Comparing Timeline and Challenges of OMOP CDM Implementation in Brazil

Juliana Araújo Prata de Faria[1], Danilo Luis Cerqueira Dias[1], Valentina Martufi[1], Julio Barbour Oliveira[2], Ricardo Felix Monteiro Neto[1], 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), Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Salvador, BA, Brazil
[2]Precision Data

Background

The innovation of the Center for the Integration of Data and Knowledge for Health (CIDACS/Fiocruz - Bahia) consists in integrating large volumes of data (BigData) routinely collected by public management into databases aimed at scientific research, ensuring quality, accuracy, and a high level of validity. The Interoperability of Data and Federated Analyses (IDAF/CIDACS) techno-scientific group, in partnership with the Provincial Health Data Center (PHDC), developed a common data model for studying infectious diseases in pregnancy, using the OMOP CDM. This work highlights the standardization of 16 million records for the study of gestational syphilis.

IDAF used as a reference the guidelines and strategies of the European Health Data and Evidence Network (EHDEN) to support the successful standardization of observational data. EHDEN is a public-private partnership with the aim of building a federated network of standardized large-scale health data for OMOP CDM. Its main success metric for source data transformation is the total number of days required to carry out each stage of the Extract, Transform and Load (ETL) process.

This study presents a comparative analysis of the duration of the stages of the ETL process between the Brazilian and European experiences, highlighting the challenges faced, the lessons learned and the solutions adopted in the local context. The findings provide relevant input for other data standardization initiatives in countries of the Global South.

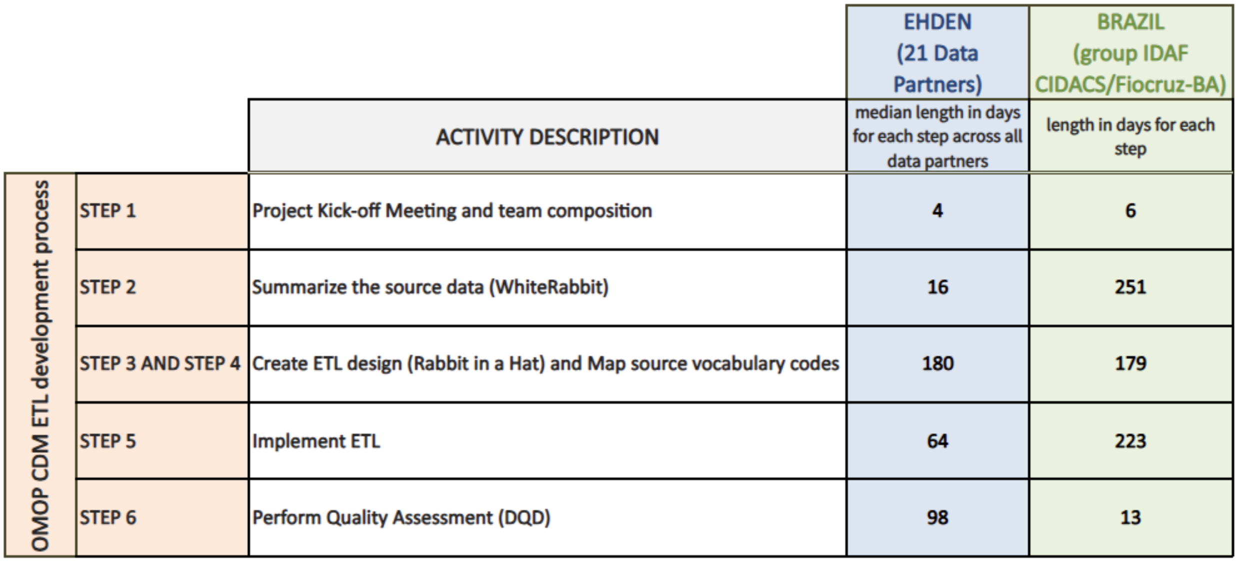
Methods

The work used a comparative approach to analyze the duration of the stages of the ETL process suggested by the Observational Health Data Science and Informatics (OHDSI) Community, based on the time metrics obtained in the EHDEN Network study. The workflow followed to implement ETL in the Brazilian context was adapted from the EHDEN and OHDSI Community methodological guide, including the following steps: (1) team formation and project kick-off meeting, (2) database exploration using the White Rabbit tool, (3) construction and execution of the ETL pipeline using Rabbit-in-a-hat, (4) definition of the vocabulary mapping using the Athena tool, (5) ETL implementation and (6) validation of the quality of the standardized database using OHDSI tools such as the Data Quality Dashboard (DQD).

The duration of each ETL stage was recorded in working days, following a methodology adapted from EHDEN. The Brazilian team, composed of six professionals from data science, medicine, epidemiology, and engineering, used an average of 4.6 people to align with EHDEN's baseline for comparison. Results were analyzed considering EHDEN’s reported averages and ranges, while accounting for differences in infrastructure and data availability. In Brazil, stages 03 and 04 were combined, so their average duration was calculated for comparison with the European network.

Results

The ETL process lasted 443 days, more than the EHDEN network average of 358 days but less than the maximum time observed of 622 days. The initial stage, from team formation to the kick-off meeting, took 6 days. The database exploration phase lasted 251 days, extending due to the substitution of synthetic data for the original database and the careful selection of variables in conjunction with the PHDC counterpart team. The ETL creation and vocabulary mapping stages were carried out in parallel, totaling 179 days, similar to the European average. The implementation of the ETL, which began before these stages were completed, took 223 days, more than the EHDEN average of 64 days. The validation and verification of data quality was completed in 13 days, with a highlight being the application of the DQD, which identified only 2.71% of faults needing correction in the pipeline. The results demonstrate adherence to the OMOP model and a commitment to data quality and interoperability.



			EHDEN (21 Data Partners)	BRAZIL (group IDAF CIDACS/Fiocruz-BA)
		ACTIVITY DESCRIPTION	median length in days for each step across all data partners	length in days for each step
OMOP CDM ETL development process	STEP 1	Project Kick-off Meeting and team composition	4	6
	STEP 2	Summarize the source data (WhiteRabbit)	16	251
	STEP 3 AND STEP 4	Create ETL design (Rabbit in a Hat) and Map source vocabulary codes	180	179
	STEP 5	Implement ETL	64	223
	STEP 6	Perform Quality Assessment (DQD)	98	13

Figure 1: The figure shows a comparison between the European network and the IDAF group in terms of the duration, in days, of the stages in the ETL development process.

The data from the European network corresponds to the average times recorded by 21 data partners in the EHDEN initiative. The data from Brazil refers to the execution carried out by the IDAF group, a working group linked to the project and the consultancy specializing in the implementation of the OMOP CDM model.

The implementation of OMOP CDM in CIDACS faced challenges such as the scarcity of adequate vocabularies for the Brazilian context and the need for technical training. The predominantly clinical nature of the OMOP CDM required adaptations to accommodate the administrative nature of CIDACS’ population data, reflecting the complexity of using large linked databases to analyze social determinants of health.

Adapting the model to data beyond the clinical context challenged the team to represent social and economic determinants of health through personalized mapping, while preserving analytical rigor. In this context, it required the team to make additional efforts in terms of translation and semantic alignment.

On the other hand, the adoption of the EHDEN workflow, the use of open-source OHDSI tools, the application of good knowledge management from the start of the project by recording every decision and changes made, elaborating experience reports and supporting documentation stand out as facilitators.

Experience has shown that although the Brazilian process takes longer in some stages, it is technically feasible and can generate products compatible with international standards. The adaptations made and the lessons learned reinforce the applicability of the OMOP CDM model in contexts of the Global South.

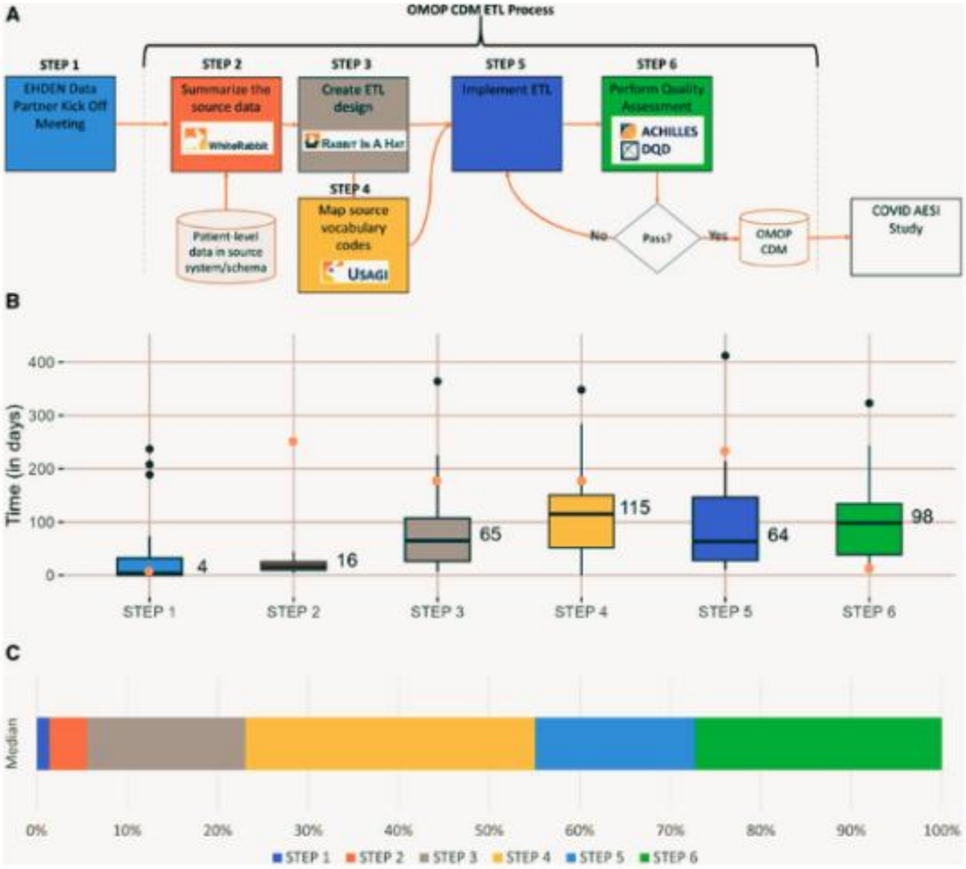
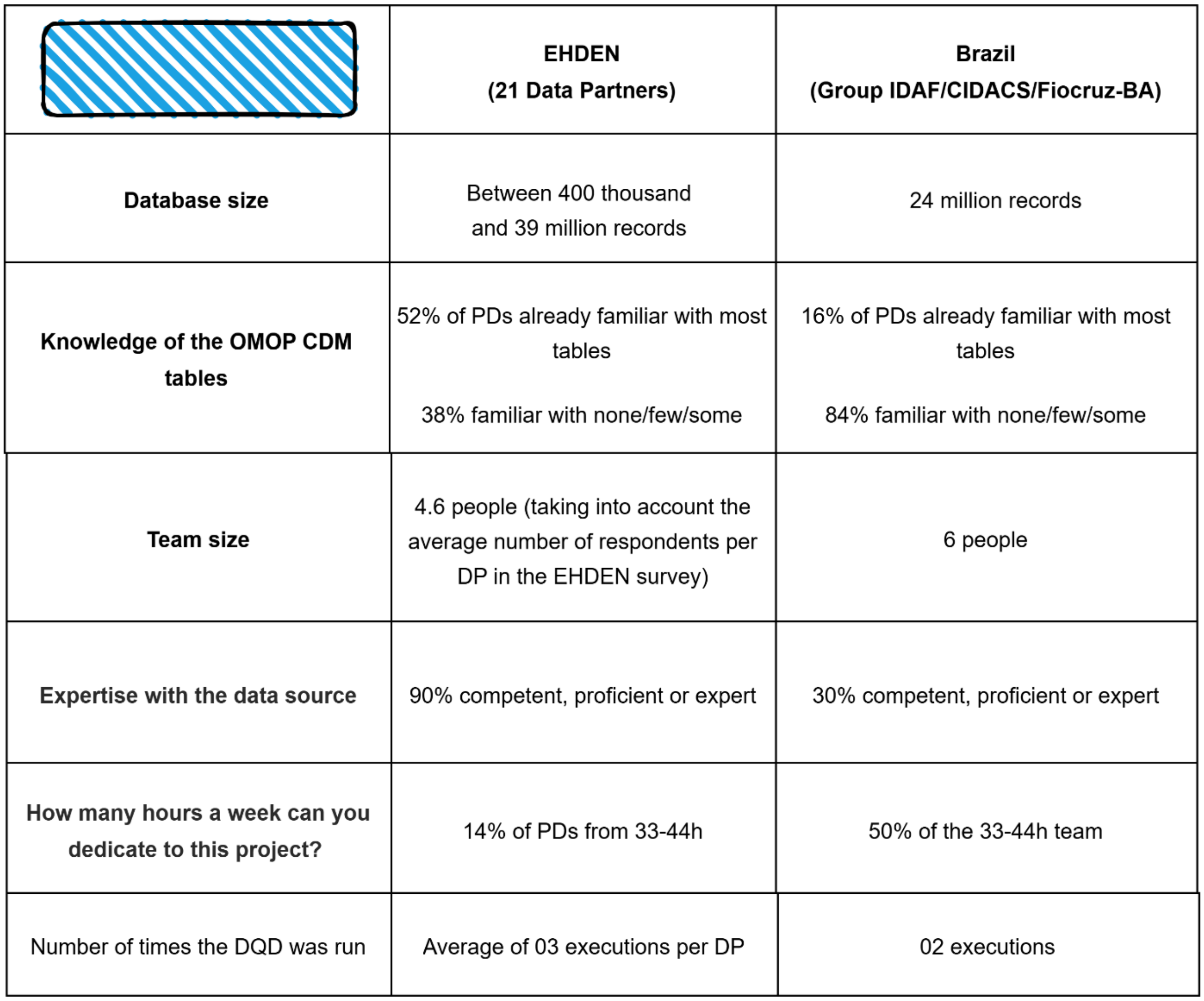


Figure 2 adapted to include the duration of the times obtained by the Brazilian team: OMOP CDM ETL development process: (A) represents the ETL process map, (B) is a box plot of the average duration in days for each step across all data partners, highlighting in orange the times obtained by the project team (orange dots; Brazil) and (C) is a stacked bar chart showing the percentage of the average time each step took. CDM, common data model; COVID AESI Study, 'Adverse Events of Special Interest within COVID-19 Subjects' study; DQD, DataQualityDashboard; EHDEN, European Health Data & Evidence Network; ETL, extract, transform and load; OMOP, outcomes partnership common data model.



▨	EHDEN (21 Data Partners)	Brazil (Group IDAF/CIDACS/Fiocruz-BA)
Database size	Between 400 thousand and 39 million records	24 million records
Knowledge of the OMOP CDM tables	52% of PDs already familiar with most tables 38% familiar with none/few/some	16% of PDs already familiar with most tables 84% familiar with none/few/some
Team size	4.6 people (taking into account the average number of respondents per DP in the EHDEN survey)	6 people
Expertise with the data source	90% competent, proficient or expert	30% competent, proficient or expert
How many hours a week can you dedicate to this project?	14% of PDs from 33-44h	50% of the 33-44h team
Number of times the DQD was run	Average of 03 executions per DP	02 executions

Figure 3 highlights relevant information comparing the EHDEN data partners and the Brazilian team in terms of the team's experience with OMOP CDM tables, knowledge of the source data, number of data source records, size of the data source, hours dedicated to the project and number of DQD iterations.

Conclusion

The experience of the IDAF group in standardizing the data and implementing the ETL process for converting gestational syphilis data into the OMOP CDM model demonstrates the technical feasibility of adopting the EHDEN network methodology in contexts of the Global South, even in the face of the challenges of adapting and accommodating information from non-clinical sources.

The comparison of execution times between Brazil and European countries shows not only the contextual and operational differences, but also the ability of the Brazilian team to adapt and learn when applying good international practices.

The challenges faced, such as complex mapping, standardization of non-clinical data and the need for technical training, were overcome through collaborative strategies, expert consultancy, intensive use of OHDSI community tools and strong interdisciplinary commitment. The lessons learned from this experience strengthen health data governance in the Brazilian context and provide relevant input for other standardization initiatives in low- and middle-income countries. This work reinforces the importance of global reference models that are sensitive to local realities and highlights the need to expand the adoption of interoperable solutions in research networks in the Global South.

References

1. CIDACS – Center for Data and Knowledge Integration for Health. About us. Oswaldo Cruz Foundation – Bahia. Available at: https://cidacs.bahia.fiocruz.br/sobre/quem-somos. Accessed June 25, 2025.

2. CIDACS – Center for Data and Knowledge Integration for Health. Partnership between CIDACS and the Provincial Health Data Centre. Oswaldo Cruz Foundation – Bahia. Available at: https://cidacs.bahia.fiocruz.br/projeto/parceria-entre-o-cidacs-e-o-centro-de-dados-de-saude-da-provincia-do-cabo-ocidental/. Accessed June 2025.

3. Voss, E. A., Blacketer, C., van Sandijk, S., Moinat, M., Kallfelz, M., van Speybroeck, M., Prieto-Alhambra, D., Schuemie, M. J., & Rijnbeek, P. R. (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. Observational Health Data Sciences and Informatics. The Book of OHDSI. The Book of OHDSI. Accessed Jun 25, 2019. Available at: https://ohdsi.github.io/TheBookOfOhdsi/

5. *OHDSI/WhiteRabbit [program].* Version V0.10.1. GitHub. Accessed Jun 25, 2023. Available at: https://github.com/OHDSI/WhiteRabbit

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