

Accelerating Translational Oncology Research Through a Reproducible OMOP CDM: Lessons Learned from Fred Hutch Cancer Center

Monica W. Gerber¹, Megan Stachura¹, Josh Southam¹, Sitapriya Moorthi¹, Christian Reich², Asieh Golozar², Amy Paguirigan¹

¹ Fred Hutch Cancer Center, Seattle, WA

² Nemesis Health

Background

The **Fred Hutch Cancer Center (FHCC) OMOP CDM** was built on the **Clinical and Research Data System (CARDS)**, a multimodal translational data platform created to unify and modernize management of cancer-related care data derived from Fred Hutch's adult oncology program. CARDS was developed following the 2022 merger of Seattle Cancer Care Alliance (SCCA) and Fred Hutchinson Cancer Research Center (FHCRC) and makes it possible to securely integrate data from electronic health records (EHRs), clinical laboratories, the cancer registry, imaging systems, and other sources.¹ It uses Databricks on AWS as a central, cloud-based analytics platform. To accelerate innovation across translational science in oncology, a core goal of CARDS is to use standard data models, open source tooling, and reproducible workflows, while being responsive and adaptive to evolving user needs.

Methods

Our approach to build this new OMOP CDM involved profiling user needs across Fred Hutch's translational research community. Analysis of these use cases helped determine the data needed to support the execution of translational data science projects. To inform the prioritization of transformation of data to the OMOP CDM, we calculated the percentage of use cases requiring data from distinct oncology data domains (overall and stratified by tumor type).

To inform our technical implementation of the ETL process that creates the OMOP CDM, high-level requirements were gathered from internal data science and engineering teams. We created workflow guidelines that specified our version control process and code review guidelines. We created markdown templates for repeatable tasks such as exploratory data analysis and table requirements specifications.

We utilized open-source tooling on Databricks to set up our pipeline. Declarative Automation Bundles (DABs) define the jobs for updating OMOP tables and run data checks on a regular basis.² As a complement to Data Quality Dashboard (DQD) for data quality monitoring, the DQX Python library enables us to define, monitor, and address data quality issues proactively.^{3,4}

Results

The case analysis confirmed that, to adequately address research questions from our translational community, the OMOP CDM will need to continue to integrate data from multiple sources (Figure 1).

- The EHR data meets critical data domains such as base diagnosis (82% of use cases), and chemotherapy regimens (71% of use cases) and provides a strong foundation for our OMOP CDM instance.
- Integration of cancer registry data will provide more robust histology and topography concepts for cancer diagnoses.
- Forty-one percent of use cases require molecular genetic and biomarker data, which will require integration of data from diverse data sources, from vendors to clinical laboratory systems, and mapping of measurement concepts to the OMOP Genomic vocabulary.

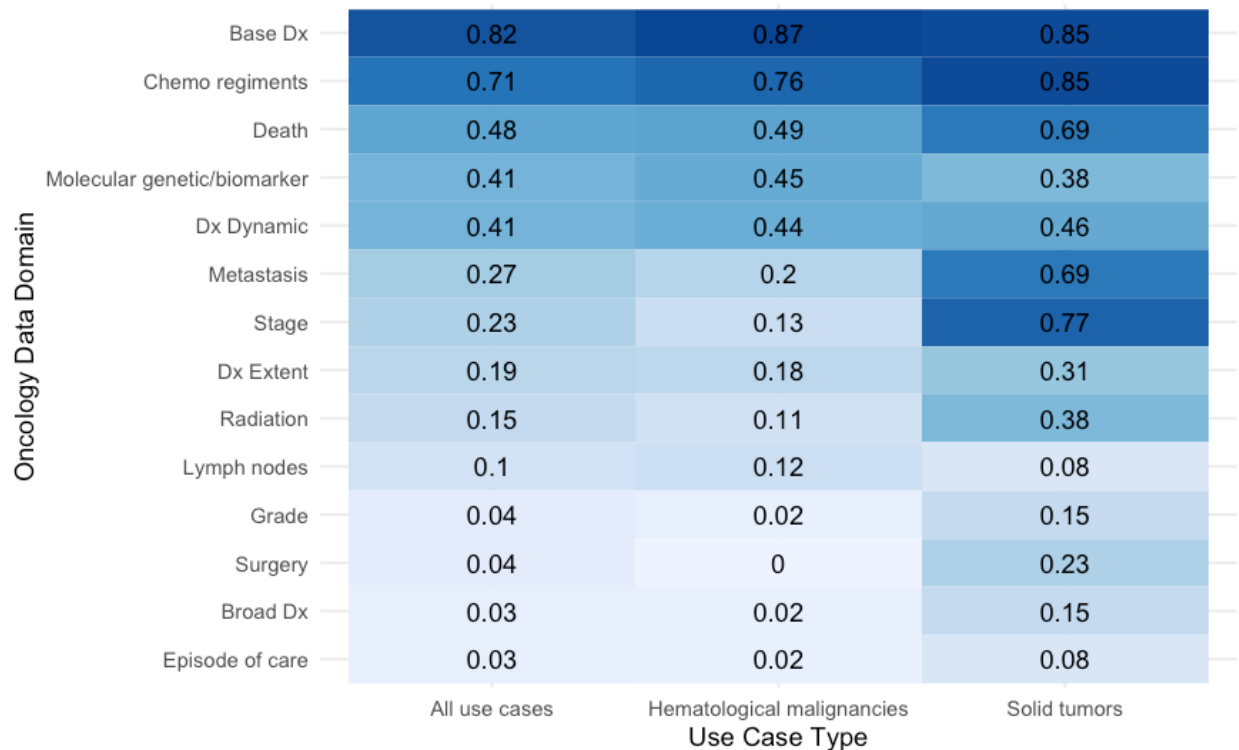


Figure 1. Proportion of use cases requiring data from each oncology data domain, by tumor type (all use cases, $n = 76$; hematological malignancies, $n = 57$; solid tumors, $n = 15$; other, $n = 6$). Boxes are shaded by proportion of uses cases, with darker colors representing higher proportions. The proportion is displayed within each box.

By fall 2026, our core developer team transformed EHR data (Epic Clarity) from the joint adult oncology program (clinically integrated hospitals and outpatient clinics across Fred Hutch and University of Washington Medical Center) to 13 clinical and health system data tables. Results from DQD indicate data quality is high (98% overall pass rate; Figure 2). Comparison of the distribution of cancer condition concepts at Fred Hutch to the distribution of cancer concepts across the FALCON network illustrates that the prevalence of cancer types at Fred Hutch is similar to that of the network, with the exception of lung cancer (Figure 3).

	Verification				Validation				Total			
	Pass	Fail	Total	Pass %	Pass	Fail	Total	Pass %	Pass	Fail	Total	Pass %
Plausibility	254	7	261	97%	234	8	242	97%	488	15	503	97%
Conformance	635	2	637	100%	66	0	66	100%	701	2	703	100%
Completeness	238	7	245	97%	12	5	17	71%	250	12	262	95%
Total	1,127	16	1,143	99%	312	13	325	96%	1,439	29	1,468	98%

Figure 2. Data Quality Dashboard (DQD) results from the 1,468 total checks applicable to the FHCC OMOP CDM instance. The 29 failed checks are primarily related to (1) columns we have not yet fully populated, (2) plausible gender for conditions, (3) record completeness for persons, (4) plausibility of date sequences.

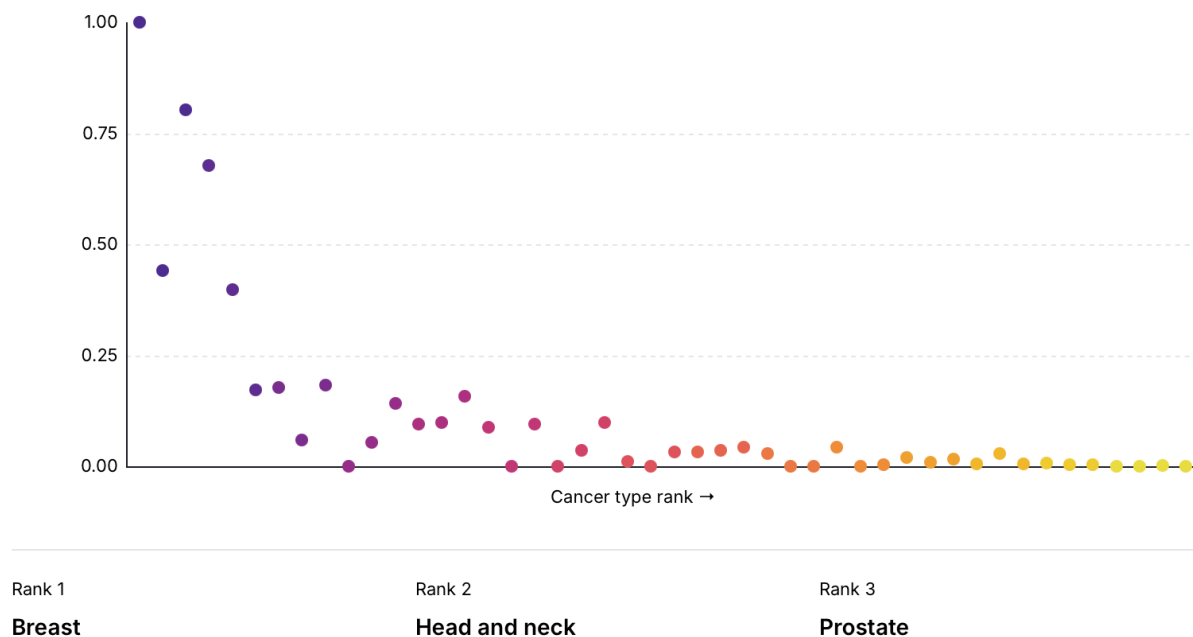


Figure 3. Distribution of cancer types at Fred Hutch. Cancer types are ordered along the x-axis by FALCON network-wide prevalence (most prevalent leftmost).⁵ The 5 most common cancer types in the Fred Hutch OMOP CDM are: (1) breast, (2) head and neck, (3) prostate, (4) lung and respiratory track, and (5) large intestine.

Essential insights from our process of building a new OMOP CDM instance include:

- Team standards and coordination were essential for moving quickly, documenting decision-making, and adapting to new technical infrastructure. Our open-source template GitHub repository (<https://github.com/FredHutch/fhcc-omop-template>) illustrates our process and is available for other health systems to “fork” and adapt for their own use.⁶ This repository includes:
 - A recommended repository structure for organizing code and documentation

- Templates for table requirements and data profiling/exploratory data analysis tasks, gathering high-level requirements, and specifying technical implementation
- Our team workflow and code guidelines
- Examples of how to set up DQX checks and monitoring
- An example of how to set up jobs using DABs
- Our pipeline implementation on Databricks allows for secure, reproducible, automated workflows that meet downstream data user needs.
 - Defining jobs: DABs provide an infrastructure-as-code (IaC) approach which defines and manages our jobs, variables, and alerts through YAML for version control and transparency.
 - ID generation: A hash function (xxhash64) provides a deterministic ID generation method that maintains consistency in the IDs between weekly fully table refreshes.
 - Data quality: DQX runs with weekly table refreshes. Binary (pass/fail) checks on individual records are defined in a YAML.
 - Table refreshes: Full table refreshes run on a weekly basis to meet user needs of up-to-date data.

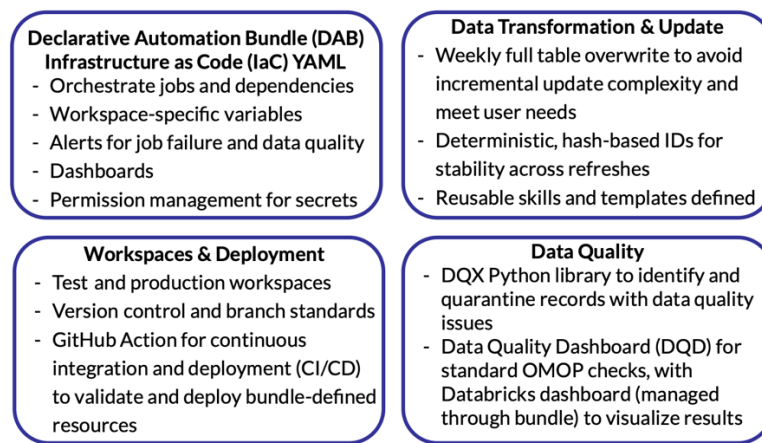


Figure 4. Key pillars of our ETL pipeline designed for reproducible, automated workflows to deliver reliable, high-quality data to meet oncology needs.

Conclusion

Launching a new OMOP CDM transformation is a major initiative but essential for accelerating innovation in the translational oncology space. Considering high-level technical needs alongside research use cases gave us a framework for prioritizing tasks and making decisions about technical implementation. Documenting team workflows, creating reusable templates, and using an IaC approach to our pipeline allowed us to move agilely and scale for integration of additional data sources and use cases. We've shared our framework in an open source GitHub repository for other OHDSI collaborators and to contribute to the Open Science goals of promoting reproducibility, transparency, sharing, and usability of scientific innovation.

GitHub repository: <https://github.com/FredHutch/fhcc-omop-template>

Acknowledgements: Thank you to the many collaborators at the OHDSI Global Symposium in 2025, especially members of the Oncology Working Group, for providing advice and encouragement as we began this project.

References

1. Marmor J. A united vision: a new working relationship between UW Medicine and Fred Hutch creates one of the nation's top adult cancer care programs [Internet]. University of Washington Magazine; 2025 Mar [cited 2026 June 4]. Available from: <https://magazine.washington.edu/feature/uw-medicine-and-fred-hutch-work-together-to-provide-a-united-vision-for-cancer-care/>
2. Databricks. What are Declarative Automation Bundles? [Internet]. Databricks; 2026 Mar 15 [cited 2026 Jun 5]. Available from: <https://docs.databricks.com/aws/en/dev-tools/bundles/>
3. Databricks Labs. DQX: Data Quality Framework [Internet]. Available from: <https://databrickslabs.github.io/dqx/>
4. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association. 2021 Oct 1;28(10):2251-7.
5. Golozar, A., et al. "146eP FALCON-bladder, from real-world evidence to impact: Piloting a global assessment of oncology guideline adoption, relevance and generalizability." ESMO Real World Data and Digital Oncology 10 (2025).
6. Lowndes JS. Forking as a worldview: how Openscapes uses the open source concept of reuse-extend-credit to support researchers and shift culture in science. AGU25. 2025 Dec 18. Available from: <https://www.openscapes.org/events/2026-03-19-leptoukh-lecture-reprise/>