**Abstract**:

Significant strides in medicine and technology have propelled the 5-year survivorship of childhood cancer in the US from 58% in the 1970s to 85% in 2023. However, childhood cancers remain the leading cause of disease-related deaths in the US among children, adolescents, and young adults. Our overarching goal is to build a computational tool that integrates and analyzes various public biomedical databases and provides physicians and scientists with the landscape of targets and available therapeutic agents that have been tested in childhood cancers. However, integrating and analyzing data from various public databases is non-trivial.

One such public database is the Clinical Trials Registry (ClinicalTrials.gov) which was launched by the National Institutes of Health (NIH) and contains 482,529 research studies from all 50 states in the US and 223 countries. While guidelines are in place to ensure data integrity, our initial inspection of the disease data revealed various inconsistencies in the form of extraneous information, typographical errors, missing values, etc. Furthermore, we observed that the tumor names do not necessarily follow standardized tumor names in the World Health Organization (WHO) Tumor Classification system or National Cancer Institute Terms (NCIT) and need to be identified from other diseases in the dataset.

Considering these factors, we designed a computational pipeline to identify and standardize tumors from the Clinical Trials Registry. The pipeline extracted over 107K diseases and identified 13,230 unique tumors, of which 6934 were pediatric. The pipeline employed different standardization techniques based on edit distances, text-embedding, and clustering, and compared their accuracy on a subset of tumor names from the registry. We found that text-embedding methods produced more accurate standardization of the tumors in the registry. The pipeline standardizes all the 13,230 tumors, which are manually validated and provided as a data table for further downstream data integration and analysis.