Scalable Big Data Workflow for OMOP CDM: Performance Optimization and Automated Quality Evaluation of Real-World Data

Danilo Luis Cerqueira Dias[1], Ricardo Felix Monteiro Neto[1,] Juliana Araújo Prata de Faria[1] , Valentina Martufi[1], Julio Barbour Oliveira[2], Karine Brito Beck da Silva Magalhães[1], Roberto Perez Carreiro[1], Maurício L. Barreto[1], Elzo Pereira Pinto Junior[1], Pablo Ivan Pereira Ramos[1]

[1]Center for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, BA, Brazil

[2]Precision Data

Background

Ensuring the quality and interoperability of health data is critical for advancing evidence-based public health policies and enabling collaborative international research. This work presents the development of a computational infrastructure designed to model health data according to the OMOP (Observational Medical Outcomes Partnership) Common Data Model. The infrastructure is tailored to maintain data integrity while supporting federated analyses focused on gestational syphilis, using datasets from the Brazilian public health information system. The study was conducted at CIDACS (Center for Data and Knowledge Integration for Health), leveraging its secure and large-scale data environment.

Methods

The data modeling process to the OMOP CDM standard was carried out within a Trusted Research Environment (TRE), ensuring strict governance, data privacy, and compliance with ethical and legal standards for handling sensitive health information. The computational infrastructure consisted of a high-performance Linux environment equipped with 126 GB of RAM and 500 GB of Disk storage, enabling efficient large-scale data processing while maintaining a secure and controlled setting. The methodology was designed to ensure scalability, transparency, and adherence to OHDSI best practices, and consisted of three main stages:

1. Data profiling and ETL Development: Data profiling, extraction, transformation, and loading were implemented using *Jupyter Notebooks* and *PySpark*, enabling distributed processing of over 24

million health records. This step included mapping source variables to standardized OMOP concepts using controlled vocabularies such as SNOMED CT, LOINC, and RxNorm, ensuring semantic interoperability.

2. Data Storage and Structuring: The transformed data was loaded into a *PostgreSQL* database structured according to OMOP CDM version 5.4. This schema provides a normalized, relational format that supports analytical consistency and facilitates integration with the OHDSI ecosystem.

3. Data Quality Assessment: Data integrity and model compliance were rigorously evaluated using the *DataQualityDashboard R package*. More than a thousand automated quality checks were performed, covering conformance, completeness, and plausibility dimensions. The process enabled iterative improvements to the ETL pipeline, enhancing the overall robustness of the final dataset.

Results

After executing the *DataQualityDashboard* (DQD), a total of 1,363 automated validation tests were applied to the OMOP CDM database, covering key dimensions such as conformance, completeness, and plausibility. Of these, only 37 tests (2.71%) resulted in failures — a low error rate that reflects strong alignment with the OMOP data model. Most of the issues were identified at the field level, such as missing or improperly formatted values, followed by concept-level inconsistencies related to vocabulary mapping. Notably, the lowest number of errors was detected at the table level, indicating structural soundness across the CDM schema.

Each inconsistency was thoroughly reviewed, traced back to its origin in the ETL pipeline, and corrected with targeted transformations. The data was then fully reprocessed and redeployed, ensuring an updated and compliant version of the dataset. This rigorous validation process highlights the robustness of the ETL implementation and underscores the project's commitment to data quality, standardization, and semantic interoperability—crucial pillars for enabling trustworthy federated analyses and cross-institutional research collaboration.

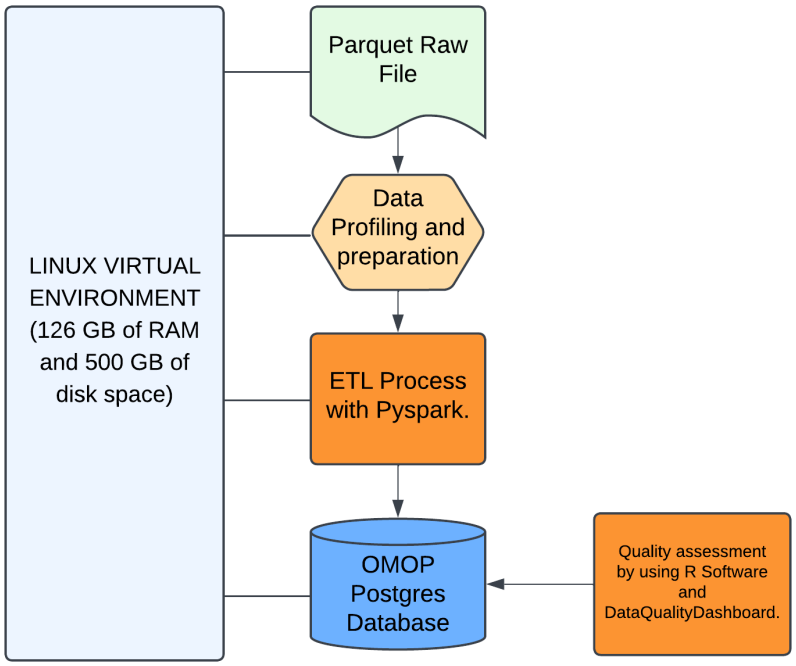

Figure 1. CIDACS OHDSI OMOP Environment and Process of OMOP Database Construction.

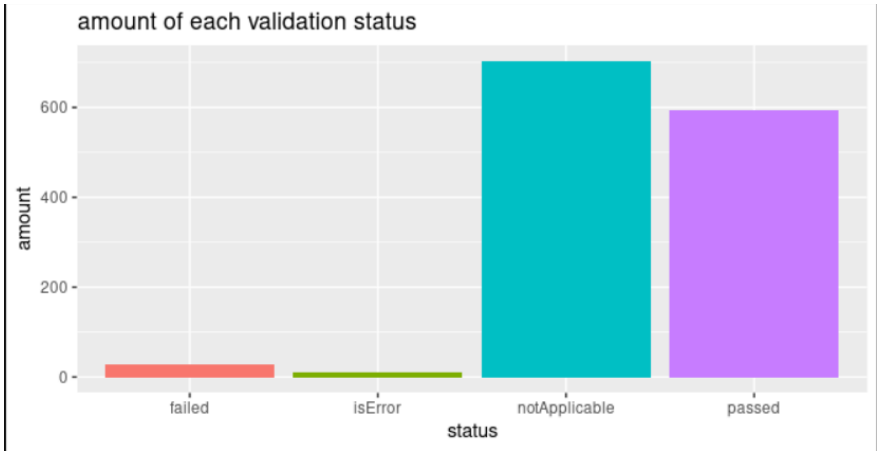

Figure 2. Data Quality Dashboard Results (validation Status)

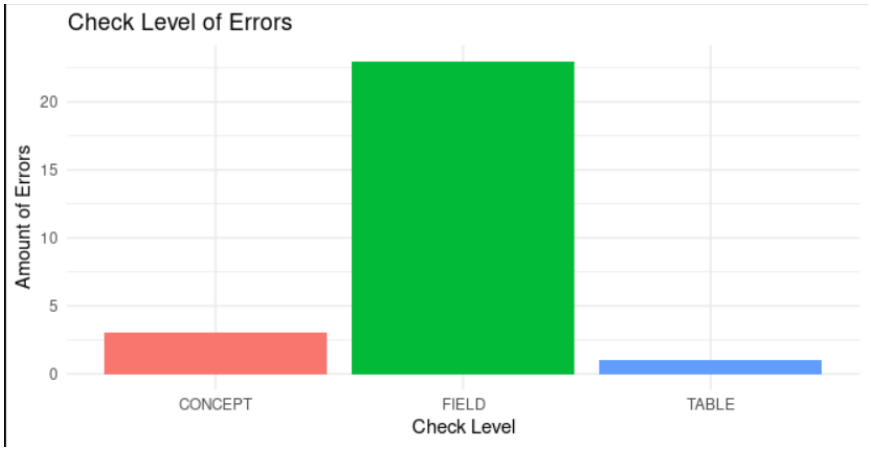

Figure 3. Data Quality Dashboard Results (Level of Errors)

Conclusion

Modeling administrative health data using the OMOP CDM standard proved not only technically feasible but also strategically valuable for enabling large-scale, population-level health analyses. The process faced inherent challenges, particularly in achieving semantic alignment across heterogeneous data sources and conforming to the structural rigor required by the OMOP Common Data Model. However, these obstacles were effectively addressed using robust and scalable technologies—*PySpark* for distributed data transformation, *PostgreSQL* for structured storage, *R software* for advanced analytics, and the *DataQualityDashboard* for systematic quality assurance.

This integrated toolchain facilitates precise mapping, rigorous validation, and iterative refinement of the data pipeline. The resulting infrastructure is not only technically sound, but also aligned with international standards for data interoperability, making it well-suited for participation in global research initiatives under the OHDSI framework. Moreover, the approach reinforces scientific reproducibility, supports evidence-based decision-making, and provides a replicable model for other institutions or countries seeking to modernize their public health data systems through OMOP adoption. The OMOP CDM-based computational infrastructure proved effective in standardizing administrative health data in Brazil. With scalable technologies and OHDSI tools, it was possible to structure an interoperable and reliable database. The low error rate highlights the quality of the process. This initiative enables federated analyses among countries, fostering data security and sovereignty, collaborative science, and evidence-based public policy support.

References

1. Hripcsak, G., Duke, J.D., Shah, N.H., Reich, C.G., Huser, V., Schuemie, M.J., Suchard, M.A., Park, R.W., Wong, I.C., Rijnbeek, P.R., Van der Lei, J., Pratt, N., Norén, G.N., Li, Y.C., Stang, P.E., Madigan, D. e Ryan, P.B. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for

Observational Researchers. Stud Health Technol Inform. 2015; 216:574-8. PMID: 26262116; PMCID: PMC4815923.

2. Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G., & Stang, P. E. (2012). Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association, 19(1), 54-60. https://doi.org/10.1136/amiajnl-2011-000376

3. Voss, E. A., Blacketer, C., van Sandijk, S., et al. (2024). European Health Data & Evidence Network—learnings from building out a standardized international health data network. Journal of the American Medical Informatics Association, 31(1), 209-219. https://doi.org/10.1093/jamia/ocad214

4. Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., & Suchard, M. A. (2018). Improving reproducibility by using high-throughput observational studies with empirical calibration. Philosophical Transactions of the Royal Society A, 376(2128), 20170356. https://doi.org/10.1098/rsta.2017.0356

5. Observational Health Data Sciences and Informatics. The Book of OHDSI. The Book of OHDSI. Accessed Jun 25, 2025. https://ohdsi.github.io/TheBookOfOhdsi/

6. *OHDSI/WhiteRabbit [program].* Version V0.10.1. GitHub. Accessed Jun 25, 2025. https://github.com/OHDSI/WhiteRabbit

7. Trusted Research Environments. Accessed Jun 30, 2025. https://www.hdruk.ac.uk/access-to-health-data/trusted-research-environments/

8. *OHDSI/DataQualityDashboard (DQD) [program]*. GitHub. Accessed Jun 25, 2023. https://github.com/OHDSI/DataQualityDashboard