# OMOPification of real world cancer data to enable privacy-preserving analytics for cancer research

*An implementation at the Maastricht University Medical Centre+ for the Digital Oncology Network for Europe*

**Background:** Converting Metastatic Non-Small Cell Lung Cancer (mNSCLC) data from Electronic Health Records (EHR) into the OMOP CDM to examine survival analysis of patients based on Metastasis Location and Line of Treatment and to showcase the feasibility of federated learning using data directly extracted from hospital's EHR.

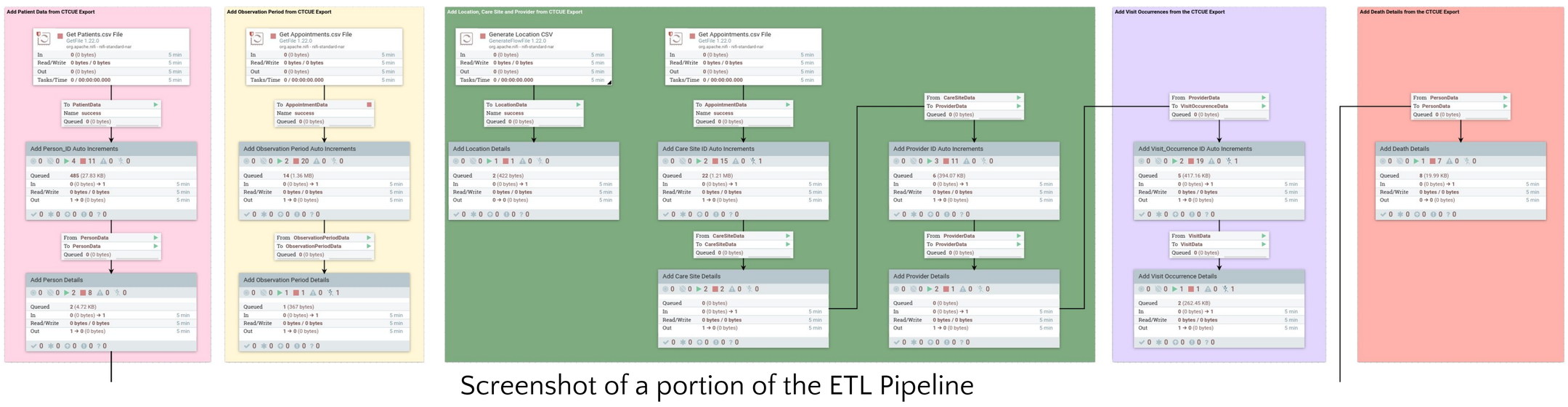
**Result**: 2000 patients converted to for the study with a set of 26 Data Concepts.

Data Quality Dashboard showing the quality to be ~ 100%

| | Verification | | | | Validation | | | | Total | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pass | Fail | Total | % Pass | Pass | Fail | Total | % Pass | Pass | Fail | Total | % Pass |
| Plausibility | 2213 | 0 | 2213 | 100% | 287 | 0 | 287 | 100% | 2500 | 0 | 2500 | 100% |
| Conformance | 949 | 2 | 951 | 100% | 157 | 0 | 157 | 100% | 1106 | 2 | 1108 | 100% |
| Completeness | 448 | 0 | 448 | 100% | 17 | 0 | 17 | 100% | 465 | 0 | 465 | 100% |
| Total | 3610 | 2 | 3612 | 100% | 461 | 0 | 461 | 100% | 4071 | 2 | 4073 | **100%** |

# Methods

- A **top-down approach** defined by the consortium was used to define how to populate OMOP tables.
- An **NLP based approach** with human validation was used with the CTcue software to find patients in the EHR system and to structure 7 data concepts available in medical notes.
- Apache NiFi was chosen to build the ETL pipeline for OMOP conversion due to its **open-source nature, web interface, and fast performance**.




Screenshot of a portion of the ETL Pipeline

**Conclusion:** A successful workflow was developed and tested for 26 data concepts for a population of mNSCLC patients. However, faulty records were found that can be explained by incorrect primary diagnosis dates. Structuring data with NLP has shown great potential, however model quality needs to improve to accelerate high-quality data curation without or with little human intervention.

Maastricht University

Prabash Galgane Banduge, Anne-Lore Bynens, Cedric Gillissen, Andre Dekker, Petros Kalendralis, Pascal Suppers, Alberto Traverso, Lizza Hendriks, Aiara Lobo Gomes

Maastricht UMC+