Title:

Building an Oncology Data Lake to Enable Cancer Research: Lessons Learned from a Large Academic Health System

Authors:

Shikha Kothari[1,2], Natasha Flowers[1,2], Hannah Morgan-Cooper[1,2], Farnoosh Sheikhi[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:

Cancer data is **notoriously fragmented** across pathology, radiology, genomics, treatment records, and registries. At **Stanford Medicine**, we sought to **bridge these silos** and **enable research** through the development of a multimodal **oncology data lake**[1-4]. Our goal was to build an **interoperable, research-ready** data lake grounded in real-world data **using standard data models**. This abstract summarizes the lessons we have learned, particularly in the areas of **infrastructure, data integration, harmonization** and **extensibility**— as relevant to the OHDSI[5] community.

Methods:

Our data lake was built within the framework of **STAnford medicine Research data Repository (STARR)**[1-4], Stanford's enterprise research data warehouse, which hosts both raw (unstructured) and analysis-ready (structured) **multimodal** clinical data and powers multiple downstream research and analytic datasets, all hosted on the **Google Cloud Platform (GCP)**. Within STARR, we developed **STARR Common**, a set of **domain-oriented** tables that aggregate **raw**, source-aligned data from **Epic Clarity** and **other clinical systems**. These tables **preserve** the native source structures and **streamline** downstream generation of **common data models (CDMs)** such as OMOP[5,6], PEDSnet[7], and PCORnet[8], as well as project-specific data deliveries.

To support cancer research, we extended STARR Common with a suite of Oncology Common Tables, which:

- Extract and structure **cancer registry data** from **NeuralFrame** (Stanford's internal registry used for state and national cancer reporting)[9]
- Ingest **image-level metadata** (e.g., **DICOM** series) from radiology systems (integrated into OMOP **using guidelines outlined by the OHDSI Imaging workgroup**).[10]
- Integrate **molecular mutation data** from **Philips IntelliSpace Precision Medicine (ISPM)**[11]

- Map supplementary **electronic health record (EHR)** diagnosis and treatment data from **Epic Clarity**[12]

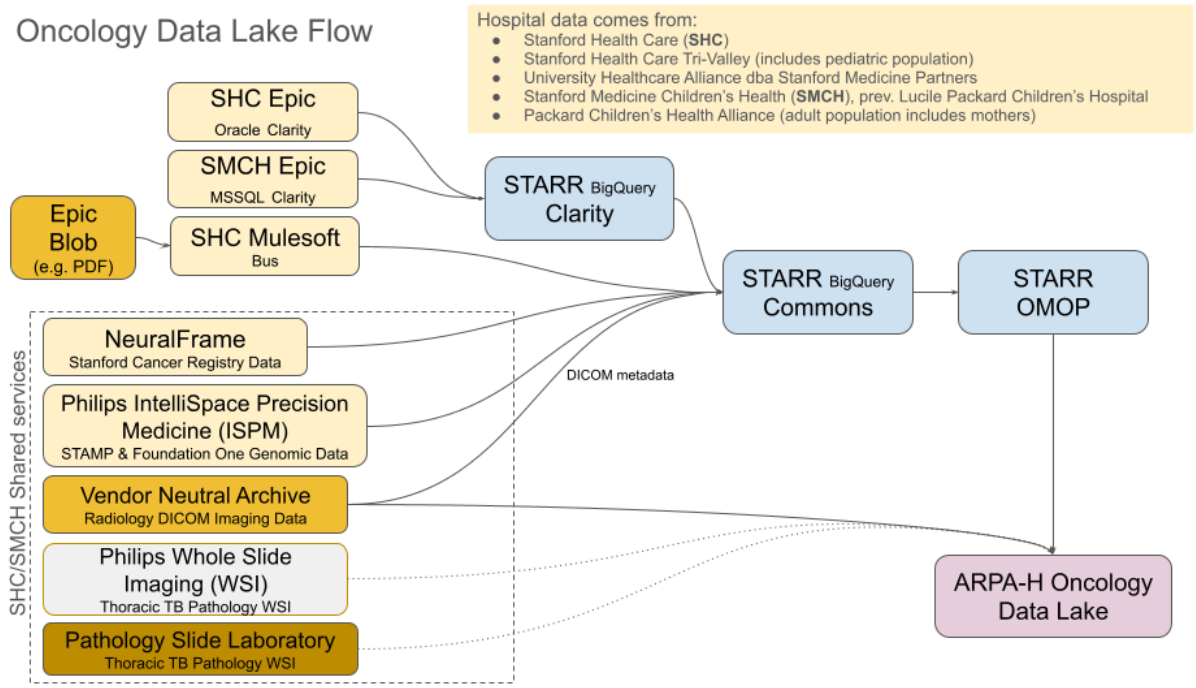

Figure 1. Oncology data lake flow

In the near future, we plan to integrate data from the new **EPIC AURA (Genomics) module**[13] and transition away from **Philips ISPM**.

We **defined our adult oncology cohort** by identifying patients who had **both** a cancer diagnosis in OMOP **and presence** in NeuralFrame. All ETL work was implemented using **modular pipelines** with **robust validation**, including **dbt tests** tailored to flag edge cases and data consistency concerns.

At present, the **Oncology Common Tables** and the **oncology OMOP** subset are **internal** and only used for this project work; **access is limited** to PIs, and delivered datasets are **PHI-scrubbed**.

Results:
Key takeaways from our development journey include:

- **Cancer registry integration required nuanced validation.** The NeuralFrame data, although structured, includes manually entered and free-text fields that necessitated custom field mapping, temporal reconciliation, and iterative quality checks.
- **Source-preserving layers like STARR Common simplify data pipeline development.** Our ability to retain raw source structure accelerated the creation of analytic-ready datasets and CDM conversions without reprocessing from scratch.

    - This **raw-data-first** layer—grouped into tables focused on various clinical as well as key oncology use cases—lets internal pipelines and the oncology OMOP subset **reuse** consistent joins and derivations **without re-extracting from source systems**.
    - Updates to raw source data (Epic Clarity, NeuralFrame cancer registry, etc) are handled just once and flow forward when we rebuild views, keeping downstream work stable and reviewable.
- **dbt-based validation supports agile iteration.** Layered, reusable dbt tests enabled us to catch event misalignments and validate assumptions quickly as pipelines evolved.[14]
    - We run dbt checks at **each stage** so the raw-to-ready path is repeatable:
        - HIPAA-oriented **age < 90 years** guardrail
        - **completeness thresholds** for key columns
        - **schema-change alerts** if our raw source feeds add unexpected columns
        - we enforce **primary-key uniqueness** and **referential (foreign-key) integrity** so **joins line up** across registry, DICOM imaging, and molecular data.
        - **unique-combination checks** are also used where appropriate to prevent duplicates.
        - **accepted values** tests help us catch new field values
        - **conditional checks** test for expressions such as whether start dates are before end dates.
- **Custom tables for oncology domains reduce downstream burden.** The modular design of Oncology Common Tables enables us to flexibly support future models (e.g., OMOP v5.4 Episode tables[15,16], MedHELM[17,18]) without duplicating ETL logic.

Conclusions:

Stanford's experience building an oncology data lake highlights the importance of **source-aligned infrastructure** and thoughtful **modular architecture**. By building reusable domain-level tables and investing in validation processes, we enabled **scalable cancer data integration** across diverse sources. These lessons may be helpful to other OHDSI collaborators seeking to support oncology research using local registry, imaging, molecular, and treatment data. We plan to expand availability of these oncology resources to the broader research community as the work matures.

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:

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

References:

1. 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.
2. 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
3. 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
4. Stanford Medicine Research IT. STARR Data Types [Internet]. Stanford (CA): Stanford University; [cited 2025 Jul 1]. Available from: https://starr.stanford.edu/data-types
5. Observational Health Data Sciences and Informatics (OHDSI). The Book of OHDSI [Internet]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/
6. Stanford Medicine Research IT. *OMOP Common Data Model* [Internet]. Stanford (CA): Stanford University; [cited 2025 Jul 1]. Available from: https://starr.stanford.edu/data-models/omop
7. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc*. 2014;21(4):602–6. https://doi.org/10.1136/amiajnl-2014-002743
8. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578–82. https://doi.org/10.1136/amiajnl-2014-002747
9. NeuralFrame. *KACI – Knowledge-Augmented Cancer Insights* [Internet]. Palo Alto (CA): NeuralFrame Inc.; [cited 2025 Jul 1]. Available from: https://neuralframe.com/kaci/#kaci
10. 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 Digit Imaging. 2024;37(2):899–908. doi:10.1007/s10278-024-00982-6. PMID:38315345.
11. Philips. IntelliSpace Precision Medicine Oncology [Internet]. Philips Healthcare. Available from: https://www.philips.com/healthcare/resources/landing/intellispace-precision-medicine-oncology
12. Epic Systems Corporation. Healthcare Intelligence [Internet]. Epic Systems Corporation; [cited 2025 Jul 1]. Available from: https://www.epic.com/software/healthcare-intelligence/
13. Epic Systems Corporation. *Life Sciences – Precision Medicine, Specialty Diagnostics, and Devices with Aura* [Internet]. Verona (WI): Epic Systems Corporation; [cited 2025 Jul 1]. Available from: https://www.epic.com/software/life-sciences/
14. dbt Labs. dbt: Analytics engineering [Internet]. Available from: https://www.getdbt.com/
15. 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

16. 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
17. Stanford Center for Research on Foundation Models. Med-HELM: Healthcare Evaluation of Language Models [Internet]. Stanford (CA): Stanford CRFM; [cited 2025 Jul 1]. Available from: https://crfm.stanford.edu/helm/medhelm/latest/
18. Posada JD, Sharma A, Nguyen M, Cook A, Altman R, Fries J, Shah NH. Cancer patient cohort retrieval from electronic health records using a hybrid ranking pipeline. arXiv [Preprint]. 2024 May 29 [cited 2025 Jul 1]; Available from: http://arxiv.org/abs/2505.23802