



Contribution ID: 311

Type: Poster

Implementation of the OMOP ETL pipeline for the standardization and integration of data from inpatients with respiratory diseases in Douala General Hospital, Cameroon

Monday 13 October 2025 19:10 (20 minutes)

Douala General Hospital is a first-class hospital in Cameroon where we meet a multidisciplinary medical team treating several thousand patients each year. This hospital hosts numerous patient records that may be useful for public health research. However, majority of these records are paper-based, hence limiting their exploitation. For some cases, particularly the pulmonology department, the data of hospitalized patients are recorded in heterogeneous datasheets mainly collected for research purposes without standardization or uniform structure. This data system also limits the exploitation for clinical research, care management and decision-making. Furthermore, the lack of standardization limited the integration of data within broader health systems, thus hindering its secure and reusable sharing.

To address this challenge, we undertook the implementation of a complete ETL pipeline aligned with the OMOP Common Data Model (CDM) version 5.4, an internationally recognized framework for standardizing health data. Our objective was to transform, standardize, and integrate patient data from the pulmonology department of this hospital into a database compliant with the FAIR (Findable, Accessible, Interoperable, Reusable) principles to facilitate their reuse for research, clinical management, and patient monitoring. This approach aimed to improve the quality of hospital data, strengthen its interoperability with other systems, and lay the foundation for advanced use in data science for healthcare. We made use of patient level dataset on Respiratory illness, which included more than 120 variables covering a wide range of clinical and administrative variables such as sociodemographic data, medical history, clinical signs and symptoms, laboratory results, final and secondary diagnoses, and medical observations, as well as information on hospital stays. This data was extracted from datasheets, medical records, and papers, presenting varied forms, heterogeneous levels of completeness, and missing observations.

To standardize this data, we made use of a number of OHDSI tools such as: WhiteRabbit for data profiling, USAGI for vocabulary mapping from our source vocabularies to OMOP standardized concepts/vocabularies and Rabbit in a Hat for data mapping of the source tables to the standard OMOP CDM tables, including: Person, drug_exposure, Measurement, visit_occurrence, condition_occurrence, and observation. The concepts used in the mappings were derived from SNOMED, LOINC, and RxNORM vocabularies, while integrating adaptations specific to the local context.

The ETL pipeline of this data was based on SQL skeleton files exported from Rabbit in Hat after the data mapping. These scripts were then customized using pgAdmin interface for PostgreSQL. After creating the OMOP tables using scripts from OHDSI's GitHub, we loaded our transformed data into the OMOP PostgreSQL databases, structuring them in accordance with the model.

To ensure the quality and compliance of our standardized database, we used the Achilles tool (developed in R), which automatically checked the completeness, conformance, and plausibility of the transformed data. This tool achieved an overall data quality score of 97%, thus attesting to the reliability of the ETL pipeline and the robustness of the database.

Standardizing Respiratory illness data to OMOP CDM in the African context is a novel and promising field that will set pace for collaboration, data sharing and interoperability across health systems. This work made it possible to establish a standardized database that meets all FAIR requirements and is ready to be used for analysis, clinical research, and data science, thus opening up a perspective for decision-makers who can benefit from data-driven decisions to improve pulmonology care practices in Cameroon.

Primary authors: KIRAGGA, Agnes (Infectious Diseases Institute, College of Health Sciences, Makerere University, Kampala, Uganda); MBATCHOU NGAHANE, Bertrand Hugo (Douala General Hospital, Data Science Without Borders project, Cameroon); MBOUAMBA YANKAM, Brenda (Douala General Hospital, Data Science Without Borders project, Cameroon); ANICET ONANA AKOA, François (Douala General Hospital, Data Science Without Borders project, Cameroon); BLAISE EBIMBE, Jean (Douala Gynaeco-Obstetric and Pediatric Hospital (DGOPEH), Cameroun); FANKOUA TCHAPTCHET, Luc Baudoin (Douala General Hospital, Data Science Without Borders project, Cameroon); BARASA, Miranda (African Population and Health Research Center (APHRC), Nairobi, Kenya); ANDESO, Pauline (African Population and Health Research Center (APHRC), Nairobi, Kenya); IDDI, Samuel (Department of Statistics, University of Ghana, Accra, Ghana)

Presenters: MBOUAMBA YANKAM, Brenda (Douala General Hospital, Data Science Without Borders project, Cameroon); ANICET ONANA AKOA, François (Douala General Hospital, Data Science Without Borders project, Cameroon); FANKOUA TCHAPTCHET, Luc Baudoin (Douala General Hospital, Data Science Without Borders project, Cameroon); BARASA, Miranda (African Population and Health Research Center (APHRC), Nairobi, Kenya); IDDI, Samuel (Department of Statistics, University of Ghana, Accra, Ghana)

Session Classification: Poster Session

Track Classification: SciDataCon Persistent Themes: Open Data, FAIR Data, Innovation, Industry and Development