CCDI Data Ecosystem and AI Readiness

Introduction

The Childhood Cancer Data Initiative (CCDI) Data Ecosystem is a connected network of tools and resources that integrates multiple data platforms—the Childhood Cancer Clinical Data Commons (C3DC), CCDI cBioPortal Cancer Data Explorer, CCDI Data Federation Resource, National Childhood Cancer Registry (NCCR) and the CCDI Hub—to support discovery, access, and reuse of pediatric cancer data. The CCDI Childhood Cancer Data Catalog supports AI in pediatric cancer research by helping researchers discover high-quality clinical, genomic, and imaging data sets needed to train and validate machine-learning models. The CCDI Participant Index (CPI) maps data from the same participant across data sets and resources.

CCDI Data Ecosystem resources present the ability to view and download:

- raw sequencing, clinical, imaging, and assay files for download in the CCDI Hub.
- cleaned and standardized metadata mapped to Common Data Elements in the CCDI-DCC data model and CCDI Data Federation API.
- curated and annotated analytics-ready data for clinical and genomic alteration data exposed via the C3DC, NCCR, or CCDI cBioPortal.

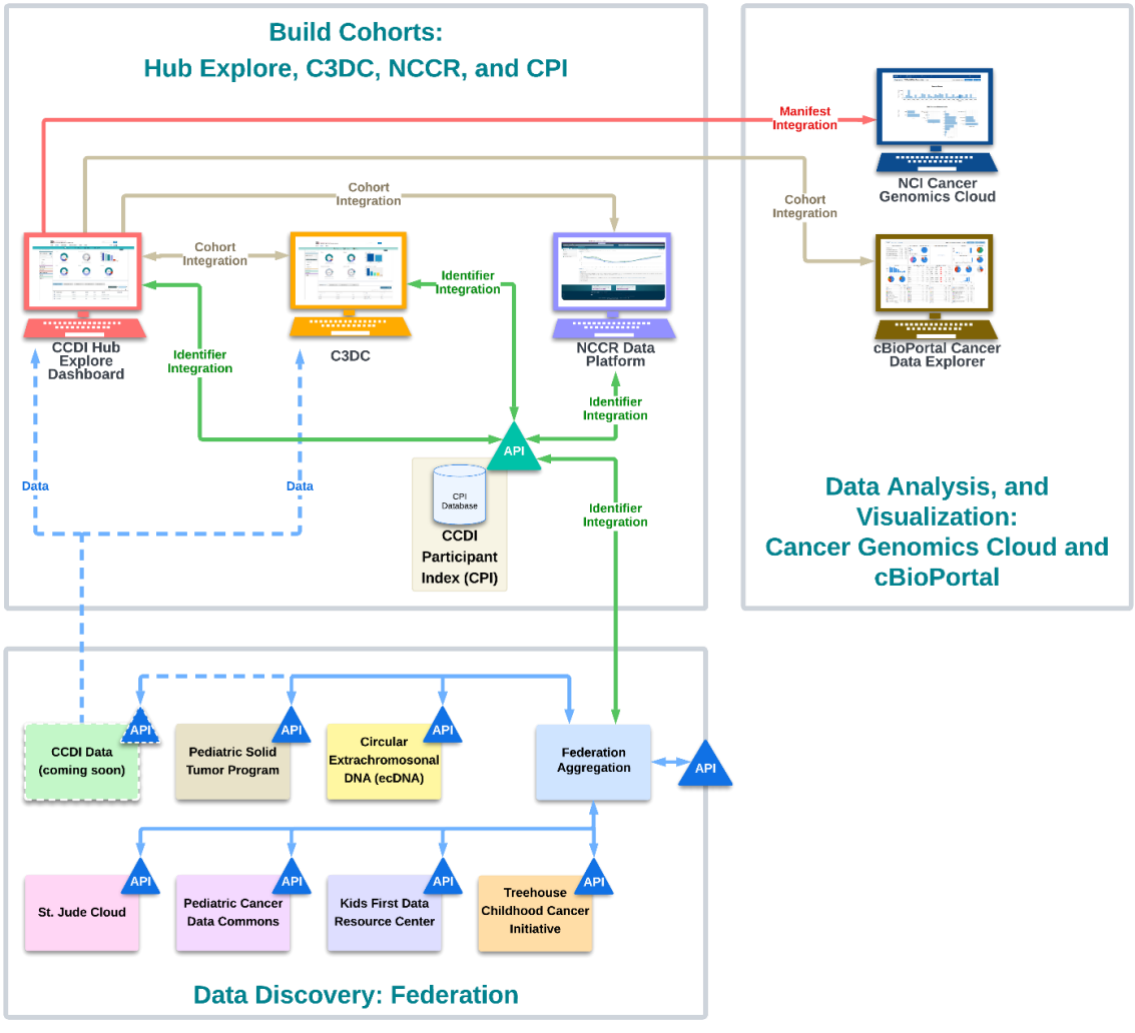

Figure 1: CCDI Data Ecosystem architecture. *The CCDI Data Ecosystem connects cohort-facing platforms (CCDI Hub Explore Dashboard, C3DC, and the NCCR Data Platform) with the CPI, which maps research identifiers and links records across data sets and resources. The ecosystem also includes a federation aggregation layer that connects external resources via APIs (e.g., St. Jude Cloud, Kids First, Pediatric Cancer Data Commons, Treehouse, ecDNA, and others). Analysis and visualization are supported through integration with the NCI Cancer Genomics Cloud and the CCDI cBioPortal Cancer Data Explorer.*

AI in Pediatric Cancer Research

Artificial intelligence (AI) transforms pediatric cancer research through the integration and analysis of multimodal data (e.g. genomic profiles, digital pathology images, clinical information). These advanced technologies play a critical role in identifying disease-specific molecular and morphological patterns, improving tumor classification, predicting disease progression and treatment response, and supporting the development of precision therapies for children with cancer.

For information on the use of AI across NCI, please refer to the AI and Cancer and AI in Cancer Research pages.

Applications of AI in CCDI

CCDI has made substantial progress in generating, harmonizing, and sharing high-value pediatric cancer data. CCDI's data sets, including prospectively collected Molecular Characterization Initiative (MCI) data, are well-curated pediatric cancer resources and offer a unique opportunity to advance AI research and development and are well suited for AI validation and benchmarking. Some CCDI-supported, AI-relevant efforts are highlighted on the CCDI-Funded Projects page.

The examples below highlight how the research community is using these data to develop and apply AI models tailored to childhood cancers.

AI Analysis or Application	**How CCDI Data Could Support It**
Predictive prognosis models	Use harmonized, deidentified pediatric clinical and outcome data from the C3DC to train AI models that predict survival, relapse timing, or treatment response in children with cancer.
Multimodal diagnosis classifiers	Combine clinical data from C3DC with pediatric tumor imaging from the Imaging Data Commons (IDC) to train deep learning models that improve diagnosis accuracy of cancer subtypes.
Radiomics and image feature extraction	Use IDC pediatric cancer imaging collections to develop AI tools that extract quantitative radiomic features (e.g., texture, shape) to correlate with clinical outcomes or molecular profiles.
Automated tumor segmentation	Train models on annotated MRI or pathology images within the IDC to automatically segment tumors for treatment planning.

Rare subtype clustering	Apply machine learning to integrated clinical, genomic, and imaging data sets in CCDI resources to identify novel rare pediatric cancer subgroups.
Cohort selection automation	Use metadata and filters from the C3DC to automate creating custom cohorts for specific AI experiments (age group, diagnosis, data types, etc.).
Multimodal survival risk scoring	Combine structured clinical variables with radiologic and histologic image features to train AI models that compute composite risk scores for individual patients.
Genotype–phenotype pattern discovery	Link harmonized clinical data from C3DC with molecular/genomic information from CCDI study data files to train models that uncover associations between genetic variants and observed cancer traits.
Imaging anomaly detection	Leverage large collections of pediatric radiology and pathology images to train unsupervised anomaly detection systems that flag unusual imaging patterns for expert review.
Treatment-response simulation models	Train reinforcement learning or survival modeling frameworks on longitudinal clinical data (treatment regimens + outcomes) in C3DC to simulate best-likely intervention strategies.
Integrate multi-omics for pediatric cancer discovery	Use models to train on and integrate high-dimensional multi-omics (genomic, transcriptomics, proteomics) data to identify patterns, predict functional neoantigens, prioritize dysregulated pathways, and uncover clinically actionable targets across pediatric cancers.
Real-world outcome prediction	Train and validate AI models using NCCR, which includes cancer registry data with longitudinal treatment and outcome information and linked medical and pharmacy claims, combined with clinical and genomics data in the CCDI Data Ecosystem, to predict survival, relapse risk, late effects, and care utilization patterns.