High robustness of OHDSI tools and OMOP CDM to support transformation of lung oncology real world data

Title: Transforming lung cancer EHR data into the OMOP CDM: A case study of Non-Small Cell Lung Cancer

Background: Common Data Models (CDMs) are essential for data harmonization, leading to significant improvements in healthcare and research domains. CDMs enhance transparency, increase the reliability of medical research, and ensure efficient, timely evidence generation for decision-making[1]. Despite the continuous progress and development of CDMs in oncology real-world data[2,3], challenges remain.
Lung cancer is the leading cause of cancer-related mortality worldwide, with an estimated 2.2 million new diagnoses and 1.8 million deaths annually cases[4]. Non-Small Cell Lung Cancer (NSCLC) accounts for 80-85% of these cases[4].
This study evaluates the adoption of the OMOP CDM for lung cancer oncology real-world data, exploring the opportunities and challenges of implementing the OMOP CDM in lung oncology data.

Results:

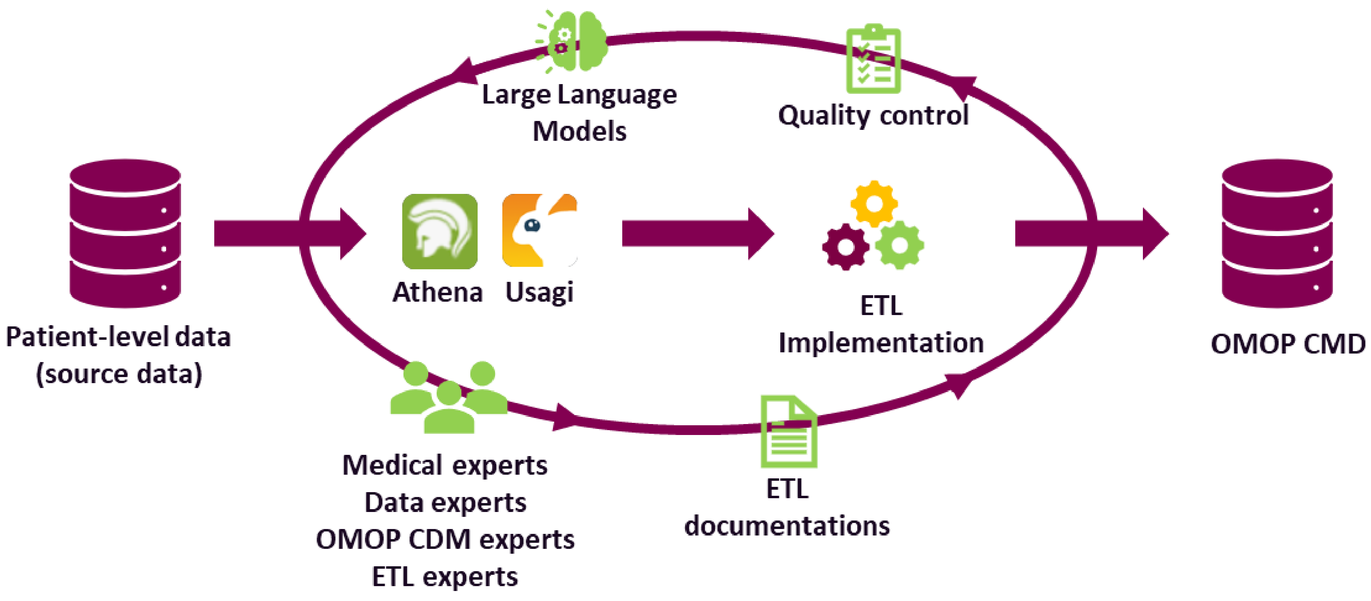
- Applied to anonymized clinical and laboratory data revealed a high success rate, with over 99% of fields effectively transformed into OMOP CDM concepts, affirming the robustness of the data transformation process.
- American Joint Committee on Cancer (AJCC) cancer staging manual (eighth edition) able to accurately translated cancer stages while retaining essential clinical details.

Methods and Material:

- Database: Flatiron Enhanced Datamart (EDM), a subset of patients with Non-Small Cell Lung Cancer (NSCLC).
- Observational and retrospective data of over 90.000 anonymized patients.
- Data mapping using ATHENA, USAGI and in-house R and SQL pipelines.
- Quality assessment using in-house R and SQL pipelines.
- Large Language Model (ChatGPT 4).

Highlights

- Tumour progression and metastasis were effectively integrated into the *Episode* and *Episode Events* tables, with additional mapping to the *Observation* and *Observation period* tables to ensure comprehensive capture of these events within the ATLAS[5].
- Drug regimens were also mapped to *Drug Exposure* and *Drug Era* tables.
- Limitations encountered during the ETL process was the transformation of general data concept or outer subsets categories (e.g., "other types of mutations").
- Line of therapy were mapped with HemOnc vocabulary.
- ChatGPT provided significant robust solutions in ELT implantation and accelerated the preparation of ETL documentation.



Conclusions:

- This study highlighted the significant challenges in mapping NSCLC patient data to the OMOP CDM and presented a framework for addressing these challenge.
- We underscored the importance of collaboration and quality assurance measures in ensuring data accuracy and reliability in oncology.
- We demonstrated the potential of a common data model to support large-scale clinical and translational research initiatives.
- Large Language Model can lead to more efficient ETL workflows and improved decision-making capabilities.

References:

1. Kent, S. et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 39, 275-285 (2021).
2. Osterman, T. J., Terry, M. & Miller, R. S. Improving cancer data interoperability: the promise of the Minimal Common Oncology Data Elements (mCODE) initiative. *JCO Clinical Cancer Informatics* 4, 993-1001 (2020).
3. Carus, J. *et al.* Mapping the Oncological Basis Dataset to the Standardized Vocabularies of a Common Data Model: A Feasibility Study. *Cancers* 15, 4059 (2023).
4. International Agency for Research on Cancer, I. W. *Globocan 2022 Fact Sheet - Cancer today*, <https://gco.iarc.fr/today/en> (2022).
5. https://www.ohdsi.org/software-tools/

Evangelos Chandakas (Handakas), Ping Sun