

Beyond Initial Integration: Evolution of an Oncology Data Lake Supporting Longitudinal Cancer Research

Authors:

Farnoosh Sheikhi^{1,2}, Shikha Kothari^{1,2}, Natasha Flowers^{1,2}, Hannah Cooper^{1,2}, Jose Posada^{1,2}, Somalee Datta^{1,2}, Solomon Henry^{1,2}, Deepa Balraj^{1,2}, David Love^{1,2}, Mina Satoyoshi^{1,2}, Joe Mesterhazy^{1,2}, Darren Guan^{1,2}, Smita Limaye^{1,2}, Alvaro Alvarez^{1,2}, Jay Chen^{1,2}, Priya Desai^{1,2}

¹ Stanford Health Care, ² Stanford School of Medicine

Background

The development of AI-enabled tools and multimodal biomarkers for precision oncology requires integrated datasets that capture the longitudinal cancer patient journey across clinical, molecular, imaging, and pathology domains.^{1,2} While advances in quantitative biomarker development have accelerated data-driven cancer research, most approaches remain limited to individual data modalities due to challenges in data integration and harmonization.^{2,3} To address these gaps, Stanford Medicine established the VISTA Oncology Data Lake, an ARPA-H funded initiative designed to create a scalable oncology research ecosystem by linking OMOP-standardized electronic health records (EHRs) with specialized oncology data sources.³⁻⁷ Building upon an initial foundation of EHR and cancer registry data, we describe the continued evolution of the platform through expansion of multimodal data assets, longitudinal treatment representations, and AI-ready research infrastructure as relevant to the OHDSI⁸ community.

Methods

The Stanford VISTA Oncology Data Lake^{4,7,9-11} was developed within the STANford medicine Research data Repository (STARR), Stanford Medicine's enterprise research data ecosystem. Built upon OMOP-standardized clinical data,^{8,9} cancer registry¹⁰ information, imaging metadata¹¹ and molecular testing results,¹² the platform underwent substantial expansion between November 2025 and August 2026 to support increasingly comprehensive multimodal oncology research.

Major enhancements included migration from OMOP Common Data Model (CDM) version 5.3 to 5.4 along with updated OHDSI vocabularies, incorporation of approximately 35,000 whole slide imaging (WSI) assets,¹³ integration of Epic Aura^{14,15} genomic data into OMOP variant occurrence representations,^{16,17} enrichment of pathology metadata from Epic Beaker, and implementation of longitudinal oncology treatment protocols using OMOP episode table.^{18,19}

Vital status ascertainment was also strengthened by incorporating external death records from the California Department of Public Health (CDPH) together with Epic Care Everywhere data, which captures death outcomes occurring outside the Stanford health system. In parallel, TiDE 2.0,^{20,21} a production-grade PHI-scrubbing framework that combines multiple anonymization strategies with cryptographic techniques, date jittering (± 30 days), and machine learning-based

entity recognition, was deployed under secure governance frameworks and data use agreements to facilitate AI-ready access to clinical narratives and associated oncology data assets while managing re-identification risk.

To accommodate these emerging oncology domains that are not yet fully represented within OMOP standards, we implemented a staged harmonization and schema extension strategy. Newly onboarded data assets are initially maintained within source-aligned auxiliary structures while preserving linkage to the OMOP ecosystem through shared patient and clinical relationships. This approach enables timely research access while supporting progressive semantic harmonization and future alignment with evolving oncology data standards. The platform further leverages modular ETL pipelines,²² provenance-aware architecture, and reusable validation frameworks to support scalable growth and long-term sustainability (Figure 1).

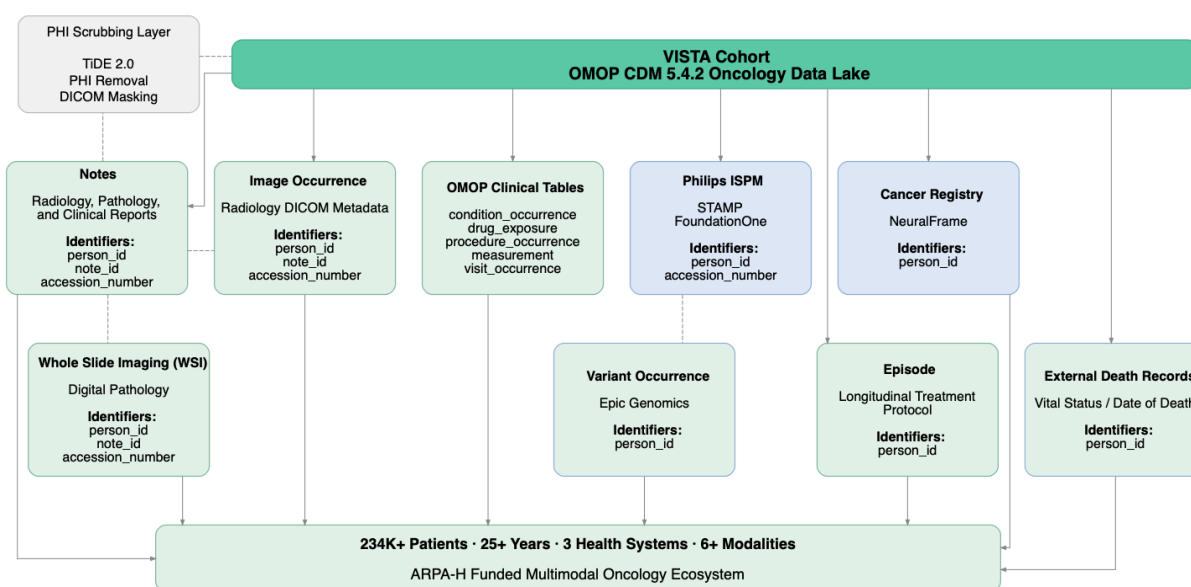


Figure 1: From Source to Standard: OMOP-Unified Oncology Data Lake

Results

As of the August 2026 release, the VISTA Oncology Data Lake supports multimodal linkage across OMOP-standardized clinical data, cancer registry records, genomics, digital pathology, imaging, and longitudinal treatment data for more than 234,342 oncology patients.⁴

From 234,342 patients, 233,143 (99.5%) have clinical notes, 208,081 (88.8%) have imaging metadata, 63,619 (27.2%) have longitudinal treatment episodes, 3,747 (1.6%) have genomic variant data, and 2,379 (1.0%) have whole slide images (Figure 2).

The data lake provides integrated access to structured clinical information, molecular testing results, pathology metadata, whole slide imaging assets, and longitudinal treatment trajectories within a unified patient-centered framework. Beyond patient-level integration through person identifiers, the platform enables finer-grained linkage across oncology modalities: 4,085,765 image records link to specific notes via note identifiers, 2,838,339 image records link to visit occurrences, 3,798 whole slide imaging assets link directly to pathology reports via note identifiers (1.62 per WSI patient), and 4,280 genomic variant records link to visit occurrences. Whole slide imaging assets and image metadata are linked to pathology reports through note identifiers and accession-based relationships, enabling specimen-level integration between digital pathology, clinical narratives, and imaging data. Genomic data from Epic Aura are represented in the custom OMOP variant_occurrence table,^{16,17} extending prior molecular data integration (Philips ISPM)¹² and enabling standardized linkage between genomic variants, clinical observations, and longitudinal patient outcomes. In addition, migration to OMOP CDM v5.4 and updated OHDSI vocabularies enhanced the support for oncology-specific concepts, resulting in improved standardization and mapping of cancer staging information across the research ecosystem.

The resulting architecture preserves connectivity across heterogeneous oncology domains while supporting incremental onboarding of future data modalities.

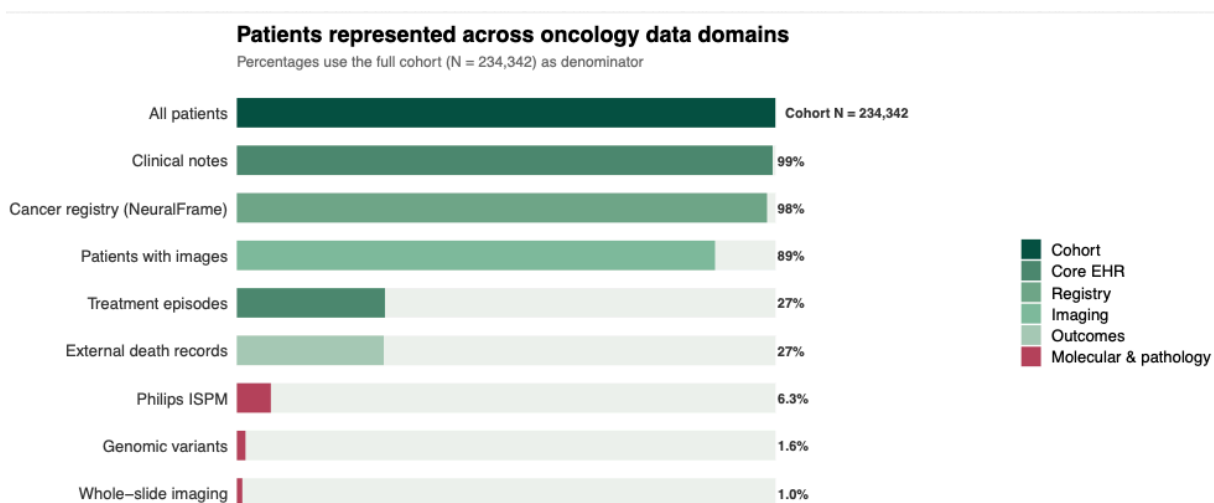


Figure 2: Data Modality Distribution

Conclusions

The VISTA Oncology Data Lake has evolved from an EHR- and registry-focused resource into a multimodal oncology research ecosystem supporting longitudinal cancer research, precision oncology, and AI-driven discovery. By integrating clinical, genomic, pathology, imaging, and treatment data within an OMOP-centered architecture, the platform addresses critical infrastructure challenges associated with multimodal oncology research. The staged harmonization and schema extension approach enables scalable incorporation of emerging data sources while maintaining alignment with the broader OHDSI ecosystem, providing a sustainable foundation for next-generation precision oncology investigations.

Acknowledgements:

This work is supported partly by the U.S. Advanced Research Projects Agency for Health (ARPA-H), as well as by a grant from Accenture.

Documentation:

Oncology Data Lake Documentation:

<https://susom.github.io/starr-oncology-data-lake-arpah/about.html>

TiDE 2.0 Core Repository:

<https://github.com/susom/tide2-core>

References:

1. Boehm K, Khosravi P, Vanguri R, Gao J, Shah S. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer*. 2022 Feb;22(2):114-126. doi: 10.1038/s41568-021-00408-3.
2. Rajesh M, Prema S, Siva Balan RV, Nilabar Nisha U, Chinnu S, Moorthy U. Multimodal artificial intelligence for early cancer detection via liquid biopsy, imaging, and clinical records. *Intelligence-Based Medicine*. 2026;14:100395. doi:10.1016/j.ibmed.2026.100395.
3. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. *arXiv*. 2020; arXiv:2003.10534. Available from: <https://arxiv.org/abs/2003.10534>
4. Stanford Medicine. VISTA Oncology Data Lake Project Documentation [Internet]. Stanford (CA): Stanford University; 2026 [cited 2026 Jun 8]. Available from: <https://susom.github.io/starr-oncology-data-lake-arpah/about.html>
5. Callahan A, Ashley E, Datta S, Desai P, Ferris TA, Fries JA, Halaas M, Langlotz CP, Mackey S, Posada JD, Pfeffer MA, Shah NH. The Stanford Medicine data science ecosystem for clinical and translational research. *JAMIA Open*. 2023 Oct;6(3):ooad054. doi:10.1093/jamiaopen/ooad054.

6. Mesterhazy J, Olson G, Datta S. High performance on-demand de-identification of a petabyte-scale medical imaging data lake. *arXiv*. 2020; arXiv:2008.01827. Available from: <https://arxiv.org/abs/2008.01827>
7. Stanford Medicine Research IT. STARR Data Types [Internet]. Stanford (CA): Stanford University; [cited 2026 Jun 5]. Available from: <https://starr.stanford.edu/data-types>
8. Observational Health Data Sciences and Informatics (OHDSI). The Book of OHDSI [Internet]. Available from: <https://ohdsi.github.io/TheBookOfOhdsi/>
9. Stanford Medicine Research IT. STARR OMOP [Internet]. Stanford (CA): Stanford University; [cited 2026 Jun 5]. Available from: <https://starr.stanford.edu/data-models/omop>
10. NeuralFrame. KACI – Knowledge-Augmented Cancer Insights [Internet]. Palo Alto (CA): NeuralFrame Inc.; [cited 2026 Jun 5]. Available from: <https://neuralframe.com/kaci/#kaci>
11. Park WY, Jeon K, Schmidt TS, You SC, Nagy P. Development of medical imaging data standardization for imaging-based observational research: OMOP Common Data Model extension. *J Imaging Inform Med*. 2024;37(2):899–908. doi:10.1007/s10278-024-00982-6. PMID:38315345.
12. Philips. IntelliSpace Precision Medicine Oncology [Internet]. Philips Healthcare. Available from: <https://www.philips.com/healthcare/resources/landing/intellispace-precision-medicine-oncology>
13. Morgan-Cooper H, Fanoush A, Desai P. Enabling Scalable Multimodal AI Research in STARR: Integrating Whole Slide Imaging Pathology Data with Clinical Information [Internet]. OHDSI 2026 Global Symposium Showcase; 2026. Available from: <https://www.ohdsi.org/2026showcase-74>
14. Epic Systems Corporation. Healthcare Intelligence [Internet]. Epic Systems Corporation; [cited 2026 Jun 5]. Available from: <https://www.epic.com/software/healthcare-intelligence/>
15. Epic Systems Corporation. Life Sciences – Precision Medicine, Specialty Diagnostics, and Devices with Aura [Internet]. Verona (WI): Epic Systems Corporation; [cited 2026 Jun 5]. Available from: <https://www.epic.com/software/life-sciences/>
16. Observational Health Data Sciences and Informatics (OHDSI). Genomic-CDM: Repository for development of the genomic module of the CDM [Internet]. GitHub; 2025 [cited 2026 Jun 8]. Available from: <https://github.com/OHDSI/Genomic-CDM>
17. Shin SJ, You SC, Park YR, Roh J, Kim JH, Haam S, Reich CG, Blacketer C, Son DS, Oh S, Park RW. Genomic Common Data Model for seamless interoperation of biomedical data in clinical practice: Retrospective study. *J Med Internet Res*. 2019 Mar 26;21(3):e13249. doi:10.2196/13249.
18. Observational Health Data Sciences and Informatics (OHDSI). OMOP Common Data Model v5.4: Episode [Internet]. Available from: <https://ohdsi.github.io/CommonDataModel/cdm54.html#episode>
19. Belenkaya R, Gurley MJ, Golozar A, Dymshyts D, Miller RT, Williams AE, Ratwani S, Reich C. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *JCO Clin Cancer Inform*. 2021;5:110–20. doi:10.1200/CCI.20.00079

20. Stanford Medicine. tide2-core: A clinical data de-identification and anonymization toolkit [Internet]. GitHub; 2026 [cited 2026 Jun 8]. Available from: <https://github.com/susom/tide2-core>
21. Posada J, Datta S, Desai P. TiDE 2.0: A Model-Agnostic, Utility-Preserving Open Source Tool for Distributed Clinical Text De-identification [Internet]. OHDSI 2026 Global Symposium Showcase; 2026. Available from: <https://www.ohdsi.org/2026showcase-112>
22. Flowers N, Posada J, Morgan-Cooper H, Sheikhi F, Kothari S, Alvarez A, Desai P. OMOP ETL Design for Transparency, Quality, and Maintainability [Internet]. OHDSI 2026 Global Symposium Showcase; 2026. Available from: <https://www.ohdsi.org/2026showcase-95>