?	Mixed granularity, incomplete hierarchy, source-to-standard loss, domain ambiguity, combination support, and duplicate standards were articulated.	Requirements start from user mapping challenges and concept-set behavior, not source vocabulary boundaries.
2021-2022 design exploration	Integrate existing vocabularies or create a new structure?	Existing hierarchy integration, generic ingredients, decomposition/formal concept analysis, manual hierarchy, and ingredient design surgery were compared.	No single structural fix solved both usability and semantic fidelity.
2022 quality evidence	Are mapping issues detectable at scale?	The term-difference-pair study generated 2,087 candidates; 95 of 200 reviewed samples were valid issues.	Implementation needs automated QA plus expert review.
2023 VO adoption	Can an existing ontology carry vaccine semantics?	CVX-to-VO mapping used automatic matching with manual curation and processed the CVX vaccine set. RxNorm/RxNorm Extension concepts were prepared as the second submission to the VO	VO became the semantic backbone for various source terminologies.
2024 submission feedback	What should be contributed to OHDSI?	Feedback prioritized curated classifications and relationships over duplicate concepts.	Shift from broad VO import to curated, use-case-driven static classification/hierarchy.
2025-2026 implementation	How can the solution become operational?	Onto-OMOP, Vocabulathon, OKRs, IQVIA/N3C pilots, and the June 2026 Vocabulary WG presentation define the next workstream.	Hybrid approach: bottom-up completeness via attribution plus top-down use-case curation.

Conclusions

The WG progress is best understood as a shift from vocabulary repair to vocabulary architecture. The early work identified why vaccine concepts resist an existing drug-domain hierarchy; the VO work converted vaccine attributes into computable structure; and the 2024-2026 feedback cycle narrowed the solution to a curated, user-facing classification layer grounded in real use cases. For the OHDSI community, the expected 2026 deliverable is not a wholesale ontology import, but a maintainable pathway for finding, grouping, and studying vaccines across heterogeneous data. The same pattern may become a reusable model for bringing other highly specified biomedical ontologies into OHDSI vocabularies.

References

1. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and

- Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574-8.
2. Reich C, Davydov A, Ryan PB, et al. OHDSI Standardized Vocabularies-a large-scale centralized reference ontology for international data harmonization. *J Am Med Inform Assoc*. 2024;31(3):583-590. doi:10.1093/jamia/ocad247.
 3. Abeysinghe R, Black A, Kaduk D, Li Y, Reich C, Davydov A, et al. Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method. *J Biomed Inform*. 2022;134:104162. doi:10.1016/j.jbi.2022.104162. PMID:36029954.
 4. Zheng J, Lin AY, Huffman A, et al. VO: the Vaccine Ontology. *Sci Data*. 2026. doi:10.1038/s41597-026-07298-w. PMID:42062309.
 5. Pan Y, Manuel W, Abeysinghe R, Zheng J, Davydov A, Yang Q, et al. Mapping and harmonization of CVX vaccine terms to the Vaccine Ontology. *bioRxiv*. 2025 Jul 18:2025.07.15.664501. doi:10.1101/2025.07.15.664501. PMID:40791563.
 6. Pan Y, Manuel W, Abeysinghe R, Hao X, Davydov A, Yang Q, et al. Harmonization of OMOP vaccine-related vocabularies through the Vaccine Ontology. OHDSI Global Symposium. 2023 Oct 20-22; East Brunswick, NJ, USA.
 7. Zheng J, Li X, Zhang X, Manuel W, Lin T, Liu L, et al. Classification of RxNorm and RxNorm Extension vaccine-related terms in the Vaccine Ontology. OHDSI Global Symposium. 2024 Oct 22-24; East Brunswick, NJ, USA.
 8. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007;25(11):1251-1255. doi:10.1038/nbt1346.
 9. Zheng J, Davydov A, Ostroplets A, Yang Q, He Y. Onto-OMOP: an automatic pipeline for generating classification level terms and relationships for OHDSI vaccine concepts using the Vaccine Ontology. OHDSI Global Symposium. 2025; showcase 140.
 10. Jackson RC, Balhoff JP, Douglass E, Harris NL, Mungall CJ, Overton JA. ROBOT: a tool for automating ontology workflows. *BMC Bioinformatics*. 2019;20:407. doi:10.1186/s12859-019-3002-3.
 11. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y. OntoFox: web-based support for ontology reuse. *BMC Res Notes*. 2010;3:175. doi:10.1186/1756-0500-3-175.