CHILDHOOD CANCER DATA INITIATIVE (CCDI) PARTICIPANT INDEX (CPI) API

1.1.0 Release Notes

Release Notes: CCDI Participant Index (CPI) API V1.2

The Childhood Cancer Data Initiative (CCDI) is excited to announce the release of Version 1.1 of the CCDI Participant Index (CPI) API. For more information, visit https://participantindex-docs.ccdi.cancer.gov/.

This update introduces the following key enhancements and additions:

- K-anonymity implementation for the data in National Childhood Cancer Registry Platform to enhance data privacy and security
- An additional 11 domains (TARGET Studies) and updated data for COG USI, totaling 11,223 additional participants.

K-anonymity Value

K-anonymity is a data anonymization technique that safeguards privacy by ensuring that data doesn't uniquely identify individuals. It uses a threshold to display data only when the number of unique participants with mapped IDs in the query result set **meets or exceeds** the specified value. The result set must include at least the k number of mappings (e.g., IDs) for a domain to be displayed.

V.1.1 Data

The CPI currently includes a total 41 domains and 429,657 participant IDs. The following table details the domain names, domain descriptions, domain types, and participant counts currently included in the CPI:

Note: New and updated data indicated with an asterisk (*)

Domain Name	Domain Description	Domain Type	Count
GENIE	AACR Project GENIE	Institutional Identifier	176
GMKF	Kids First IDs	Institutional Identifier	5,301
PCDC	Pediatric Cancer Data Commons	Institutional Identifier	22,328
StJude_CompBioID	St. Jude Comp Bio IDs	Institutional Identifier	1,679
Treehouse	Treehouse Childhood Cancer Initiative	Institutional Identifier	885
USI*	COG USI	Institutional Identifier	365,975
phs000463*	TARGET: Acute Lymphoblastic Leukemia (ALL) Pilot Phase 1	Dataset	24
phs000464*	TARGET: Acute Lymphoblastic Leukemia (ALL) Expansion Phase 2	Dataset	1,731
phs000465*	TARGET: Acute Myeloid Leukemia (AML)	Dataset	2,492
phs000466*	TARGET: Kidney, Clear Cell Sarcoma of the Kidney (CCSK)	Dataset	13
phs000467*	TARGET: Neuroblastoma (NBL)	Dataset	1,132
phs000468*	TARGET: Osteosarcoma (OS)	Dataset	383
phs000469*	TARGET: Cancer Model Systems (MDLS): Cell Lines and Xenografts (including PPTP)	Dataset	132

Domain Name	Domain Description	Domain Type	Count
phs000470*	TARGET: Kidney, Rhabdoid Tumor (RT)	Dataset	69
phs000471*	TARGET: Kidney, Wilms Tumor (WT)	Dataset	652
phs000720	dbGaP accession: Genomic sequencing of Pediatric Rhabdomyosarcoma	Dataset	403
phs001228	dbGaP accession: Gabriella Miller Kids First (GMKF) Pediatric Research Program in Susceptibility to Ewing Sarcoma Based on Germline Risk and Familial	Dataset	276
phs001436	dbGaP accession: Discovering the Genetic Basis of Human Neuroblastoma: A Gabriella Miller Kids First Pediatric Research Program (Kids First) Project	Dataset	614
phs001437	dbGaP accession: Pediatric Preclinical Testing Consortium (PPTC)	Dataset	267
phs001714	dbGaP accession: Gabriella Miller Kids First Pediatric Research Program: An Integrated Clinical and Genomic Analysis of Treatment Failure in Pediatric	Dataset	5
phs002187	dbGaP accession: Gabriella Miller Kids First Pediatric Research Program in Germline and Somatic Variants in Myeloid Malignancies in Children	Dataset	1,290
phs002276	dbGaP accession: Gabriella Miller Kids First Pediatric Research Program in Pediatric T-Cell Acute Lymphoblastic Leukemia	Dataset	1,358
phs002330	dbGaP accession: Kids First and INCLUDE: Down Syndrome and Heart Defects and Acute Lymphoblastic Leukemia	Dataset	517
phs002371	dbGaP accession: Human Tumor Atlas Network (HTAN)	Dataset	30
phs002430	dbGaP accession: Childhood Cancer Data Initiative (CCDI): Genomic Analysis in Pediatric Malignancies	Dataset	231
phs002431	dbGaP accession: CCDI U Michigan Study	Dataset	611
phs002504	dbGaP accession: UCSF Database for the Advancement of JMML - Integration of Metadata with "Omic" Data	Dataset	188
phs002517	dbGaP accession: Childhood Cancer Data Initiative (CCDI): Molecular Characterization across Pediatric Brain Tumors and Other Solid and Hematologic Mal	Dataset	1,229
phs002518	dbGaP accession: Childhood Cancer Data Initiative (CCDI): OncoKids - NGS Panel for Pediatric Malignancies	Dataset	1,039
phs002529	dbGaP accession: Childhood Cancer Data Initiative (CCDI): Comprehensive Genomic	Dataset	200

Domain Name	Domain Description	Domain Type	Count
	Sequencing of Pediatric Cancer Cases (CMRI/KUCC)		
phs002599	dbGaP accession: NCI CCSG CCDI Supplement Additional Genomic Submission	Dataset	104
phs002620	dbGaP accession: Feasibility and Clinical Utility of Whole Genome Profiling in Pediatric and Young Adult Cancers	Dataset	114
phs002790	dbGaP accession: Childhood Cancer Data Initiative (CCDI): Molecular Characterization Initiative	Dataset	2,807
phs003111	dbGaP accession: Childhood Cancer Data Initiative (CCDI): Clonal Evolution During Metastatic Spread in High-Risk Neuroblastoma	Dataset	129
SD_AQ9KVN5P	Kids First: T Cell ALL (KF-TALL)	Study	1,358
SD_DYPMEHHF	Kids First: Neuroblastoma (KF-NBL)	Study	614
SD_PET7Q6F2	Kids First: Myeloid Malignancies (KF-MMC)	Study	1,290
SD_YGVA0E1C	Kids First: Ewing Sarcoma - Genetic Risk (KF-ESGR)	Study	276
SD_Z6MWD3H0	Kids First: Leukemia & Heart Defects in Down Syndrome (KF-CHDALL)	Study	517
SD_ZXJFFMEF	Kids First: Osteosarcoma (KF-OS)	Study	5
TARGET*	Therapeutically Applicable Research to Generate Effective Treatments	Study	6,618
		Total	**429,657**

Access and Support

The CPI API will be made available to applications and services authorized by NCI. API Documentation is available at https://participantindex-docs.ccdi.cancer.gov/. Integration of the CPI service within the Hub is currently under design, with implementation planned for early next year.

Other interested system owners may initiate a request for access by emailing NCIChildhoodCancerDataInitiative@mail.nih.gov.

Appendix – Definitions and API Methods

Identifiers and Domains

Participant ID: Public identifier that appears in a research dataset accessible to researchers (e.g., Kids First ID or AACR-GENIE ID). This also includes global identifiers used more broadly (e.g., COG USI).

Domain: Groups of unique participant IDs, Domains can be categorized as:

- **Organizational Identifier**. Typically refers to a larger project/network namespace of unique participant identifiers. Current institutional identifiers include GENIE (AACR Project GENIE), GMKF (Gabriella Miller Kids First), PCDC (Pediatric Cancer Data Commons), StJude_CompBIOID (St Jude Comp Bio IDs), Treehouse (Treehouse Childhood Cancer Initiative), and USI (COG USI).
- **Dataset.** Refers to a single dataset for a particular study where data files for the participants can be found. This is typically a dbGap phs accession number – for example phs000720 (dbGaP accession: Genomic sequencing of Pediatric Rhabdomyosarcoma).
- **Study.** Represents programmatic references to research initiatives, such as TARGET, or an institution's internal study synonyms for dataset domains such as SD_AQ9KVN5P, equivalent to phs002276.

API Methods

The API provides mappings of IDs and domains to public identifiers and exposes the following methods:

Method	Description
Get status for Participant	This method validates the participant IDs for the given domains. It checks if the provided participant IDs are valid within their respective domains and returns the validation status for each.
Get Associated Participant IDs	This method retrieves a list of associated participant IDs based on input participant IDs and their respective domains. For each input participant, the method returns associated participant IDs across various domains, helping to link and track participants across different datasets and studies.
Get Domain Details	This method retrieves detailed information about various domains within a dataset. Each domain is described by its name, a brief description, its status, category, and a URL pointing to more detailed data. This information can be used to understand the purpose and current status of each domain, along with accessing additional resources.
Get Associated Domains For Participant	This method retrieves a list of domains associated with a given participant ID. Each associated domain includes details such as the domain name, description, status, and category. This helps in understanding the various domains in which a participant's data is present and provides additional context for each domain.
Get Participant Counts By Domain(statistics)	This method retrieves the counts of participants mapped to various domains within a dataset. It provides a breakdown of the number of participants associated with each domain name, along with the total counts of mapped and unique participants.