The Onco Health Data Dock:

A Data Lake for Real-Time mapping and harmonization in pediatric oncology

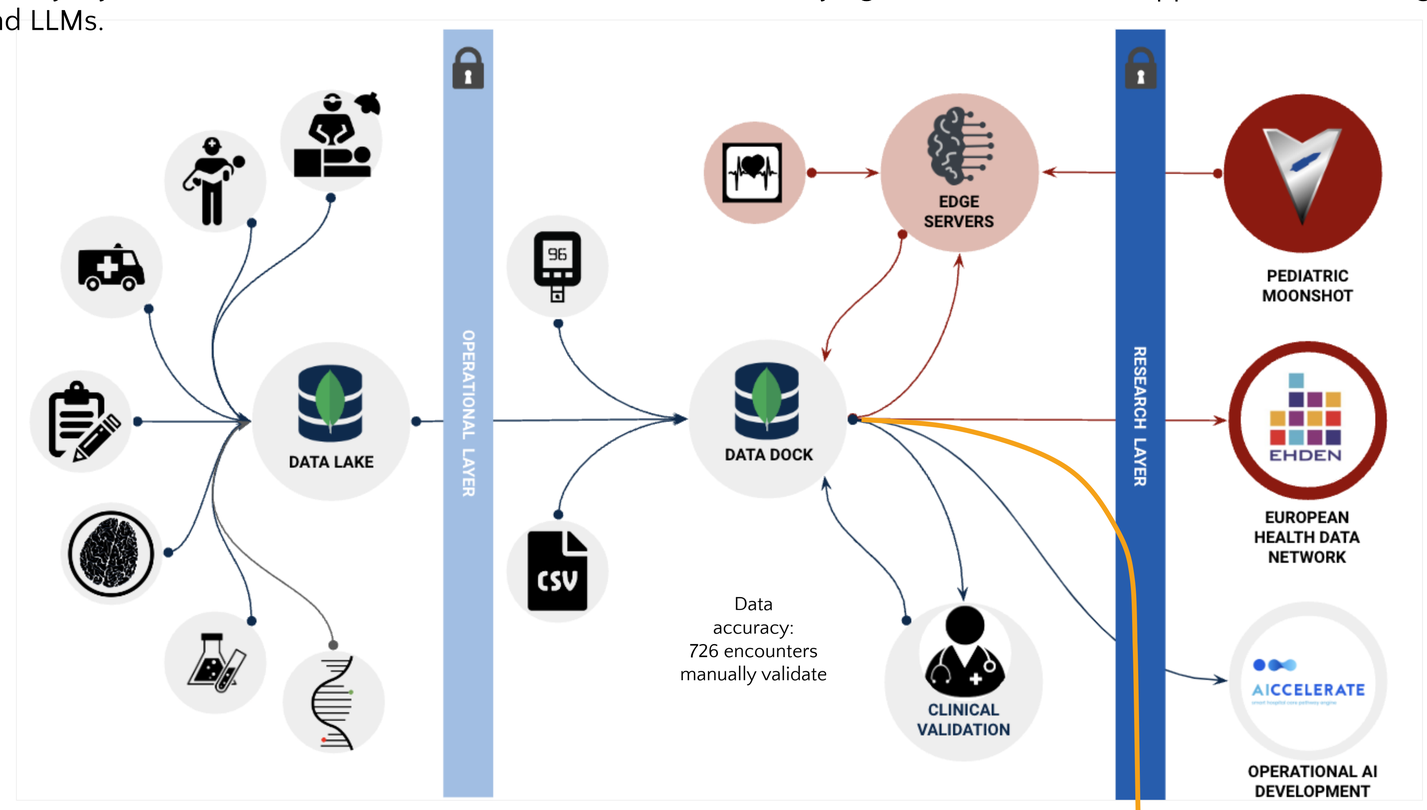
Background

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

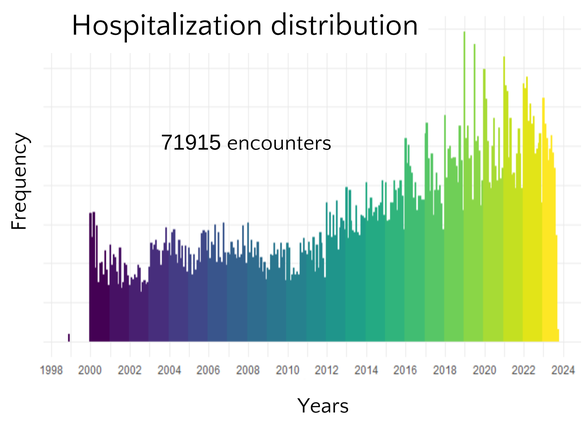
The OHDD employs a sophisticated three-layer design implemented in MongoDB:

- 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.



Our OPBG population

	Overall (N=2907)
Sex	
Female	1340 (46.1%)
Male	1567 (53.9%)
Age (years)	
Mean (SD)	19.2 (8.90)
Median [Min, Max]	18.4 [1.94, 60.4]
Period of observation (days)	
Mean (SD)	2720 (1960)
Median [Min, Max]	2160 [365, 8670]
Total admissions (per person)	
Mean (SD)	22.8 (37.7)
Median [Min, Max]	11.0 [2.00, 604]
Total operations (per person)	
Mean (SD)	1.11 (3.14)
Median [Min, Max]	0 [0, 28.0]
Total diagnosis (per person)	
Mean (SD)	6.09 (5.05)
Median [Min, Max]	5.00 [1.00, 38.0]



Transformer-based NLP pipeline for NER in Clinical Notes

OPBG dataset preparation

DL Translation
Chunked text is translated with model opus-mt-tc-big-it-en, published by EAMT. No efficient model for NER has ever been trained on Italian medical notes before.

Data Pre-processing
Data Cleaning, Normalization, Tokenization, Stop words Removal, Stemmization and Lemmization

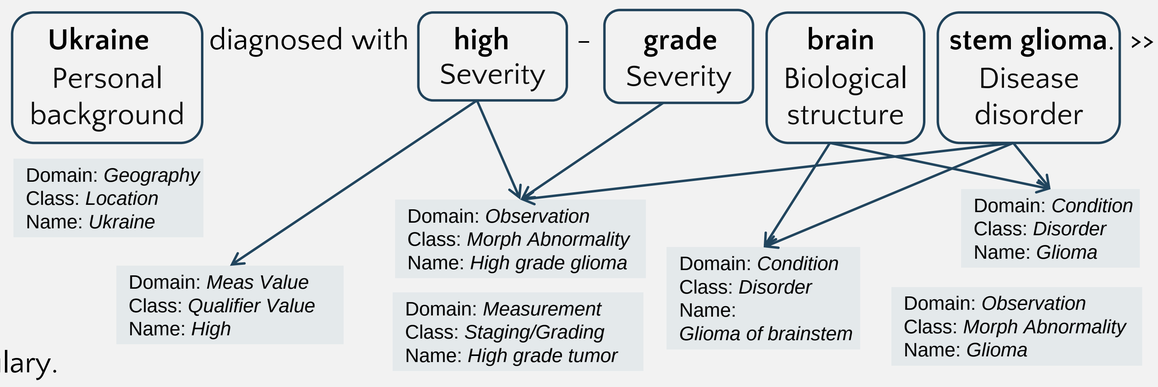
NLP model for NER
d4data/biomedical-ner-all trained on the MACCROBAT dataset.

Mapping to OMOP CDM, SNOMED Vocabulary.

Results

- Near-real-time data harmonization to the OMOP CDM;
- Secure and efficient integration of heterogeneous data sources;
- Data are easibly 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.

Bambino Gesù
OSPEDALE PEDIATRICO

OHDSI

Diana Ferro[1], Silvia Capuzzi[1], Giulia Nardini[1], Francesco Fabozzi[2], Antonella Cacchione[2], Andrea Carai[3], Alberto Tozzi[1], Angela Mastronuzzi[2]

[1] Predictive and Preventive Medicine Research Unit, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy
[2] Department of Oncology/Hematology, Gene and Cell Therapy and Hemopoietic Transplant, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy
[3] Department of Neuroscience and Neurorehabilitation, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy