OMOPSYNC - NABC Action Plan

LOF TA: Merlin Simoes

Team Members:

- Aum Sathwara (A20543213)
- Vishnu Priyan Sellam Shanmugavel (A20561323)
- Harneet Kaur Dehiya (A20548613)
- Bjoern Ove Sagstad (A20557181)

An article by RBC Capital Markets states that healthcare already produces ≈ 30 % of the planet's data, and volume is set to leap from 2.3 ZB in 2020 to 10.8 ZB by the end of 2025 (36 % CAGR); traditional, hand-built ETL pipelines simply cannot keep up with that growth [1]. According to a data brief by the Office of the National Coordinator for Health Information Technology fewer than half of U.S. hospitals (43 %) routinely exchange and integrate external patient information, and roughly 30 % remain not fully interoperable, hampering cross-institutional studies. A 2024 HIMSS survey found that 47 % of healthcare leaders are dissatisfied with their organizations' data quality, citing budget limits and manual workflows as the top barriers.

Accessing and sharing patient data across various platforms is a common issue that affects the quality of care. We aim to unlock the full potential of healthcare data by enabling standardized, interoperable, and privacy-preserving analytics across synthetic and real-world patient records. Our solution bridges this gap by simplifying the process of converting diverse data sources into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), making it actionable for clinical research, while ensuring privacy and data quality.

Problem:

Current healthcare datasets are fragmented, and end-users like researchers and healthcare providers struggle to analyze and integrate them. Health data from different sources come in incompatible formats with different schemas, vocabularies, and structures. A survey by Verato indicated that 49% of healthcare leaders report their patient data is stored in fragmented, siloed systems.

In short, we're solving the problem of data fragmentation in healthcare by providing a user-friendly, scalable, and privacy-aware solution to enable high-quality, reproducible health research.

Need:
There is a significant need to simplify the process of standardizing healthcare data to improve research and clinical outcomes. Inconsistencies in data make it difficult to perform reliable, large-scale, and reproducible health data analysis across systems or institutions. Without standardized data, large-scale analysis is error-prone and limited in scope. However, converting healthcare data into the OMOP CDM format is a labor-intensive and complicated process. Researchers need tools to easily convert and unify data to avoid slower research timelines and inconsistent patient care outcomes, and for accurate, reproducible studies. This ensures that they can focus on the actual research rather than spending time cleaning and transforming data.

Without standardized data:

- Researchers spend a lot of time and effort to manually clean and map datasets.
- Analysis may be error-prone or non-compliant with regulatory standards.
- It's nearly impossible to compare or merge data across studies or institutions.

Researchers and clinicians lack a simple and compliant way to transform heterogeneous health datasets into a unified OMOP CDM format to support longitudinal, cross-institutional analysis.

Approach:

We integrate various tools like the ETL pipelines in Python / R / SQL and a Large Language Model (LLM) using a Model Context Protocol (MCP) server. This solution enables users to initiate data processing with simple natural language prompts that are parsed by the LLM and mapped to the appropriate MCP tool. The MCP server manages execution by providing a catalog of asynchronous tools such as Run SQL Queries, Plot Graphs, and Run ETL Pipelines.
For example, a user can simply input: "Convert the CSV file at [directory] into the OMOP CDM format."
This prompts the LLM to trigger the corresponding backend ETL process (in this case, a Python script). The resulting transformed data is then relayed back to the LLM, which presents the output directly to the user.
Currently, we use the Synthea dataset as our raw data. The ETL pipeline itself can be updated / changed to cater to different datasets without affecting the overall system. A Data Quality Dashboard is displayed after the ETL process is completed to evaluate the accuracy of data conversion. We have also integrated OHDSI's Atlas to facilitate OMOP CDM data analysis.

Benefits:

- Accelerates clinical and population health research by reducing the effort spent on data cleaning and integration.
- Investors benefit from reusable IP, future product integration, and reduced time-to-insight in clinical research settings.
- Enhances reproducibility and privacy compliance.
- Provides a modular and scalable solution that could be used by healthcare institutions, researchers, and public health agencies.

Competition:

- IQVIA OMOP Converter: Customized SQL-based solutions to convert data to the OMOP CDM format.
- The Hyve's Delphyne: Open-source OMOP ETL. Efficient, yet lacks adaptability for custom data needs.
- Inhouse ETL Pipelines: Custom academic tools built in-house at universities / hospitals / research institutions

Our solution requires minimal set-up, is fully customizable, and transparent. IT offers an easy solution – once it has been set up, users need only to give simple prompt-like commands to perform ETL. The entire system resides on a local machine, thereby ensuring data privacy.

Risk Mitigation:

Name	Description	Risk	Risk Mitigation Plan
Data Integration	Difficulties in standardizing diverse healthcare datasets.	High	Increase testing on data quality and consistency.
Regulatory Compliance	Delays or issues with ensuring full regulatory compliance.	Low	Engage legal and compliance teams early in the process.
Vocabulary Management	Frequent vocabulary updates to OMOP CDM	Medium	Regular checks on vocabulary updates.
Technology Scaling	Challenge in scaling the ETL process for larger datasets.	Medium	Optimize ETL scripts and explore cloud-based solutions.

How can this value proposition be improved?

- Integrating more data sources into the OMOP CDM to expand the data pool.
- Automating vocabulary updates and ETL validation steps.
- Provide templated queries and reporting.

Why hasn't it been solved?

- The Model Context Protocol server architecture is fairly new.
- Limited open-source tools that are user-friendly and well-maintained
- Integrating heterogeneous datasets into OMOP CDM requires complex, customized ETL logic, evolving vocabularies, and strict compliance safeguards, which most existing tools don't handle in a unified or scalable way.
- Variability in data sources and evolving healthcare vocabularies.
- Strict privacy and compliance regulations make general solutions difficult.
- Lack of one-size-fits-all ETL due to different schema designs.

What more would end-users need to know?

- How the ETL pipeline ensures data fidelity, vocabulary accuracy, and compliance readiness, along with clear documentation, runtime performance metrics, and guidance on how to use the resulting OMOP datasets with tools like Atlas.
- Security protocols for handling PHI under HIPAA.
- How to interpret OMOP-coded data for clinical research.
- Tutorials on integrating Atlas and running common cohort queries.

What more would investors need to know?

- Potential for this solution to be generalized and scaled for other datasets.
- Market size: number of institutions adopting OMOP/CDMs (e.g., academic medical centers, pharma, and public health agencies)
- Reusability across research networks
- Potential for commercial licensing or integration with existing healthcare analytics platforms
- Plans for long-term maintenance, support, and monetization (e.g., SaaS model).
- Projected revenue growth and market penetration over the next 5 years.