

Extensible Refactoring of OHDSI Cancer Modifier "Nodes" concept class replicating SNOMED CT Lymph Node Anatomy behavior

Vlad Korsik¹; Iryna Zherka¹; Maryia Khitrun²; Asieh Golozar³; Christian Reich³

¹ Omophoria Inc., ² Odysseus, an EPAM company, ³ Nemesis Health Inc.

Background

Lymph node involvement is a core determinant of cancer staging, prognosis, treatment planning, and longitudinal disease characterization. In real-world oncology data, nodal disease is captured with variable granularity across cancer registries, pathology reports, imaging, structured EHR fields, and narrative documentation. The current OMOP Oncology Cancer Modifier "Nodes" concept class contains historically accumulated terms from heterogeneous sources and is largely represented as a flat list. This limits consistent ETL mapping, descendant-based cohort definition, semantic validation, and reuse across analytic use cases. A further challenge is that "regional" versus "distant" nodal spread is not an intrinsic property of a lymph node location alone; it depends on tumor context, primary site, and staging convention. Therefore, anatomically specific nodal findings should not be forced into regional or distant categories at ETL time when that semantics is not explicitly available.

Methods

We propose a location-first refactoring of the Cancer Modifier "Nodes" hierarchy using the SNOMED CT Body Structure lymph node anatomy axis as the primary organizing backbone. The model introduces a general root concept, "Spread to lymph node" (concept_id 36768587), and a curated hierarchy of location-only modifier concepts expressed as "Spread to [lymph node location]". Hierarchical "Is a" relationships in this layer mirror SNOMED CT lymph node anatomical subsumption after excluding non-target structures such as follicles, capsules, lymphatic vessels, and other concepts that do not represent clinically mappable nodal locations. Regional and distant semantics are modeled as an orthogonal typing axis using "Regional spread to lymph node" (concept_id 36769243) and "Distant spread to lymph node" (concept_id 36769269). Typed leaf concepts, such as "Regional spread to [lymph node location]" or "Distant spread to [lymph node location]", are represented as polyhierarchical descendants of both their spread-type archetype and their corresponding location-only concept. SNOMED CT is treated as a robust and computable anatomical backbone, not as an exhaustive source for every oncology-specific topographic abstraction. When source vocabularies, registries, or clinical workflows require generic, composite, or convention-specific nodal regions that are not adequately represented by the SNOMED CT lymph node anatomy axis, OHDSI-managed Topography concepts can be introduced and linked to one or more appropriate SNOMED CT lymph node Body Structure anchors. These concepts would replicate the same semantic pattern: a location-only modifier, optional regional/distant typed descendants, and lateral definitional relationships such as "Has finding site" and "Has type".

Results

The proposed design transforms the "Nodes" concept class into a scalable, ontology-aligned vocabulary structure that supports both high-level and granular analytics. The SNOMED CT lymph node body-structure subtree (anchoring SNOMED CT concept_code 59441001) provides the main substrate for constructing the location hierarchy, while OHDSI-curated Cancer Modifier Topography concepts provide a controlled community extension mechanism for gaps, composite nodal basins, or oncology-specific topographic conventions. This mixed model preserves computable roll-up through SNOMED-anchored anatomy wherever possible while allowing the OHDSI community to extend the vocabulary in a governed and reproducible way. The resulting hierarchy supports standard OMOP descendant expansion for queries such as "any lymph node involvement", "abdominal lymph node involvement", "cervical lymph node involvement", or level-specific nodal disease. Because regional/distant semantics are separated from location, ETL can preserve source meaning without adding unsupported staging assertions. Records with explicit spread-type semantics are mapped to typed concepts; records containing only anatomical location are mapped to the most

specific location-only concept. Downstream regional-versus-distant interpretation can then be computed when tumor context, primary site, episode information, and site-specific staging rules are available. The model also enables systematic vocabulary quality control: each nodal spread concept can be checked for an anatomical anchor, each location hierarchy can be compared against its SNOMED CT or OHDSI Cancer Modifier Topography parentage, and each typed concept can be validated for both a location parent and a spread-type parent. Existing legacy "Nodes" concepts can be migrated to the nearest location-only or typed concept, with deprecated concepts remapped where appropriate and backward-compatible roll-ups preserved for existing cohort definitions.

Conclusions

This refactoring provides an extensible, ETL-safe, and analytically scalable representation of lymph node involvement in OMOP Oncology. By combining the SNOMED CT lymph node anatomy axis with governed OHDSI Topography extensions, the approach avoids overclaiming anatomical coverage while enabling the community to add missing or oncology-specific nodal abstractions using a shared semantic model. Separating location from regional/distant spread type preserves source granularity, supports computable reasoning, and improves consistency across registry, pathology, imaging, and EHR-derived oncology data. The proposed model is compatible with current OMOP oncology modifier and episode-based representation, is intended for review and implementation through the OHDSI Oncology and Vocabulary Working Groups, and may provide a reusable foundation for future OMOP-compatible modeling of radiotherapy nodal regions.

References:

1. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574-578. PMID: 26262116.
2. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc.* 2012;19(1):54-60. doi:10.1136/amiajnl-2011-000376.
3. OHDSI Oncology Working Group. OncologyWG repository and wiki: OMOP Oncology modeling, cancer modifiers, episodes, and ETL resources. GitHub. <https://github.com/OHDSI/OncologyWG>. Accessed June 5, 2026.
4. SNOMED International. SNOMED CT documentation. <https://docs.snomed.org/>. Accessed June 5, 2026.
5. OHDSI. OMOP Common Data Model documentation and Standardized Vocabularies conventions. <https://ohdsi.github.io/CommonDataModel/>. Accessed June 5, 2026.