Advancing Learning Health Systems Through Integrated Machine Learning Operations: A Novel Extension of the OHDSI Research Infrastructure

Boudewijn Aasman[1], Selvin Soby[1], Adil Ahmed[1], Chandra Nelapatla[1], Manuel Wahle[1], Parsa Mirhaji[2]
[1] Montefiore Medicine, [2] Albert Einstein College of Medicine at Montefiore

Background

Healthcare delivery is transitioning from protocol-driven care to dynamic, continuous learning systems. The Institute of Medicine's (IOM) definition of Learning Health Systems, which combines evidence generation and clinical practice to form a continuous improvement cycle, has been defined for over a decade, yet the technical infrastructure required to operationalize this vision has been inconsistent[1]. The primary challenge lies in creating architectures that can both complete rigorous observational research as well as real-time clinical decision support.

The Observational Health Data Sciences and Informatics (OHDSI) collaborative includes hundreds of participating organizations across six continents who standardizes around the OMOP Common Data Model and the ATLAS analytical platform. This infrastructure has facilitated multi-site networked studies involving millions of patients and generating evidence that has informed clinical guidelines and regulatory decisions. The existing OHDSI architecture was designed initially for retrospective analysis rather than the dynamic, operational analytics required for learning health systems implementation.

Machine learning operations (MLOps) describes several frameworks for productionizing AI in enterprise environments, emphasizing reproducibility, automated testing, continuous integration, and robust governance. This is similar to OHDSI's mission towards reproducible research but have not been systematically integrated within the clinical research infrastructure. The result is a gap between research insights and clinical application[2], where valuable predictive models developed through rigorous research remain separated from the clinical workflows it could potentially optimize[4].

Methods

We developed a comprehensive extension to the OHDSI research platform that enables end-to-end machine learning workflows while preserving the collaborative, standards-based approach that's been advanced by OHDSI for the last 17 years[3]. Our approach intends to improve how observational research infrastructure can support both traditional retrospective analysis and clinical applications within a unified platform. Previously we established incremental daily OMOP-CDM uploads, which allows for recurring updates by the ETL from the source data[5].

We continued to develop an extension to the OHDSI WebAPI that removes caching limitations, enabling dynamic cohort generation necessary for just-in-time inference applications. We introduced the "data basket" construct which is a reusable abstraction for feature engineering that enables clinical researchers to define a "model ready" dataset[5]. These baskets encapsulate complex logic for extracting demographics, laboratory values, vital signs, conditions, procedures, and other clinical variables from OMOP-standardized data. Furthermore, the system supports high temporal granularity, allowing researchers to extract data relative to specific events or cohort entry/exit points, all configured through ATLAS's familiar user-interface.

We use Dagster as an orchestration framework[1], providing pipeline management with comprehensive data lineage tracking, automated testing, and flexible execution patterns. The architecture does not enforce a particular modeling or data framework. Instead, it focuses on maintaining strict versioning and provenance tracking required for clinical applications. Importantly, we implemented bidirectional data flow through our "Interrogator" utility, enabling the creation of analytical-specific concepts, which are then written back into the OMOP CDM's observation table, where they inform downstream cohort definitions and clinical decision rules[5].

Enterprise-level security controls are implemented throughout the pipeline, with IRB integration enforced at multiple levels to ensure compliance with institutional governance policies[5]. Additional functionality for audit logging and secure multi-environment deployments patterns was developed.

Results

The platform has been successfully deployed across multiple healthcare institutions, supporting both research and direct clinical applications. Real-time sepsis phenotyping algorithms process continuous data streams and generate risk scores automatically integrated into clinical decision support workflows, with preliminary validation studies demonstrating improved detection sensitivity while maintaining operational specificity requirements. Automated chart abstraction pipelines have replaced manual quality reporting processes, reducing abstractor workload by approximately 60% while improving data completeness and accuracy for regulatory reporting.

One example was the integration with the Montefiore Health System's Epic EHR for real-time inference for Sepsis patients across 10 hospital sites, which led to improvements in care team response times and more consistent application of the hospital's treatment protocols. These implementations demonstrate the platform's ability to combine both research insights and clinical practice, which are components of a learning health system.

Beyond clinical applications, the platform improved the entire research workflow by streamlining cohort definition, feature engineering, model development, and integration back into the EHR. This resulted in 3-4^{x} reductions in time-to-insight for exploratory analyses while maintaining reproducibility and version control. The reusable data basket construct allows for better collaboration across research teams, with feature definitions shared and refined across multiple projects, improving consistency while avoiding redundant work.

Conclusion

This work creates a pathway for healthcare institutions to evolve their data science capabilities while preserving existing investments in OHDSI infrastructure and OMOP data standardization. The modular architecture and open-source foundation allows the platform to grow within the OHDSI community, creating opportunities for collaborative development of standardized feature libraries that benefit the entire network.

The integration of modern ML-Ops practices with established clinical research infrastructure demonstrates that learning health systems can be implemented without abandoning the rigorous, collaborative approaches that have made observational research successful. This project shows how these approaches can be further extended and enhanced to support the continuous insight ingestion that defines a learning health system.

By successfully bridging the gap between research and clinical operations through standards-based, scalable architecture, we have created a practical framework for advancing the vision of the learning health system that has been articulated by leading health policy organizations for over a decade. The platform's ability to support daily clinical applications while accelerating research workflows positions it as important infrastructure for the future of evidence-based healthcare delivery.

References:

1. Budrionis A, Bellika JG. The learning healthcare system: where are we now? A systematic review. *J Biomed Inform.* 2016;64:87-92.
2. Rajagopal A, Ayanian S, Ryu AJ, Qian R, Legler SR, Peeler EA, et al. *Machine Learning Operations in Health Care: A Scoping Review.* Mayo Clin Proc Digit Health. 2024; **2**(3): 211–224
3. Henninger EM, Soby S, Wahle M, Aasman B, Goriacko P, Nelapatla C, et al. *Using OMOP-CDM to Develop Dynamic Disease Registries and Analytic Data Enclaves to Share and Use Real-world Evidence.* OHDSI Global Symposium 2023: Software Demonstration 415
4. Lee TC, Shah NU, Haack A, Baxter SL. *Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review.* Informatics (MDPI). 2020; **7**(3):25
5. Mirhaji P, Soby S, Henninger E, Nelapatla C, Wahle M, Aasman B, Bellin E. *Data quality monitoring, transparency and governance: Enterprise process for data quality stewardship and governance for real-world data* [Internet]. Presented at: OHDSI Symposium; 2022 Oct 10; Washington, DC