Framework and approach to establish a multi-registry OMOP cluster with shared OHDSI tooling and access specificity

Jared Houghtaling[a], Emma Gesquiere[a], Lisa McDowell[b], Cathy Enright[b], Claire Donohoe[b], Tom Feusels[a], and Lars Halvorsen[a]

[a] edenceHealth NV
[b] Trinity Saint James Cancer Institute, Dublin (TSJCI)

Background:

The Trinity Saint James Cancer Institute (TSJCI), located in Dublin, Ireland, manages and curates Irish cancer registries that catalogue patients with specific oncological presentations **[1]**. As part of the European Health Data Evidence Network (EHDEN), TSJCI and edenceHealth are transforming eight of those registries into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM); namely for upper gastrointestinal, urological, colorectal, breast, skin, head-and-neck, gynaecological, and lung cancers. Moreover, once these registries are transformed, they are merged into a ninth 'merge' data source that captures all registry members and their OMOP events in a single dataset. The work presented here details this transformation process and aims to describe the following: (1) semantic and structural aspects of the data harmonization as well as associated challenges, (2) technical details regarding infrastructure and data access, and (3) prospects for collaboration with cancer registries both within and outside of Europe.

Methods:

To transform all eight cancer registry data sources, we created a single, flexible Extract-Transform-Load (ETL) process with an environment variable defined at the beginning of the run that specifies which source registry should be transformed. Some transformations are source dependent; information regarding demographics (PERSON) and DEATH varies slightly across registries. Filling of OMOP event tables, however, is consistent across sources and depends on a central semantic mapping file. The source data is organized in a wide format, as is typical for (cancer) registries. We employed a STEM table as an intermediate transformation step, which we constructed by pivoting column-value pairs into extended rows using information about dates, domains, and standard mappings in the semantic mapping document. TSJCI mapped more than 2500 unique source variables using tools created by edenceHealth called edenceMapper (suggestion generator) and edenceReviewer (web application for collaborative semantic mapping) **[2]**. Once all eight sources were transformed, we executed a 'merge' ETL in order to extract the OMOP CDM data out of the eight independent, registry-specific OMOP databases, and then to load and consolidate that data into a single OMOP source.

We launched the ETLs as Docker processes on a dedicated virtual machine (Ubuntu, VMWare) hosted on TSJCI premises; we deployed all other OHDSI and EHDEN tooling in a similar fashion. Each data source has its own separate PostgreSQL database instance; the

environment variable described previously routes the output from the transformations to the respective database instance. In order to access the necessary web applications, users create an ssh connection to the virtual machine (via Putty on Windows workstations) and tunnel the necessary ports. Users can then interact with the applications at their localhost address and the application-specific port(s).

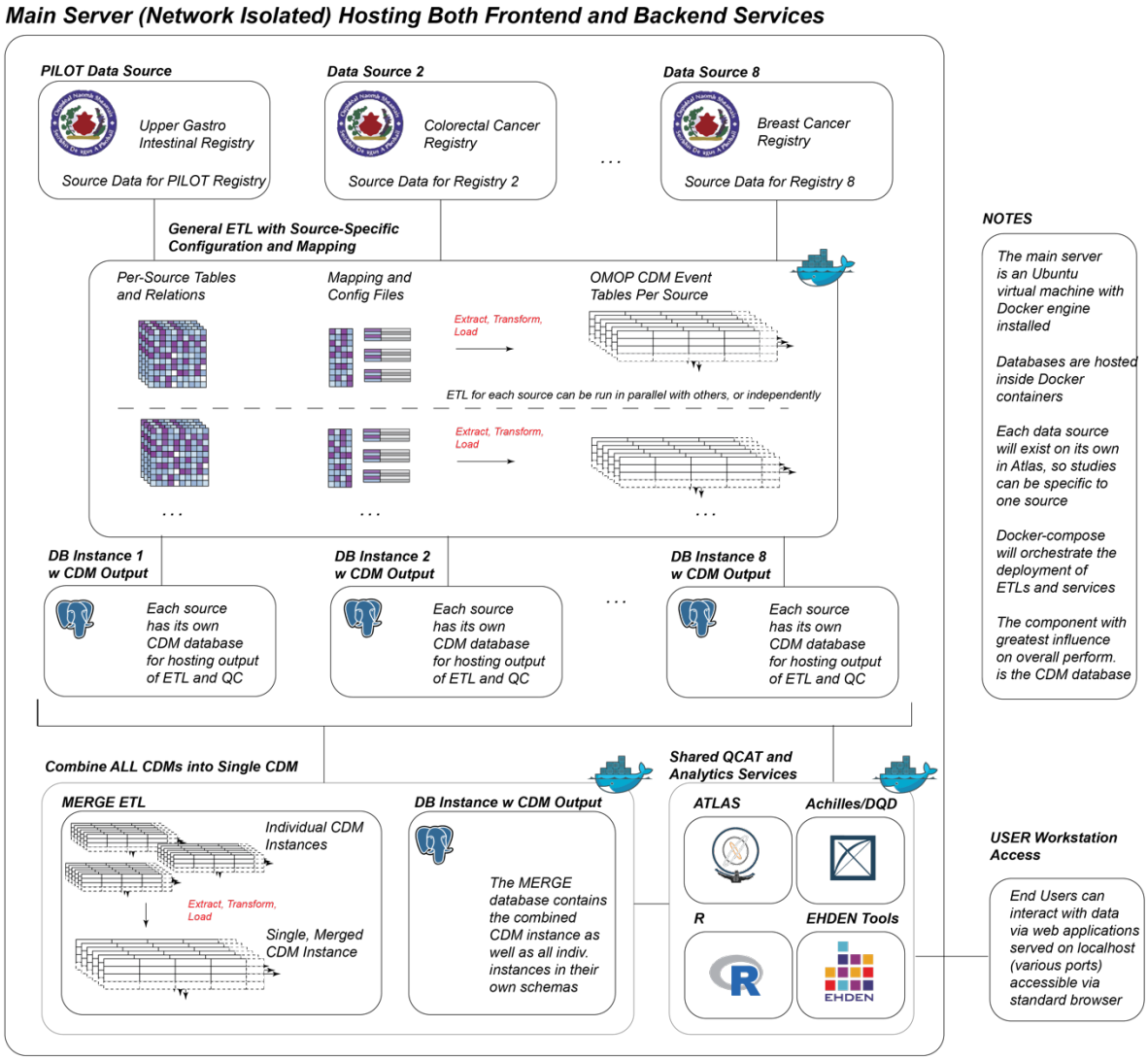

Figure 1. Overview of the technical deployment structure of the data transformation and tooling. All processes – ephemeral and persistent -- are executed inside of Docker containers, and bash scripts automate deployment steps, using docker-compose for source-specific configurations.

Results:

As mentioned above, this ETL design is centered around the STEM table. This table captures all clinical events, which are then routed to a relevant OMOP table based on the domain of their mapped standard concept identifier. TSJCI mapped more than 2500 source variables to OMOP standards using the edenceReviewer platform. Each of these source codes also needed to be linked to other relevant information, such as the associated start and end dates (via column

reference), *value_as_concept_ids*, *unit_concept_ids*, etc. This ETL design allows for easy modification of mappings, future updates to any of the source tables (e.g. adding a field with relevant information for a particular study), and future additions of new cancer registries, all without extensive technical expertise. Importantly, user access within Atlas can be restricted per registry source by changing the user role configuration within the Atlas application, which is critical for enabling and encouraging Atlas use across the organization for users with varying degrees of data access privileges. **Table 1** shows an overview of top 5 most common concepts per domain in the MERGE dataset, with record and person counts rounded to the nearest 100.

Table 1. Most common OMOP concepts and their associated domain/counts in the MERGE dataset

Concept Description	OMOP Domain	Records[a]	Persons[a]
No evidence of	Observation	11800	5000
Body mass index	Measurement	7500	7100
Body height measure	Measurement	7500	7100
Follow-up encounter	Observation	7100	6100
Tumor site	Observation	6200	5200
Body weight measure	Measurement	6100	5800
cM category	Observation	5200	5000
cN category	Observation	5000	4700
Pre-operative chemotherapy	Procedure	3340	2000
Postoperative chemotherapy	Procedure	3270	2000
Combined radiotherapy	Procedure	3140	2000
Regional lymph nodes positive [#] Specimen	Measurement	3000	1400
Histopathology finding	Condition	2800	2300
Number of lymph nodes examined	Measurement	2700	2500
Surgical procedure	Procedure	2640	1080
Radiation oncology AND/OR radiotherapy	Procedure	2400	1100
Tumor size finding	Condition	1700	800
Moderately differentiated histological grade finding	Condition	1700	600
Primary malignant neoplasm of prostate	Condition	1500	1400
Venous (large vessel) invasion by tumor present	Condition	1200	200
fluorouracil	Drug	300	300
oxaliplatin	Drug	300	300
leucovorin	Drug	200	200
docetaxel	Drug	100	100
Cisplatin	Drug	100	100

[a] Counts rounded to nearest 100

One major challenge we have encountered working with multiple data sources is that each source has its own intricacies and quirks. For this reason, we have recently deployed the *Ares* network quality tool to gauge the mapping coverage and source/release-specific quality issues across the nine sources **[3, 4]**. This tool has been invaluable for (1) comparing sources with regard to their quality, (2) updating mappings to ensure consistency across the network, and (3) pre-coordinating study design by investigating measurement values and other OMOP event counts across the different sources in one place.

Conclusion:

Through the harmonization and subsequent validation processes, we have uncovered new insights into our source registry data. We are in the process of connecting with other registries across Europe to coordinate federated studies and discuss best practices. The details shared in this work are merely a glimpse into the ongoing work at TSJCI to integrate OMOP CDM data into daily analytical routines; we expect that moving forward, this new dataset will continue to grow in quantity, quality, and importance within our organization and beyond.

References:

[1] *Trinity St. James's Cancer Institute - Trinity Development & Alumni - Trinity College Dublin*. (n.d.). https://www.tcd.ie/alumni/inspiring-generations/cancer-institute/

[2] edenceHealth. (n.d). *HOME - edenceHealth NV*. edenceHealth NV. https://edence.health/

[3] Ohdsi. (n.d.). *GitHub - OHDSI/AresIndexer: R package that creates the index and relevant files for an Ares deployment.*GitHub. https://github.com/OHDSI/AresIndexer

[4] Ohdsi. (n.d.-a). *GitHub - OHDSI/Ares: A Research Exploration System*. GitHub. https://github.com/OHDSI/Ares