The Onco Health Data Dock: A Data Lake for Real-Time mapping and harmonization in pediatric oncology

Background
Pediatric oncology stands at a critical juncture, where the integration of AI harbors the potential to revolutionize patient care. Despite the burgeoning interest and investment in AI applications within this domain, the absence of a strategic foundation significantly hampers the realization of its full potential. Addressing this gap, we introduce the data lake Onco Health Data Dock (OHDD), a pivotal shift towards an integrated and AI-enhanced healthcare.

Methods
The OHDD employs a sophisticated three-layer design implemented in MongoDB:
- the operational layer aggregates diverse data from various software systems;
- the research layer orchestrates collected data mapping them to CDM vocabularies;
- the delivery layer standardizes data into the OMOP CDM format, readying them for AI and ML applications, including NLP and LLMs.

Results
- Near-real-time data harmonization to the OMOP CDM;
- Secure and efficient integration of heterogeneous data sources;
- Data are easily accessible, processed, visualized and analyzed;
- A significant leap forward in the ethical and effective use of patient data, ensuring quality and relevance;
- AI-driven Data Integration.

On the right is reported a translated sample taken from our clinical notes and the respective model output mapped to standard concepts in SNOMED Vocabulary.

Limitations
- Mapping to OMOP CDM challenges:
  - automation of the entire mapping process;
  - combination of different entities for a meaningful association between predicted tokens and vocabularies;
  - definition of a mapping decision criteria when multiple associations are available.

AI enabled future perspectives
- identify specific patient characteristics that could predict total length of stay, time-to-surgery, and time to diagnosis, thus improving healthcare delivery.
- provide decision support, personalize treatment plans and optimize resource allocation within diagnostic-therapeutic care pathways.
- fine-tuning a pre-trained Italian Transformer for NER in the clinical domain.