**Abstract**:

Advancements in medicine and technology have significantly increased the 5-year survivorship of childhood cancer in the US, from 58% in the 1970s to 85% in 2023. Despite this progress, childhood cancers remain the leading cause of disease-related deaths among children, adolescents, and young adults in the US. Our primary objective is to develop a computational tool that integrates and analyzes various public biomedical databases, offering physicians and scientists a comprehensive view of targets and therapeutic agents tested in childhood cancers. However, integrating and analyzing data from diverse public databases is non-trivial.

One such database is the National Institutes of Health's (NIH) Clinical Trials Registry (ClinicalTrials.gov). The registry contains 482,529 research studies from all 50 US states and 223 countries. Although guidelines exist to ensure data integrity, our initial analysis revealed that many fields lack adherence to nomenclature standards for diseases and interventions. In the context of childhood cancer, tumor names often do not align with standardized names from the World Health Organization (WHO) Tumor Classification System or National Cancer Institute Terms (NCIT). Therefore, these inconsistencies complicate any exploration into clinical trials for childhood cancers and their interventions.

Considering these factors, we designed a computational pipeline to identify and standardize tumor names from the NIH-Clinical Trials Registry. This pipeline extracted over 107,000 diseases and identified 13,230 unique tumor names, with 6,934 identified as pediatric. We then sought to categorize each tumor name according to their respective WHO and NCIT classifications based on edit distances, text-embedding, and clustering and compared their accuracy on a subset of tumor names from the registry. Our findings indicated that text-embedding methods significantly improved the accuracy of tumor name standardization. The pipeline successfully standardized all 13,230 tumor names, which were manually validated and provided as a data table for further integration and analysis of cancer clinical trials.