Abstract title: Genetic Data Modeling for the Childhood Cancer Clinical Data Commons (C3DC)

Author(s) - repeat as many times as needed per author:

First Name: Michael

Last Name: Watkins

Degree(s): PhD

Organization: University of Chicago

First Name: Brian

Last Name: Furner

Organization: University of Chicago

First Name: Samuel

Last Name: Volchenboum

Degree(s): MD, PhD

Organization: University of Chicago

First Name: Patrick

Last Name: Dunn

Degree(s): PhD

Organization: Frederick National Laboratory for Cancer Research

First Name: Sean

Last Name: Burke

Degree(s): PhD

Organization: Frederick National Laboratory for Cancer Research

First Name: Maureen

Last Name: Ryan

Organization: National Cancer Institute

First Name: Janice

Last Name: Knable

Organization: National Cancer Institute

First Name: Austin

Last Name: Fitts

Degree(s): PharmD

Organization: National Cancer Institute

First Name: Subhashini

Last Name: Jagu

Degree(s): PhD

Organization: National Cancer Institute

Presenting author: Brian Furner

Institution: University of Chicago

Email address: bfurner@bsd.uchicago.edu

Abstract (250 word max):

The Childhood Cancer Clinical Data Commons (C3DC) is an important addition to the CCDI Data Ecosystem. With its 6th release in June 2025, its available data currently includes 18,594 participants from the Molecular Characterization Initiative (MCI) and the TARGET Initiative. The scope of the data has initially been limited to disease site and diagnosis, treatment types and agents, treatment response, and survival status. However, recent modeling has been undertaken in order to augment this data with genetic findings. The C3DC data model is adapted from the Data for the Common Good (D4CG) Pediatric Cancer Data Commons (PCDC) data model, which was built through iterative consensus by dozens of international oncology subject matter experts. The source format of clinical genomic data varies greatly depending on the specific test/panel and the proprietary structure/format of the test reports that each laboratory uses. This model is able to accommodate genetic findings at three general levels of granularity. The least granular is simply a test name and an unstructured text blob of results. The middle level (which we have seen to be the most common) also includes a test name and unstructured results, but adds additional fields for the standardized representation of the specific alterations in either ISCN (chromosomal) or HGVS (genic) nomenclatures. The most granular is to represent the elements of the unstructured text blob as discrete fields in addition to the ISCN or HGVS strings. These fields cover a wide breadth of information, from copy number status to allele frequency.