Challenges and opportunities in adopting OMOP-CDM in Brazilian healthcare: a report from Hospital Israelita Albert Einstein

Maria Tereza Fernandes Abrahão[1], Uri Adrian Prync Flato[1,2], Mateus de Lima Freitas[1], Diogo Patrão[1], Amanda Gomes Rabelo[1,2], Cesar Augusto Madid Truyts[1,2], Gabriela Chiuffa Tunes[1], Etienne Duim[1], Gabriel Mesquita de Souza[1], Soraya Yukari Aashiro[1], Edson Amaro Jr[1], Adriano José Pereira[1,2]

1. **Big Data - Hospital Israelita Albert Einstein, São Paulo, SP, Brazil**
2. **Einstein Network of Intensive Care Medicine, São Paulo, SP, Brazil**

Background

Brazil has one of the largest and most complex public health systems in the world[1] with concerns related to discrepancies in data quality, lack of interoperability among federated states and lack of informatization (paper based medical records). Broader use of real-world data and classification standards are particularly challenging, for instance, since health conditions and related services or procedures are not linked. Implementing a common standardized data model is seen as an opportunity to integrate a multifaceted healthcare system in our country[2,3]. Since 2018, Hospital Israelita Albert Einstein (HIAE), in Sao Paulo - Brazil, started a journey to implement OMOP standards, and since 2022 has become a partner of the OHDSI Community. This report describes our approach to overcoming data format and ontology challenges during the implementation of an OHDSI-OMOP database in a not-for-profit quaternary hospital in Brazil.

Materials and Methods

A descriptive case study of the OMOP - Common Data Model implementation in a Brazilian hospital was performed from 2018 to 2023 after implementation of CERNER Eletronic Health Record platform (2017). The local HIAE clinical data vocabulary was translated into English and imported into USAGI tool (V1.4.3) for mapping standard vocabulary OMOP, by multidisciplinary groups (gathering data experts and representatives of different hospital stakeholders). Additionally, we describe the direct conversion processes from the information contained in the Electronic Medical Record (EHR) and other internal databases. Data mapping, extraction, Transform and Load (ETL) processes were developed to enable vocabularies embedded in our EHR. The descriptions of the steps taken highlight the challenges observed, and the actions adopted.

Results

We identified four main challenges:

Availability of qualified and dedicated team: The implementation of OMOP-CDM requires expertise in mapping, transformation, and data quality assessment. Organizations may struggle to find professionals who possess the necessary skills and experience to work on this type of project. The consequences can range from delivery delays to increased costs. We managed to overcome this challenge with a dedicated team consisting of 1 project manager, 2 data scientists, 2 data engineers, 2 clinical consultants, and 1 analyst consultant.

Lack of standardized internal vocabulary: The absence of a standardized internal vocabulary and barriers related to Portuguese language pose challenges during implementation. Local terminology and coding systems may not align with international standards, making mapping and harmonizing the data difficult. This can hinder data interoperability and affect the accuracy and consistency of analyses. To solve this issue, we direct efforts to teach the teams responsible for data collection and storage to convert local codes into standardized vocabularies using the USAGI tool in each sector (ex: pharmacy, laboratory).

Need for adequate infrastructure: Implementing an OMOP-CDM requires appropriate computing resources and infrastructure. This includes storage, memory, and processing capabilities to handle large-scale healthcare data. Ensuring the availability and scalability of infrastructure can be a significant challenge for organizations, especially when dealing with complex and diverse datasets. The HIAE OMOP architecture is based on Hadoop cluster with 168 cores, 655 GB of memory, and 32.4 TB of storage (Figure 1). This cluster is shared with additional applications and can be reached through Impala or Spark. The Oracle Database serves as the transactional database for the Hospital Information System (HIS). Spark scripts copy specific tables to Hadoop Datalake daily.

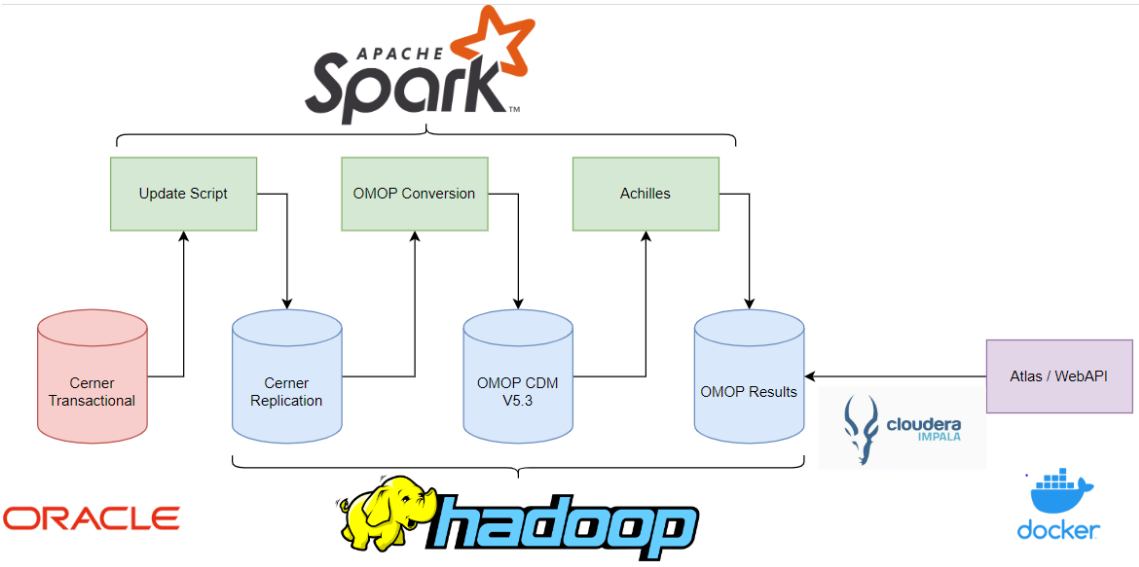

Figure 1: overview of data flow from source database to OMOP tables (Cerner) HIAE

A new structure has been designed as a dedicated infrastructure. Leasing cloud facilities is not an easy option because of the high volumes of sensitive data, and it must be planned at the institution level. The support received was fundamental for directing resources and facilitating the implementation of OMOP at the institutional level.

Development of Extract, Transform, Load (ETL): Extracting data from different sources, transforming it into the OMOP-CDM format, and loading it into the target database requires well-designed and efficient (ETL) processes. Developing these ETL pipelines can be complex, time-consuming, and require data integration and transformation expertise. ETL construction implies a deep knowledge of the current hospital information system (HIS) of the institution and a knowledge of the CDM model. The queries were hand-written by HIAE ETL professionals, and all data connected with patient records that failed the quality review has been deleted (though it may be imported again if the entries are repaired in the transactional database). An R-based OHDSI library then ran the Achilles script and Data Quality Dashboard, delivering this analysis and facilitating database visualization in accordance with the OMOP-CDM. The total amount of

data mapped per domain and the number of records with mapped terms are presented in Table 1.

Table 1: OMOP CDM-HIAE mapping vocabularies, April 2023

Table	Total Rows	Rows with mapped terms	% Row with mapped terms
Conditions_Occurrence	10,632,396	9,944,269	93.5
Procedure_Occurrence	12,684,959	6,476,230	51.0
Measurements	561,293,236	38,860,701	6.9
DrugExposure	52,657,745	40,133,631	76.2
DeviceExposure	1,817,221	-	-
Observation	495,727,962	35,963,951	7.2
Visit_Occcurrence	14,553,204	12,154,085	83.5
Visit_Detail	25,506,192	12,865,096	50.4
Person	2,863,208	2,863,208	100.0

Conclusion/final remarks

The value of this submission is that it represents new energy in adopting OMOP-CDM and committing to the principles of open science. The feasibility of implementing OMOP-CDM at HIAE can be attributed to three main factors: presence of a committed staff, adequate IT resources and institutional / financial support. Throughout all phases of the study, meticulous attention was devoted to safeguarding the confidentiality of sensitive data. This was achieved by adhering to a naming structure that guarantees the anonymity and security of data in accordance with Brazilian legislation, specifically the Data Protection and Privacy Law (DPGL - inspired in the General Data Protection Regulation - GDPR implemented in Europe). The potential impacts within the institution can go beyond research, for example, exploring new uses in the health network management system, clinical assistance for rare diseases and even in commercial partnerships. The following steps involve looking for ways to expand the experience across the country incorporating data of our Brazilian public and private health systems and conducting large-scale studies and contributions to the community through international partnerships. We believe sharing this experience can encourage other institutions to adopt this model, as well as provide perspectives on dealing with the inherent challenges of the process.

References

1. Brazilian National Digital Health Strategy 2020-2028. http://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf; acessed june 2023
2. Abrahao MTF, Freitas ML, Flato UAP, et al; Common data models in intensive care medicine during COVID-19 pandemics: the Hospital Israelita Albert Einstein experience. EINSTEIN (SÃO PAULO), v. 20, p. S1-S16, 2022
1. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahão MTF et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun. 2020 Oct 6;11(1):5009. doi: 10.1038/s41467-020-18849-z. PMID: 33024121; PMCID: PMC7538555