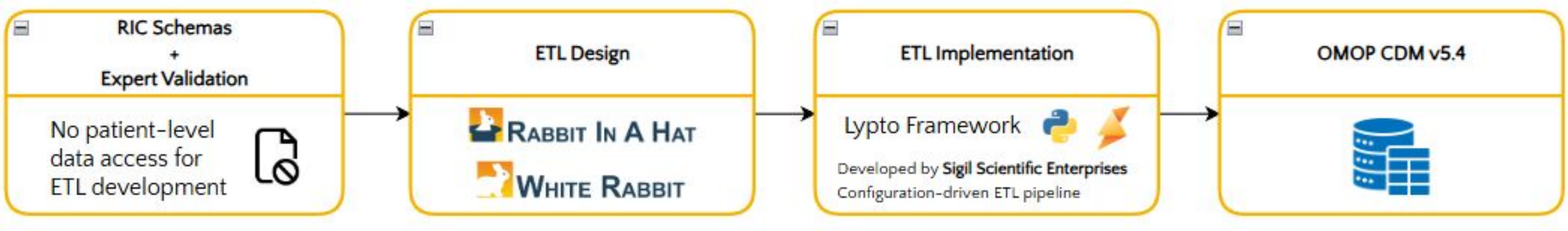
Standardizing the "Repositório Integrado de Conhecimento" at IPO Porto: Design and Implementation of an OMOP CDM ETL for Oncology Data

Mariana Canelas-Pais[1,2,3], Renata Silva[1], Sofia Gomes[1], Tiago Taveira-Gomes[2,4], Rita Rb-Silva[5], Maria José Bento[5,6,7], Teresa Garcia[5,6]
[1]MTG Research and Development Lab, Porto, Portugal
[2]Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine, University of Porto, Porto, Portugal
[3]RISE-Health - Centre for Health Technologies and Services Research, Porto, Portugal
[4]SIGIL Scientific Enterprises, Dubai, United Arab Emirates
[5]Group of Epidemiology, Outcomes, Economics and Management in Oncology – Research Center, Porto Comprehensive Cancer Center (Porto.CCC) & RISE@CI-IPOP (Health Research Network), Portuguese Oncology Institute of Porto (IPO Porto), Porto, Portugal
[6]Department of Epidemiology, Portuguese Oncology Institute of Porto (IPO Porto), Porto, Portugal
[7]Population Studies Department. School of Medicine and Biomedical Sciences, ICBAS, University of Porto, Porto, Portugal.
mariana.pais@mtg-research.com

Background

IPO Porto is a national reference center for oncology care in Portugal. Its clinical registry, Repositório Integrado de Conhecimento (RIC), consolidates oncology data through a mix of automated ingestion and extensive manual entry by clinical experts. Despite its quality, coding and structural heterogeneity limit interoperability and secondary use[1]. This work presents the OMOP-CDM ETL design and planned next steps for extension[2].

Methods



Results

The first version cover the **PERSON**, **VISIT_OCCURRENCE**, **CONDITION_OCCURRENCE**, **MEASUREMENT**, **DEATH**, **CDM_SOURCE**, **OBSERVATION_PERIOD** domains. The ETL design decisions included:

- **ICD-O-3 Mapping:** Histology, behaviour, and topography concatenated for CONDITION_OCCURRENCE mapping.
- **Wide-to-Long Transformation:** Tumor characteristics were transformed into multiple longitudinal records across MEASUREMENT and CONDITION_OCCURRENCE.
- **Cancer Modifiers:** Staging and tumor attributes mapped to OMOP Cancer Modifier concepts in MEASUREMENT.
- **Temporal Proxies:** Tumor diagnosis date used as proxy when comorbidity onset dates were unavailable.

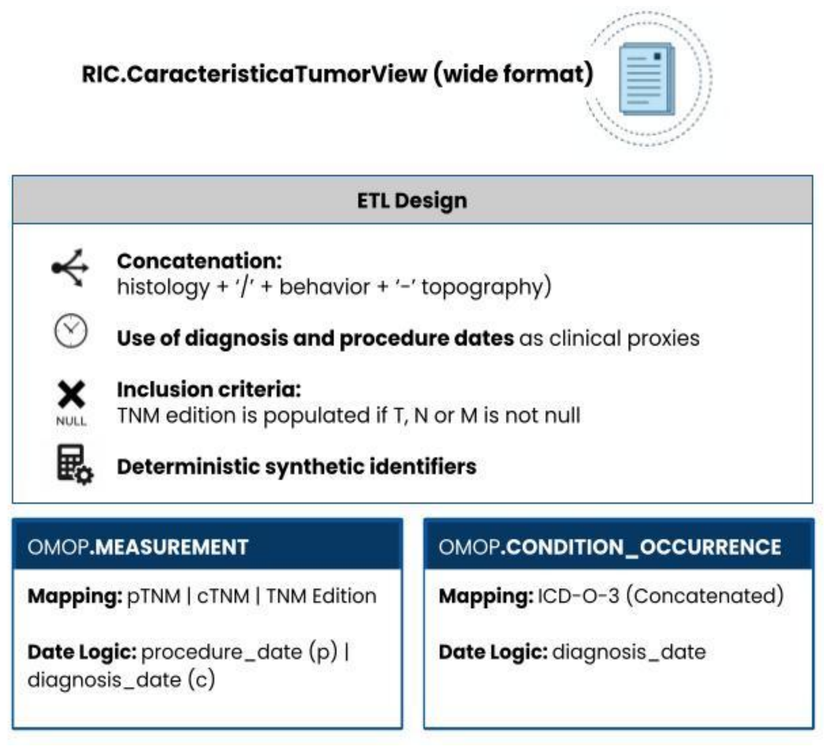


Figure 1. ETL design specification for the RIC.CaracteristicaTumorView source, detailing the transformation logic for mapping oncological data, TNM and ICD-O-3 classifications, into the OMOP Measurement and Condition_Occurrence tables.

Conclusion

This work demonstrates the feasibility of transforming a curated oncology registry into OMOP CDM through a formally specified ETL process. Future iterations will expand the analytical scope by capturing systemic therapies, surgical procedures, and critical clinical performance indicators like ECOG

References

1. Ajmal A, Bouissou O, Brash J, Cheeseman S, Banduge PG, Gomes AL, et al. Establishing standards: harmonising coding principles for a minimal cancer dataset in the OMOP Common Data Model. ESMO Real World Data and Digital Oncology [Internet]. 2025 Sept;9(100179):100179. Available from: http://dx.doi.org/10.1016/j.esmorw.2025.100179
2. Observational Health Data Sciences, Informatics. The Book of OHDSI [Internet]. 2021 [cited 2026 Feb 2]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/