# Curating AI-ready datasets for pediatric oncology: the Pediatric Cancer Data Commons

Kaitlyn Ott, MS[1], Jooho Lee, PhD[1], Mei Li, MS[1], Luca Graglia, MS, MBA[1], Michael Watkins, PhD[1], Brian Furner, MS[1], Kirk D. Wyatt, MD, MAS[2], Samuel L. Volchenboum, MD, PhD[1]
[1]Department of Pediatrics, University of Chicago, Chicago, IL, USA, [2]Department of Pediatric Hematology/Oncology, Sanford Health, Fargo, ND, USA

**Background:** Small sample size and a lack of standardized data limits our ability to train AI models for pediatric oncology use cases. Opportunities for AI usage in pediatric oncology include improving diagnostic classification, risk stratification, and prediction of treatment response. Data standardization and harmonization efforts facilitate pooling of datasets from disparate sources into large datasets primed for artificial intelligence tasks. We present the efforts of the Pediatric Cancer Data Commons (PCDC) to generate large "AI-ready" datasets for pediatric cancer.

**Methods:** The PCDC collaborated with six disease groups to develop a working target data model. This process included building a consensus data dictionary in collaboration with clinical and data standards experts, using research study case report forms, and leveraging existing standards where possible for each disease group. This culminated in a target data model that was coordinated and formed from the six different consensus data dictionaries. Once a stable data dictionary was established for each disease group, de-identified datasets were harmonized to the target data model, transformed, and loaded into the data commons. Shared data governance structures–including data access policies–were developed.

**Results:** As of April 2024, the PCDC includes the largest known collection of harmonized data from over 40,000 children with cancer. Data include demographics, medical history, disease characteristics, molecular alterations, treatment, imaging, laboratory values and outcomes. Clinical and imaging data from the PCDC enabled one study that utilized the AI-ready dataset to train a neural network to predict response to chemotherapy based on initial imaging studies. The PCDC datasets include over 35,338 individual laboratory test values derived from approximately 17,517 patients that can be leveraged for biomarker studies.

**Conclusion:** We have demonstrated the utility of a standards- and consensus-based target model to develop AI-ready datasets for pediatric cancer. A structured common data model provides a solid foundation for AI initiatives by ensuring data quality, consistency, and interoperability. As additional disease groups and patients are represented in the data commons, we anticipate uses of the dataset to train machine learning models to increase. We expect future machine learning models trained on this dataset to uncover new insights that will improve diagnosis, risk stratification, and treatment response prediction for pediatric cancer patients.