

# Accelerating Analytics with Multi-source, OMOP-Conformed Data and Community Tools in a Cloud-Native Platform

Aleksandra Petkova <sup>1</sup>, Glynn Dennis <sup>1</sup>, Lance Dowling <sup>1</sup>, and Chris Baldwin <sup>2</sup>  
<sup>1</sup> Kythera Labs, <sup>2</sup> The Hyve

## Background

The growing volume and diversity of real-world data (RWD) present immense opportunities for healthcare analytics, but also significant integration challenges. Data are often siloed across sources - e.g., medical claims, electronic health records (EHRs), and lab results - each with inconsistent formats, vocabularies, and structures, making combined analysis time-consuming and error-prone. Standardizing such disparate data to a common model can enable a more comprehensive understanding of patient journeys and outcomes. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)<sup>2</sup> has emerged as a global standard for observational healthcare research data<sup>1</sup>, offering a uniform schema and standardized vocabularies to facilitate multi-source analytics. However, simply adopting OMOP doesn't by itself resolve data fragmentation issues - careful data re-mapping and a strong infrastructure are needed to truly operationalize multi-source RWD at scale.

Kythera Labs, in collaboration with The Hyve, an OMOP specialist partner, undertook an initiative to unify and standardize large-scale multi-source RWD to the OMOP CDM. Our goal was to streamline the analysis of multi-source, multi-model RWD by leveraging a unified, single-model strategy. Central to this effort was Kythera's cloud-native platform, Wayfinder, built on Databricks, which provided the scalable, Spark-based compute environment and a FAIR (Findable, Accessible, Interoperable, Reusable)<sup>3</sup> data catalog needed for handling billions of records of data. Aligning the OMOP transformation pipeline with the FAIR data principles enabled a searchable data catalog, role-based access controls, and lineage and provenance tracking of the resulting multi-source OMOP dataset.

Data quality is paramount for observational healthcare research studies. While OHDSI offers a suite of valuable open-source tools, these have not traditionally been configured to run natively in cloud-based environments. We began by enabling

WhiteRabbit to run within Wayfinder, supporting pre-ETL data scanning. To extend this approach and provide a scalable path for the ETL development teams, we containerized additional tools, specifically the Data Quality Dashboard and Achilles, so those can operate effectively in a Spark-based big data environment. This approach ensured that familiar OHDSI tools are seamlessly accessible in modern, cloud-native workflows, bridging the gap between research usability and enterprise-scale data infrastructure.

It is the platform-driven approach, coupled with OMOP subject matter expertise, that facilitated the efficient, reproducible, and reusable transformation of multi-source RWD to the OMOP standard. Therefore, it is important to highlight that our focus was not just on developing the ETL process to produce the OMOP-formatted data, but also to do so at scale, speed, quality and with computational efficiency. In what follows, we describe our specific approach and highlight results to illustrate the data composition of the final OMOP dataset at a high level. We conclude by outlining the next steps in this work, including assessment and preparation of the OMOP data for comparative studies as well as incorporating additional data types including clinical notes.

## **Method**

### **Data Sources**

Data from four U.S.-based nationwide RWD sources to standardize within the OMOP model: one nationwide medical claims dataset and three sources of EHR and clinical laboratory data. Data from all sources was de-identified and tokenized to ensure patient privacy and anonymity is preserved. Raw medical claims data was transformed into patient-level event records to enable the construction of clinical event-based representations - for example, linking diagnoses, procedures, and providers from separate claims into structured patient journeys. The EHR and lab data from three additional sources were standardized into a staging layer, organizing information into relevant domains (e.g., patients, providers, encounters, vitals, results, and measurements).

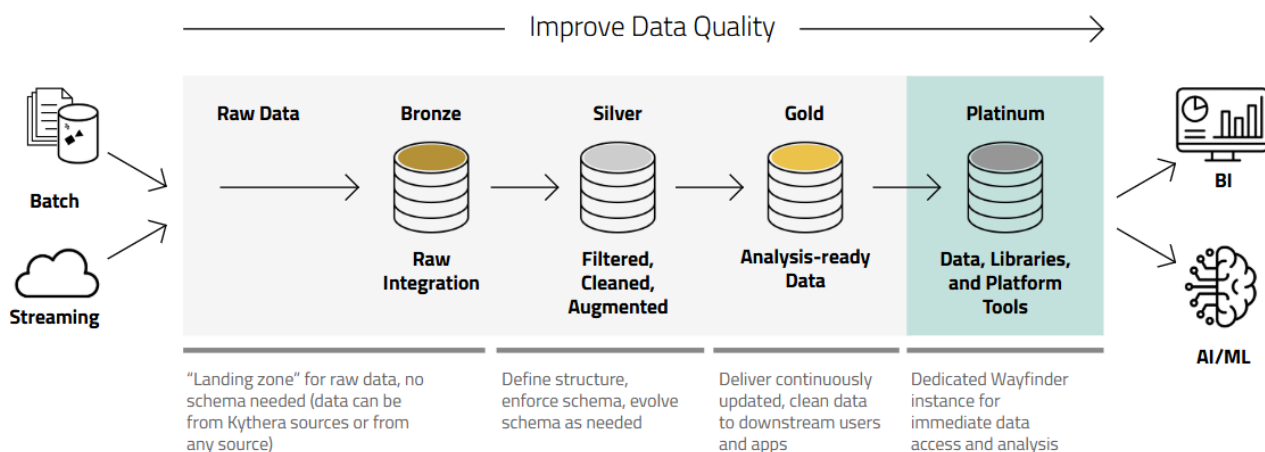
## Data Pipelines

The data transformation pipeline followed a layered, medallion architecture (Bronze - Silver - Gold - Platinum) within the Wayfinder platform (see Figure 1 at the end of this section). In the Bronze layer (raw integration), source data was ingested “as-is”, without imposing schema or other requirements. Next, in the Silver layer (remastered data), we standardized each source into a structured intermediate schema. For the claims data, this meant transforming raw claim lines into patient-level event records, linking together all diagnoses, procedures, medications, and providers that pertained to the same healthcare encounter or episode, even if those were listed in separate claims. Claims data then was further processed to produce patient event tables (e.g. aggregated surgical events, physical therapy, diagnostic, and hospital events, composed from multiple claims). For the EHR and lab sources, we similarly organized the data into a common staging schema with normalized tables for patients, encounters, providers, observations, measurements, etc., ensuring that disparate clinical source systems were mapped to a consistent structure before transformation to OMOP. This established a uniform base from which we could apply OMOP mappings in a source-agnostic manner. Importantly, this shared data staging layer eliminated the need for four parallel OMOP ETL builds (one per source). The final end-to-end ETL run completes within 90 minutes across 2.75 TB of structured RWD, making onboarding new sources much faster (stage the data once, reuse the OMOP mappings).

We performed transformations to the OMOP CDM (version 5.4) in the Gold layer. This involved applying semantic and code mapping logic (developed jointly with The Hyve) to map source-specific code systems (ICD-9/10, CPT, LOINC, etc.) and clinical concepts to the standardized OMOP vocabularies (SNOMED, RxNorm, etc.) and to fit the data into OMOP’s relational schema. The collaboration with The Hyve was instrumental in creating robust mapping algorithms and lookup tables to handle the complexity of multiple coding systems and to ensure fidelity of clinical meaning across sources.

Finally, the Platinum layer in our architecture refers to the delivery of the curated OMOP dataset and associated services and tools for end-users, such as analytics and data science subject matter experts. All data processing and transformations were run

on Kythera’s Wayfinder platform. Throughout the ETL process, each OMOP table was version-controlled, annotated and linked to provenance metadata, providing transparent lineage in line with FAIR<sup>3</sup> data principles. Secure data access and large-scale data sharing was enabled by Wayfinder’s Delta Share capabilities without the need for data duplication, hence allowing authorized users to query the data directly in the cloud. This was particularly important for scaling collaboration between Kythera and The Hyve, while maintaining transparent data governance.



**Figure 1:** *Kythera’s medallion approach to real-world data processing.*

## Tools Integration

The WhiteRabbit tool was first enabled to connect to data stored in Wayfinder for initial data scans. In addition, two open-source OHDSI tools Achilles and Data Quality Dashboard (DQD) were containerized to run natively in Wayfinder to ensure data quality and consistency at scale within the same platform. By doing so, these R-based tools could run natively in the cloud on Wayfinder, overcoming their typical performance limitations on large datasets. Post ETL and final OMOP transformations, we executed Achilles to generate summary statistics and plots for the OMOP data and ran DQD to systematically scan for any data issues or conformance problems.

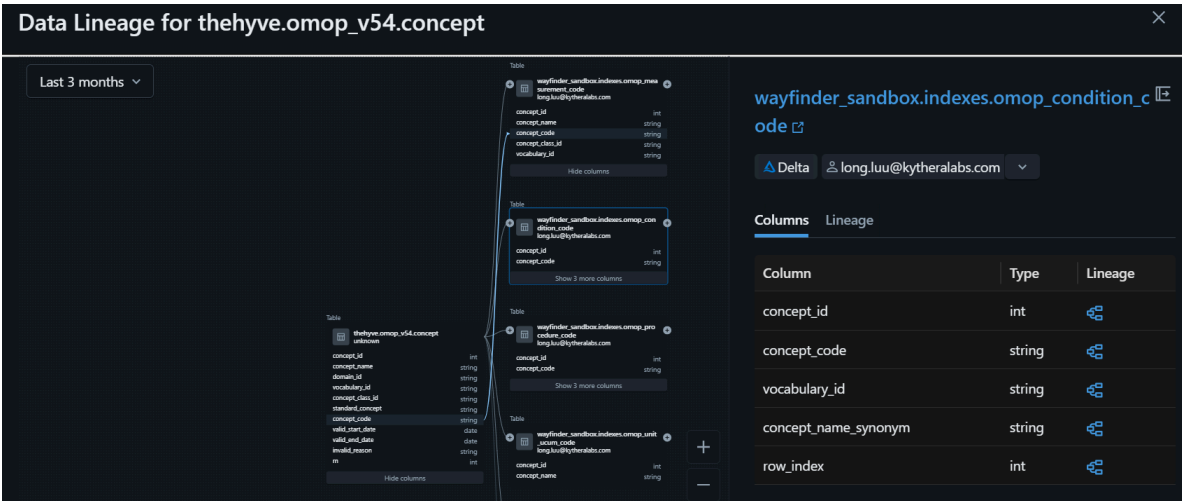
Quality findings (e.g., unexpected or missing data or mapping errors) were reviewed with The Hyve and corresponding OMOP transformation updates were applied. This integration of community QA tools ensured that our OMOP output was not only standardized in schema, but also in quality – aligning with OHDSI community

standards. These tools have remained integral in Kythera’s continued efforts to scale and maintain the current OMOP implementation.

## Results

Over the span of three months, four national-scale heterogeneous datasets (medical claims, EHR records from two separate sources, and lab results) were mapped to OMOP. As a result of the source data remastering approach, each data source was standardized into intermediate event tables with a consistent structure. These intermediate steps streamlined the transformation to OMOP across datasets and reduced the need for separate extract-transform-load (ETL) pipelines per source, saving considerable cost, time, and resources.

All transformation logic (SQL notebooks and Python orchestration) is version-controlled in AWS CodeCommit, ensuring systematic updates and reproducibility into the development and execution of OMOP transformation steps. Wayfinder also provides built-in lineage and audit trails, enabling visibility into downstream use of OMOP tables (see Figure 2 for an example of how Wayfinder automatically captures lineage). In addition, Wayfinder’s Delta Sharing functionality allows external partners to securely query and interact with the OMOP dataset in a privacy-preserving environment, removing the need to copy or redistribute large datasets.



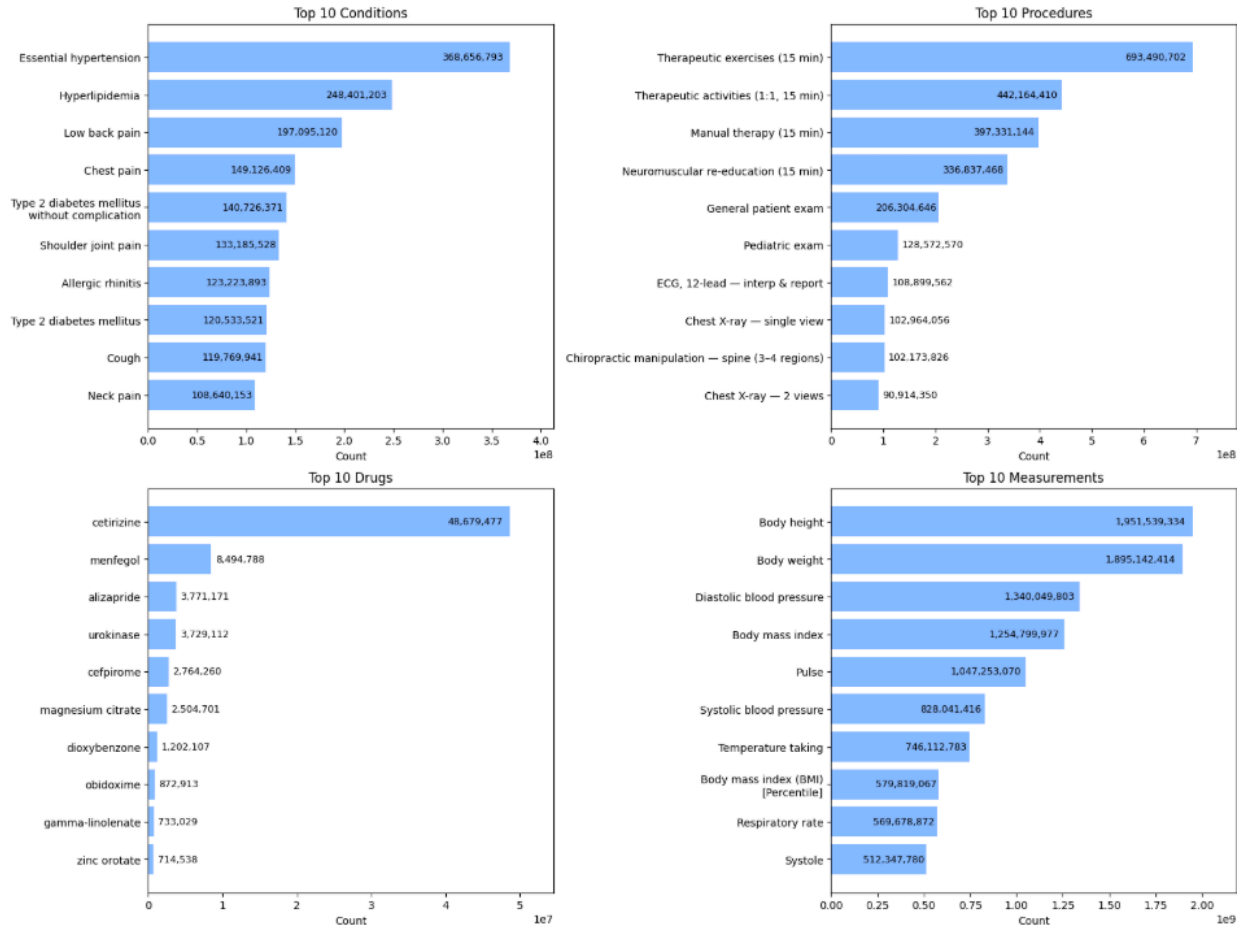
**Figure 2:** *Example of data lineage captured in Wayfinder for the OMOP concept table. The visualization shows how downstream assets are derived from the table and tracks which columns are used, ensuring transparency in how users query the data. This built-in lineage allows teams to trace dependencies, monitor changes, and maintain reproducibility across analytics environments.*

The resulting OMOP dataset includes 582,755,368 unique person identifiers<sup>1</sup>, spanning a wide range of ages and clinical domains. Patients are represented across birth years from 1931 to 2025, with a concentration among those born between 1950 and 2000 (top five decades). The dataset captures 844,920 unique condition concepts, with the most frequently observed diagnoses including hypertension, chest pain, and cough. Visit data reflected a broad range of care settings, with inpatient, outpatient, and emergency department encounters being the most common. See Figure 3 below for data characterization statistics.

Finally, we observed that containerizing the Achilles and DQD tools greatly accelerated their runtime on the full dataset, allowing us to complete data characterization with Achilles and Data Quality Dashboard in hours rather than days, without incurring excessive compute cost. Data quality checks with DQD identified and resolved potential issues early, ensuring a high-fidelity OMOP translation and enabling us to continually refine the ETL.

---

<sup>1</sup>The reported patient count exceeds the total U.S. population, likely due to the presence of potentially duplicated individuals across separately acquired RWD sources. The absence of universal patient identifiers makes it difficult to determine when two records from different data vendors may refer to the same person. We are actively developing algorithms to detect and reconcile the same individuals from different datasets.



**Figure 3:** Top-10 frequency concepts across four OMOP domains (conditions, procedures, drugs, measurements), illustrating coverage and scale across combined claims and EHR sources.

## Discussion

Adopting the OMOP Common Data Model for integrating multi-source real-world data (RWD) has significantly accelerated Kythera’s internal analytics operations. Transitioning from a multi-source, multi-model approach to a unified multi-source, single-model strategy has enabled the use of reusable, standardized queries built upon a robust, large-scale data infrastructure. Our experience demonstrated that collaboration with specialized OMOP partners, such as The Hyve, significantly complements internal data and technology expertise, facilitating efficient and accurate data mapping.

We identified that the close integration of domain knowledge from subject matter experts, coupled with the technical skills of data and technology specialists creates an optimal environment for successfully managing complex data transformations and analytics workflows. Furthermore, our modern, cloud-native architecture has demonstrated scalability, reliability, and efficiency, serving not only Kythera's internal needs but also offering a practical reference model for the broader OHDSI community.

Importantly, the incorporation of transparent versioning and sharing mechanisms aligns with FAIR principles<sup>3</sup>, ensuring continuous availability and usability of data assets for diverse stakeholders.

## **Future Directions**

Following the initial integration and its successful outcome, we have several next steps planned to further enhance and leverage this multi-source OMOP dataset and platform technology. Next steps include algorithmically applying patient de-duplication and linkage (i.e., identifying the same patients across various data sources); enabling regular incremental updates to the final OMOP dataset, and containerizing additional OHDSI tools for data quality monitoring, as well as leveraging AI to streamline cohorting. Our roadmap also includes expanding the current data coverage to add additional intelligence on drugs and other data domains relevant for life science and healthcare research (e.g., clinical notes).

Importantly, this structured QA process also lays the foundation for applying AI/LLM-based assistants to help interpret results from standard tools, assist with triage, and prioritize quality issues at scale.

## **Conclusion**

In conclusion, this project highlighted that effective cross-functional collaboration, combined with a scalable cloud-based infrastructure, is pivotal for realizing the full potential of large-scale OMOP implementations. We have demonstrated that with a robust architecture and strong adherence to standards, even extraordinarily large and disparate healthcare datasets can be integrated into a common model. The outcome is

a dataset that is versatile and broad in scope that can in the future drive a variety of observational and real-world evidence studies. Notably, there are clear areas for further improvement (especially patient de-duplication and deeper evaluation of data completeness), but the progress so far provides a foundation upon which incremental enhancements (and new data additions) can be added. Furthermore, we envision this foundational work and the adoption of the OMOP CDM, to enable more federated, privacy-preserving, approaches to the analysis of disparate RWD sources, that will enable larger scale evidence generation. We consider this a journey rather than a one-time project, which mirrors the broader OHDSI mission of continually improving the ecosystem of tools and data for observational healthcare research that promotes better health decisions and better care.

## References

1. Hripcsak G, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform.* 2015;216:574-8.
2. Observational Medical Outcomes Partnership. The Common Data Model (CDM). Foundation for Observational Health Data Sciences and Informatics (OHDSI). Accessed September 2025. Available at: <https://www.ohdsi.org/data-standardization/>.
3. Wilkinson MD, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.

## Contact Information

Aleksandra Petkova ([aleksandra@kytheralabs.com](mailto:aleksandra@kytheralabs.com)), Senior Product Manager, Kythera Labs  
Glynn Dennis ([glynn@kytheralabs.com](mailto:glynn@kytheralabs.com)), Chief Science Office, Kythera Labs  
Lance Dowling ([lance@kytheralabs.com](mailto:lance@kytheralabs.com)), Director of Analytics, Kythera Labs  
Chris Baldwin ([chris@thehyve.nl](mailto:chris@thehyve.nl)), Solutions Consultant, The Hyve