Developing Pediatric Cancer Data Standards

CCDI Webinar Series

NIH NATIONAL CANCER INSTITUTE

February 26, 2024

Introductions

Dr. Gregory Reaman

Agenda

1. *Pediatric Data Commons*
2. *PCDC Consensus Data Modeling*
3. *D4CG, NCI Semantic Infrastructure, and CCDI Collaboration*
4. *Pediatric Cancer Core Elements*
5. *Childhood Cancer Clinical Data Commons (C3DC)*
6. *Q&A*

Today's Speakers

Dr. Sam Volchenboum
Principal Investigator and Pediatric Oncologist
Data for the Common Good

Dr. Michael Watkins
Manager of Data Standards and Modeling
Data for the Common Good

Brian Furner
Senior Director of Data and Technology
Data for the Common Good

THE UNIVERSITY OF CHICAGO | DATA FOR THE COMMON GOOD

Pediatric Cancer Data Commons (PCDC)

Dr. Sam Volchenboum

CCDI Harmonization Objectives

- Establish consistent and interoperable standards for collecting, storing, and sharing pediatric cancer data.
- Facilitate data integration across studies for streamlined analysis.
- Develop core pediatric elements for clinical and research purposes.
  - Curate key data elements for retrospective harmonization and prospective collection.

Image Source: Shutterstock

D4CG: Who We Are

DATA FOR THE COMMON GOOD

PEDIATRIC CANCER DATA COMMONS

GEARBOx

Sociome

PREDICT
monogenic diabetes

other rare disease commons

…and more to come

Service Providers

PEDIATRIC CANCER DATA COMMONS

Advisory Groups

SAC | EAB | ARC

D4CG Team

D4CG

Consortia

LT, AML, ALL, FA, Bone, CNS, GCT, HL, NBL, NPC, Pre, FPRH, RB, STS

PCDC Data Portal

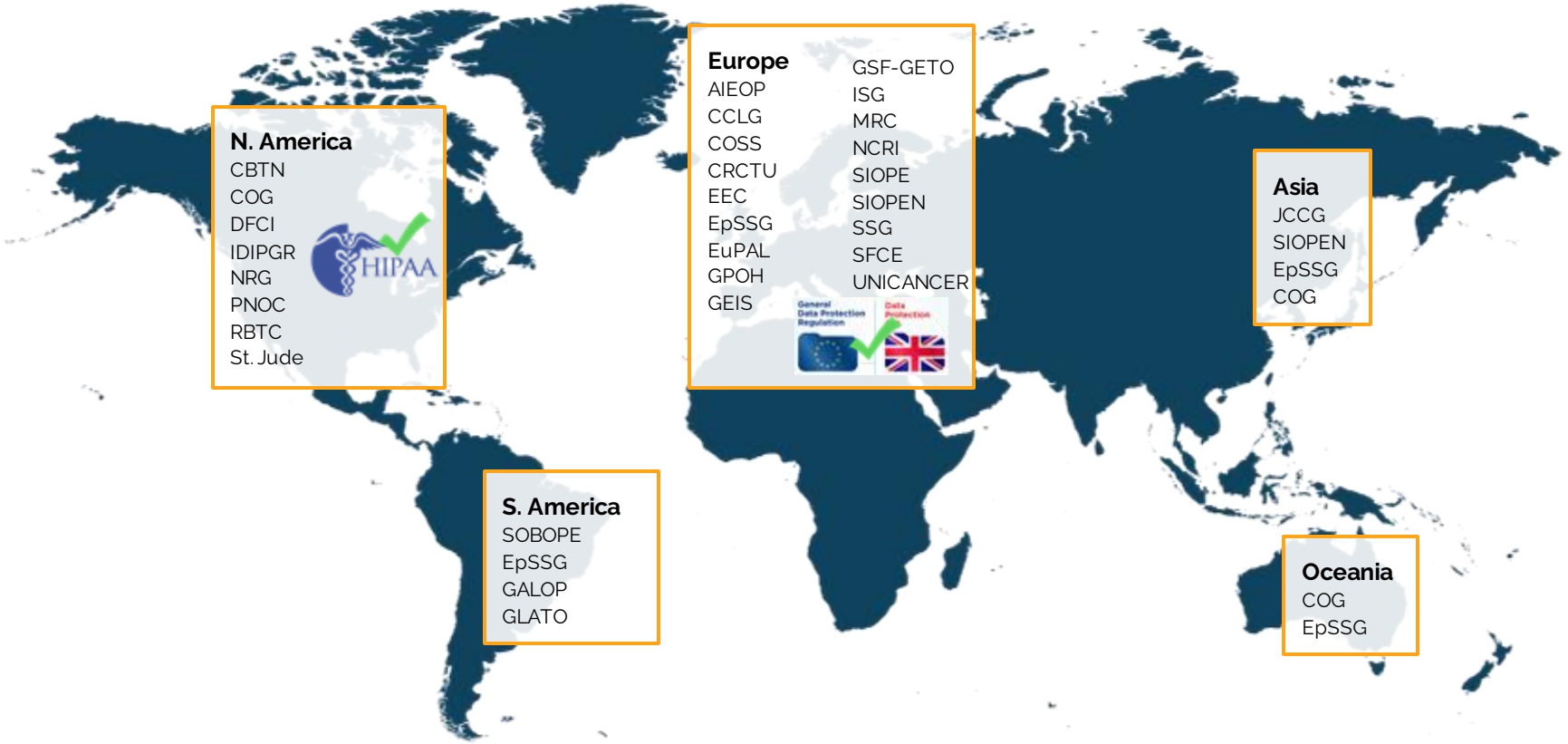
Imaging data | Biobank data | Genomic data

Governance, Data Modeling, Technical Infrastructure

PCDC Consensus Data Modeling

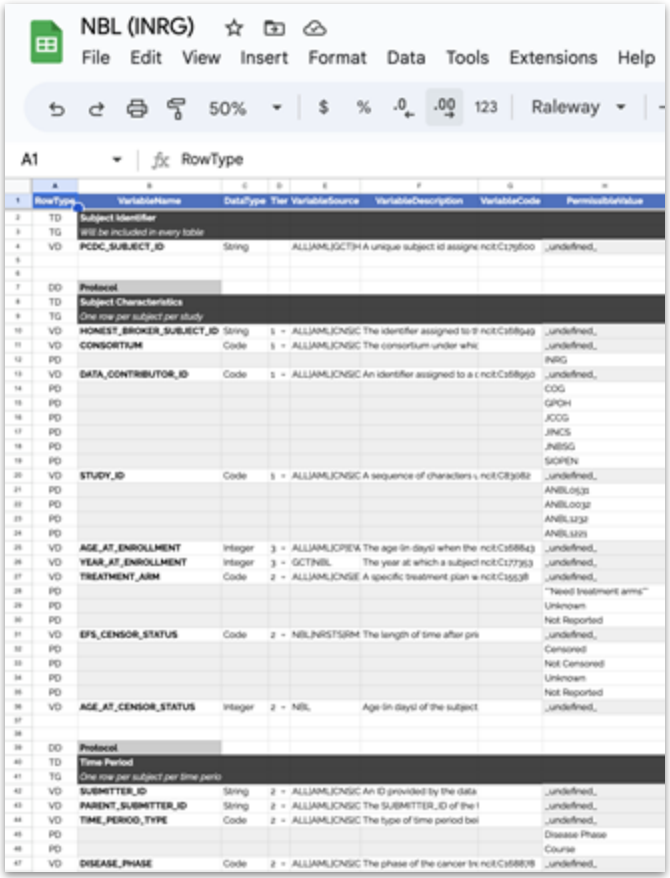
Dr. Michael Watkins

An International Collaboration



Data Dictionary Meetings

- Prioritize the re-use of existing data elements to enable cross-disease research.



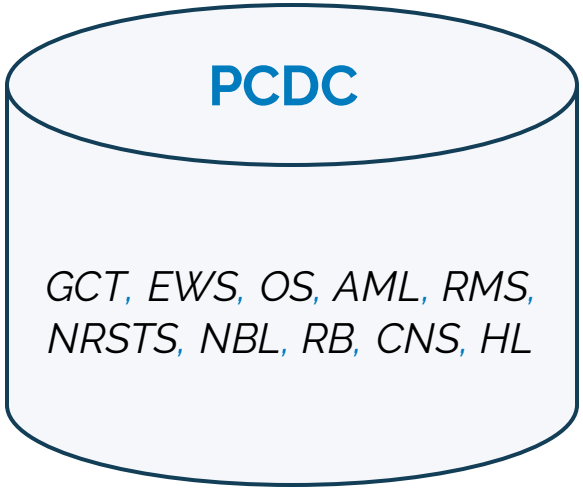
Cross-Domain Stakeholders

We address the challenge of complex rare diseases by convening focused communities of diverse experts:

- Pediatric oncologists
- Surgeons
- Pathologists
- Radiation oncologists
- Genetic counselors
- Statisticians
- Database engineers
- Terminologists
- Regulatory and governance experts

A Commons of Commons

https://commons.cri.uchicago.edu/pcdc/



COG
UCL
EURO-EWING
CRCTU
EORTC
FSG/SFCE
GPOH
ISG
SSG
UK

Ewing Sarcoma

AIEOP
COG
MRC
CCLG
NRG-Oncology
SFCE
SOPOBE
DFCI

Germ Cell Tumors

AIEOP
BFM-SG
COG
DCOG
MRC
NOPHO
PPLLSG
SJCRH
JACLS
JPLSG

AML

BOCG
COG
COSS-GPOH
CRCTU
FSG
GALOP
GEIS
GPOH
ISG
MRC
SSG
UCL
UK

Osteosarcoma

Neuroblastoma

COG
GPOH
JCCG
JINCS
JNBSG
SIOPEN

Soft-Tissue Sarcoma

AIEOP
COG
EpSSG
SIOP MMT
CWS
STSC

Retinoblastoma

CHLA
COG
DFCI
DEPICT
EuRBG
GALOP

Embryonal CNS

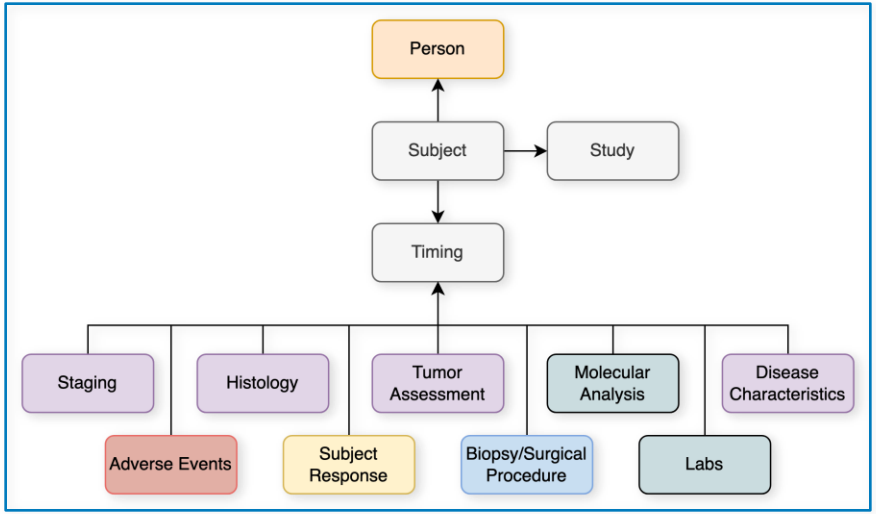
CBTN
COG
IDIPGR
PNOC
RBTR
SIOPE

Hodgkin Lymphoma

COG
SJCRH
PHC

Changes to the PCDC model

- Adjustments across 16+ groups are frequent.
- Versioning and change management are critically important but complex.
- High volume of changes over the years as new groups have joined.
- The rate of change requests is now slowing—an indication of maturity in the consensus elements.



Subset of the PCDC Model

D4CG, NCI Semantic Infrastructure, and CCDI Collaboration

Dr. Michael Watkins

NCI's Enterprise Vocabulary Services: Promote Semantic Interoperability

- The Cancer Data Standards Registry and Repository (caDSR): A structured repository for clinical and research metadata, which semantically defines data through controlled terms and vocabularies.
- Common Data Elements (CDEs) are standardized metadata elements used to describe specific pieces of information in clinical and research settings.
- NCI Thesaurus (NCIt): A controlled vocabulary of biomedical concepts including terms related to diseases, anatomy, procedures, and more.
- CDEs incorporate terms or concepts from NCIt, which ensures consistency and interoperability across different data collection (clinical and research) efforts.
- Semantic interoperability enables easy sharing and understanding of data across systems and research domains.

PCDC Disease-Specific Dictionaries Alignment

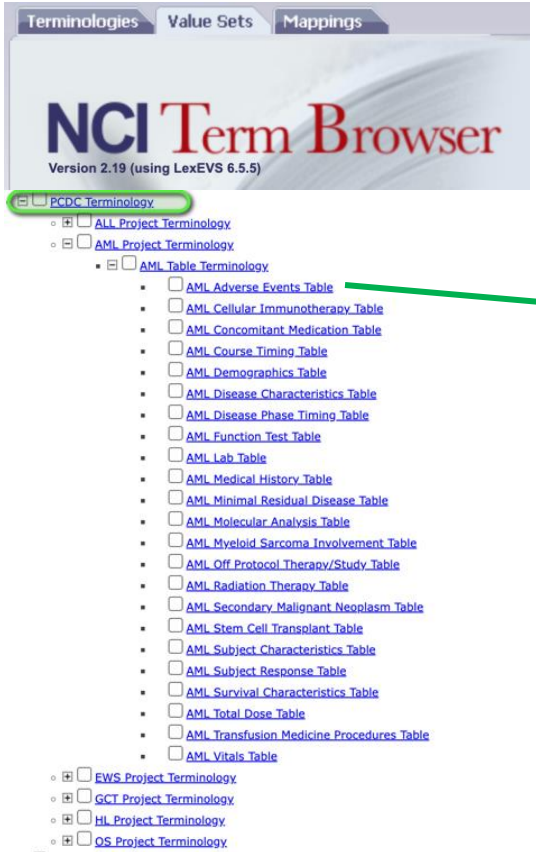
Domain	Tables	ALL	AML	CNS	CP	EWS	FA	FPRH	GCT	HL	LT	NBL	NPC	NRSTS	OS	RB	RMS
Protocol	Subject Characteristics	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Protocol	Time Period	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Protocol	Off Protocol Therapy Or Study	☑	☑	☑	☐	☑	☑	☐	☐	☑	☑	☐	☑	☐	☑	☑	☐
Demographics	Demographics	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Demographics	Medical History	☑	☑	☐	☑	☑	☑	☑	☑	☑	☑	☑	☐	☑	☑	☑	☑
Demographics	Survival Characteristics	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Demographics	Family Medical History	☐	☐	☐	☐	☐	☑	☐	☐	☑	☐	☐	☑	☐	☐	☑	☐
Testing	Vitals And Anthropometrics	☑	☑	☑	☑	☐	☑	☑	☐	☐	☐	☐	☑	☐	☐	☐	☐
Testing	Laboratory Test	☑	☑	☐	☑	☑	☑	☑	☑	☑	☑	☑	☑	☐	☑	☐	☑
Testing	Genetic Analysis	☑	☑	☑	☑	☑	☐	☑	☑	☐	☐	☑	☐	☑	☑	☑	☑
Testing	Function Test	☐	☑	☐	☐	☐	☐	☐	☐	☑	☐	☐	☑	☐	☑	☐	☐
Testing	Immunohistochemistry	☐	☐	☐	☐	☑	☐	☐	☑	☑	☐	☐	☐	☐	☐	☐	☐
Testing	Imaging	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐	☑	☐
Disease Attributes	Diagnosis	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Disease Attributes	Disease Characteristics	☑	☑	☐	☐	☑	☑	☑	☑	☑	☑	☑	☑	☐	☐	☑	☐
Disease Attributes	Disease Site Assessment	☐	☐	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Disease Attributes	Staging	☐	☐	☑	☑	☐	☑	☑	☐	☑	☑	☑	☑	☐	☐	☑	☑
Treatment	Radiation Therapy	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☐	☑	☑	☑	☑	☑
Treatment	Stem Cell Transplant	☑	☑	☐	☐	☑	☐	☐	☑	☑	☑	☐	☐	☐	☐	☑	☐
Treatment	Medication	☐	☑	☑	☐	☑	☑	☑	☑	☑	☑	☐	☑	☐	☑	☑	☐
Treatment	Transfusion Medicine Procedures	☐	☑	☐	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐	☐	☐
Treatment	Cellular Immunotherapy	☐	☑	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Treatment	Biopsy And Surgical Procedures	☐	☐	☑	☑	☑	☑	☑	☑	☑	☑	☐	☑	☑	☑	☑	☑
Treatment	Protocol Treatment Modifications	☐	☐	☐	☑	☐	☐	☐	☐	☑	☐	☐	☑	☐	☐	☐	☐
Response	Minimal Residual Disease	☑	☑	☐	☐	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐	☐
Response	Subject Response	☐	☑	☑	☐	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Events	Adverse Events	☑	☑	☐	☑	☑	☑	☑	☑	☑	☑	☐	☑	☐	☑	☑	☐
Events	Subsequent Malignant Neoplasm	☑	☑	☑	☐	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Events	Patient Reported Outcomes Metadata	☐	☐	☐	☐	☐	☐	☐	☑	☑	☐	☐	☐	☐	☑	☐	☐
Events	Late Effects	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐	☐	☐

Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Central Nervous System Tumors (CNS), Ewing Sarcoma (EWS), Fanconi Anemia (FA), Fertility Preservation and Reproductive Health (FPRH), Germ Cell Tumors (GCT), Hodgkin's Lymphoma (HL), Liver Tumors (LT), Neuroblastoma (NBL), Nasopharyngeal Carcinoma (NPC), Non-rhabdomyosarcoma Soft-Tissue Sarcomas (NRSTS), Osteosarcoma (OS), Retinoblastoma (RB), Rhabdomyosarcoma (RMS)

https://ncithesaurus.nci.nih.gov/ncitbrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&version=23.12d&ns=ncit&code=C168547&key=1832152029&b=1&n=null

https://evs.nci.nih.gov/ftp1/PCDC/

Disease-Specific Dictionaries Published to NCI Thesaurus



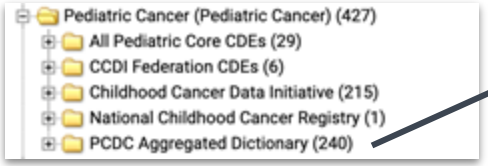
- AML Adverse Events Table

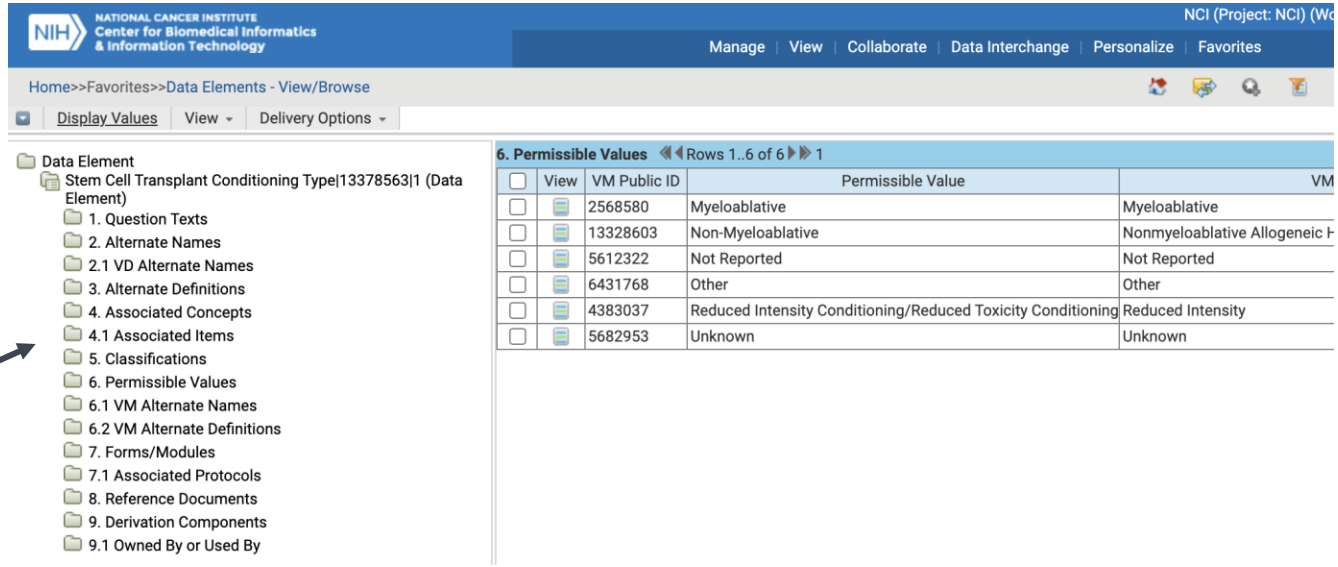
Export Excel Export CSV

Project	NCIt Code of Table	PCDC Table PT	NCIt Concept Code	NCIt PT	NCIt SY	PCDC PT	PCDC SY	NCIt Definition	PCDC
AML	C173217	Adverse Events Table	C175600	PCDC Subject Identifier	Pediatric Cancer Data Commons Unique Subject Identifier	PCDC_SUBJECT_ID		A unique identifier for a subject in a Pediatric Cancer Data Commons (PCDC) study.	
AML	C173217	Adverse Events Table	C172677	Age in Days at Onset Adverse Event		AGE_AT_AE		Age of subject (in days) at the onset of the adverse event.	
AML	C173217	Adverse Events Table	C168878	Disease Phase		DISEASE_PHASE		The stage or period of an individual's disease.	
AML	C173217	Adverse Events Table	C156813	Initial Diagnosis		Initial Diagnosis		The first diagnosis of the individual's condition.	

PCDC Data Elements Published in the caDSR

PCDC common data elements (CDEs) are published alongside other pediatric cancer CDEs in caDSR, fostering interoperability.



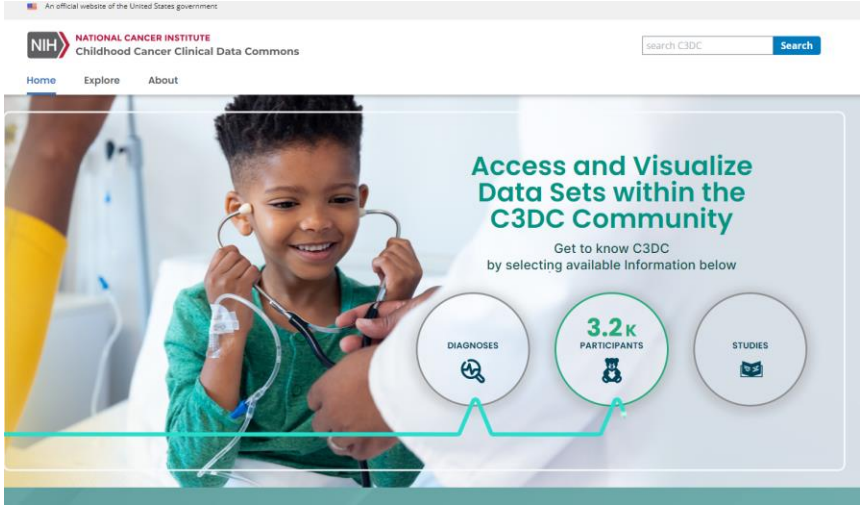


Pediatric Cancer Core Elements

Dr. Michael Watkins

Pediatric Cancer Core CDEs

- ~75 concepts (in active development)
- Includes CDEs from a number of high-level clinical domains:
  - Administrative, Demographics, Diagnosis, Treatment, Specimen, Molecular, and Imaging
- Importance will be indicated through tiering:
  - Tier 1–must include, regardless of the resource cost
  - Tier 2–include if resources are available
  - Tier 3–include if resources permit



Core Data Elements Overview–Molecular Features

CDE Public ID	CDE Long Name	Alt Name	PCDC Variable	Closest FHIR Path(s)	In mCODE?
13362328	Subject Gene Mutation Assessment Status	Status	STATUS	GenomicVariant.value	Y
11379445	Gene Human Chromosome Name	Chromosome	CYTOGENETIC_LOCATION	GenomicVariant.cytogeneticLocation	Y
11280318	Gene Occurrence HGNC Symbol Name	Gene Symbol	GENE	GenomicVariant.geneStudied	Y
13367968	Gene Mutation Second Gene HGNC Symbol Name	Second Gene Symbol	GENE2	GenomicVariant.geneStudied	Y
13367930	Gene Mutation Detection Type	The type of gain, loss or alteration	ALTERATION_TYPE	GenomicVariant.genomicDNAChangeType	Y
13367935	Gene Mutation Abnormality Type	Gene Mutation Abnormality Type	ALTERATION_EFFECT	GenomicVariant.molecularConsequence	Y
13367965	Gene Mutation Copy Number Variation Assessment Type	Copy Number Variation	COPY_NUMBER_STATUS	GenomicVariant.molecularConsequence	Y
13367961	Gene Mutation ISCN Karyotype Text	ISCN	ISCN	GenomicVariant.cytogeneticNomenclature	Y
13367959	Gene Mutation HGVS Protein Text	HGVS Protein	HGVS_PROTEIN	GenomicVariant.aminoAcidChange	Y
13367956	Gene Mutation HGVS Coding Text	HGVS Coding	HGVS_CODING	GenomicVariant.genomicDNAChange	Y
6142510	Molecular Analysis Genetic Zygosity Type	Zygosity	ALLELIC_STATE	GenomicVariant.allelicState	Y

- Closely aligns with HL7 Genomic Reporting Profiles used in mCODE.
- Excluded concepts: sample type, method, clinical significance, external IDs, allele frequencies.
- National Childhood Cancer Registry (NCCR) elements:
  - Core concepts present but are spread across many lab-specific fields.

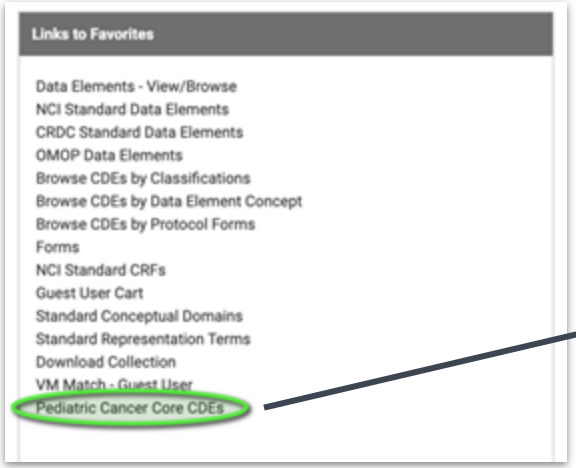
Core Data Elements Overview–Diagnosis

CDE Public ID	CDE Long Name	Alt Name	PCDC Variable	Closest FHIR Path(s)	In mCODE?
12318179	Subject Age at Histology Assessment Day Count	Age at Hist Assessment	AGE_AT_DIAG_ASSESSMENT	PrimaryCancerCondition.assertedDate	Y
14714145	Diagnosis Level of Certainty Status	diagnosis_certainty	-	PrimaryCancerCondition.verificationStatus	Y
13606067	Disease Morphology ICD-O-3 Label Text	Diagnosis Classification	DIAGNOSIS	PrimaryCancerCondition.code	Y
13606067	Disease Morphology ICD-O-3 Label Text	Diagnosis Classification	MORPH_CODE	PrimaryCancerCondition.histologyMorpholo	Y
-	-	Diagnosis Classification System	MORPH_CODE_SYSTEM	PrimaryCancerCondition.histologyMorpholo	Y
13279286	Diagnosis Primary Anatomic Site	Tumor Site	SITE	PrimaryCancerCondition.bodySite	Y
13381583	Biospecimen Histologic Grade Type	Histology Grade	HISTOLOGY_GRADE	-	N
12137353	Imaging Technology DICOM Modality Type	Imaging Modality	DETECTION_METHOD	TumorSize.method	Y
12922545	Tumor Classification Category	Tumor Classification	CLASSIFICATION	CancerDiseaseStatus.value	Y
13382770	Disease or Disorder Staging System Name	Stage System	STAGE_SYSTEM	PrimaryCancerCondition.stage.type	Y
-	-	Stage System Version	STAGE_SYSTEM_VERSION	PrimaryCancerCondition.stage.type.version	Y
13382767	Disease Tumor Stage Name	Stage	STAGE	PrimaryCancerCondition.summary	Y
3123069	Longest Tumor Surgeon Diameter Measurement	Tumor size longest diameter (as c	LONGEST_DIAM_DIM1	TumorSize.tumorLongestDimension	Y

- Most concepts present and mappable to mCODE.
- Diagnosis design complications:
  - Many new rare disease definitions not published as ICD-O morph codes.
  - Hierarchical disease definitions difficult without ontology bindings.
  - Biopsies can be diagnostic, prognostic, and therapeutic.

Pediatric Cancer Core CDEs: Draft

- These CDEs have a dedicated space within the caDSR.
- Developed in concert with curators from CCDI, NCI Semantic Infrastructure (SI) team, D4CG, and others.



Home>>Favorites>>Pediatric Cancer Core CDEs

Display Values | View | Delivery Options

Pediatric Cancer Standard Data Elements

View	CDE Public ID	Version	2 Pediatric Cancer Name	1 Domain	VD Type
	12960571	1.00	Study ID	Administrative	Enumerated
	12988910	1.00	Treatment Arm	Administrative	Enumerated
	2847330	1.00	Vital Status	Clinical Observation	Enumerated
	2192217	2.00	Ethnicity	Demographics	Enumerated
	2192199	1.00	Race	Demographics	Enumerated
	2690165	2.00	Relationship	Demographics	Enumerated
	11341616	1.00	Anatomic Site	Diagnosis	Enumerated
	14714145	1.00	diagnosis_ certainty	Diagnosis	Enumerated
	12217251	1.00	Disease Phase	Diagnosis	Enumerated
	13209646	1.00	Medical History Condition Category	Diagnosis	Enumerated
	13212141	1.00	Medical History Condition Type	Diagnosis	Enumerated
	13279286	1.00	Tumor Site	Diagnosis	Enumerated

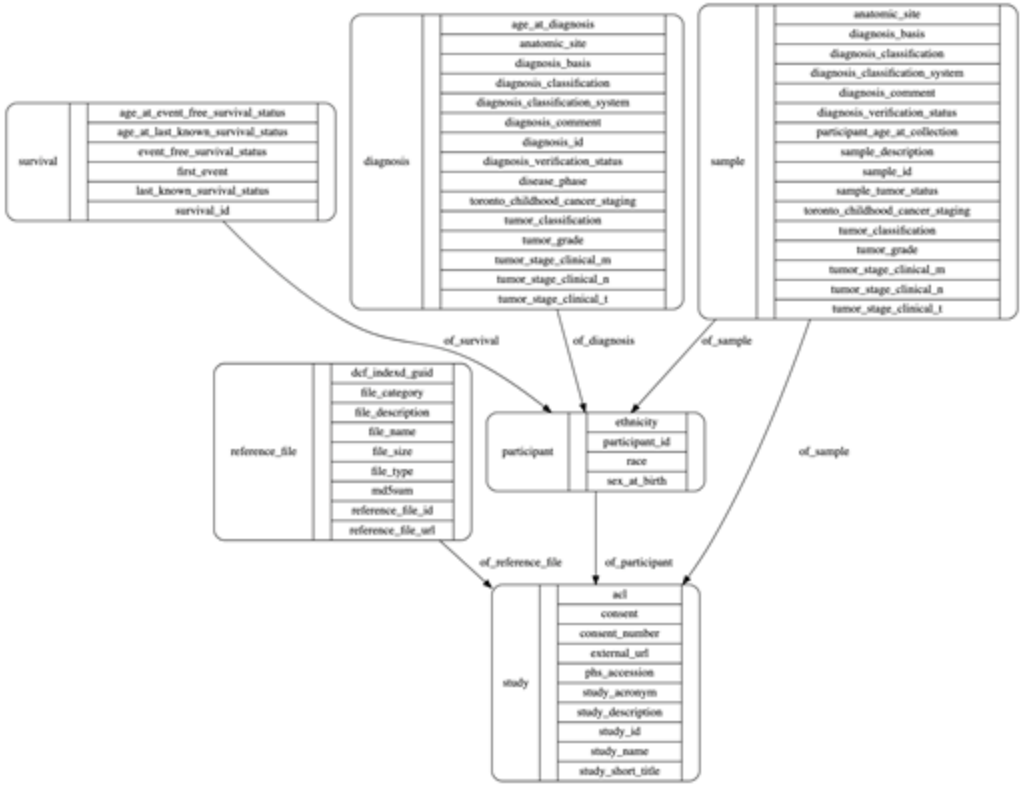
https://cadsr.cancer.gov/onedata/Home.jsp
https://cadsr.cancer.gov/onedata/dmdirect/NIH/NCI/CO/PCDC%20Browser

Childhood Cancer Clinical Data Commons (C3DC)

Brian Furner

C3DC Model

- C3DC includes coverage of study, participant, diagnosis, sample, and survival data.
- Model will grow as more studies are included.
- Reference file node contains metadata that describe lineage and provenance of data during mapping and extract, transform, and load (ETL) process.



https://github.com/CBIIT/c3dc-model

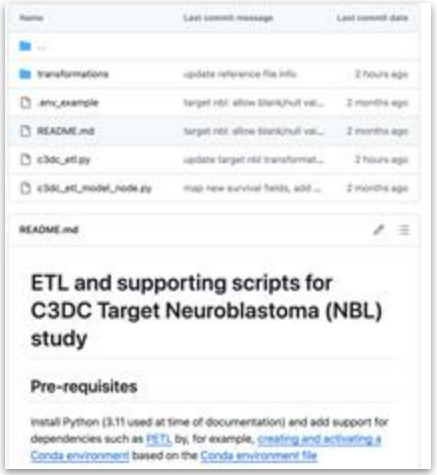
C3DC ETL

- Source data and harmonized data are kept separate to simplify transformations and downstream use of data.

- Mapping and transformation logic is documented and publicly available through GitHub, **ensuring transparent provenance and lineage.**
  - Source-to-destination mappings are encoded in JSON.
  - Transformation logic encapsulated in Python.

- To date, completed ETL and delivery of data for prioritized studies (TARGET NBL, MCI CNS, OncoKids Cancer Panel).

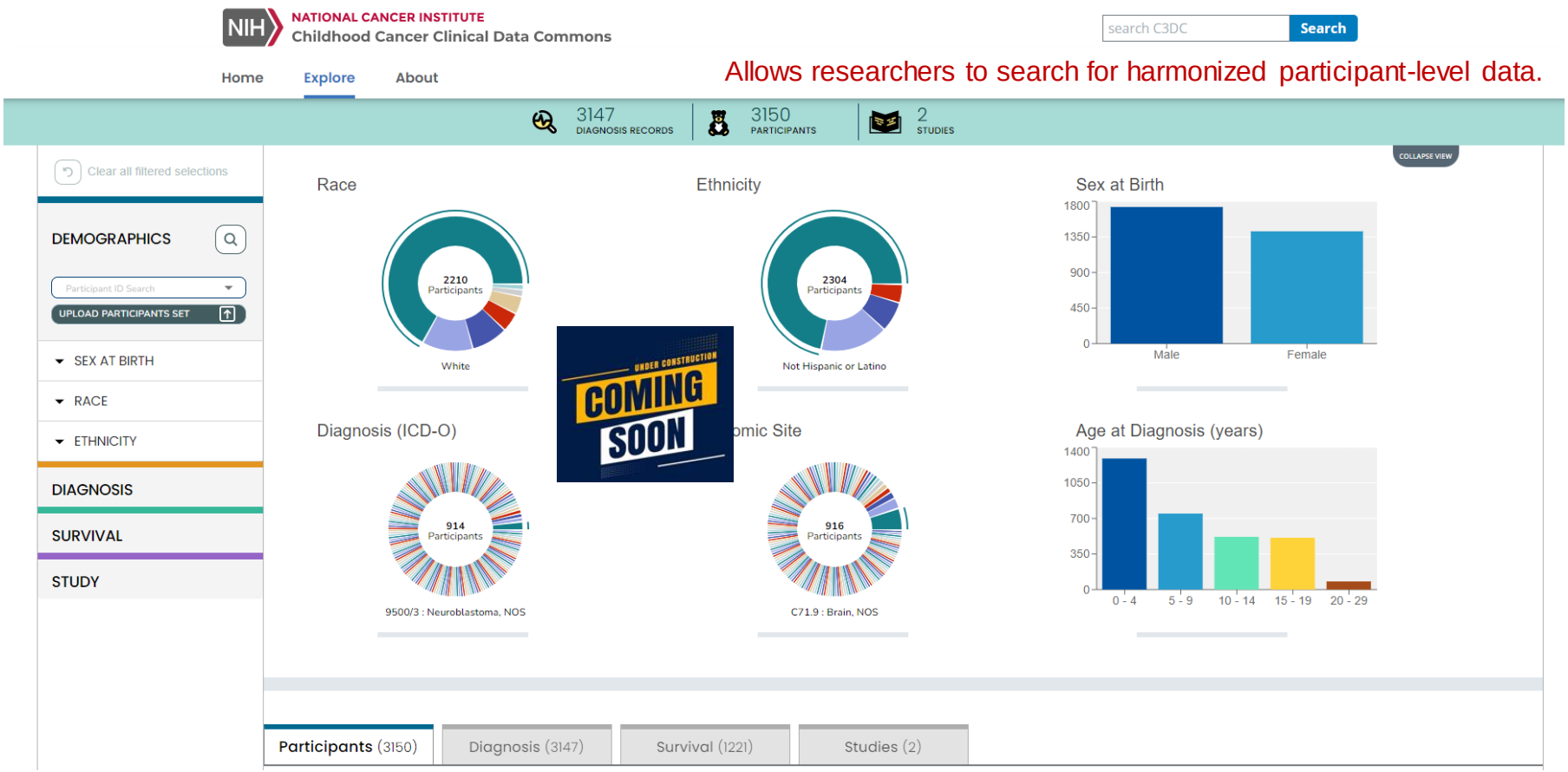
https://github.com/chicagopcdc/c3dc_etl



GitHub repository for ETL

Mapping example

Childhood Cancer Clinical Data Commons (C3DC)



Find Out More About CCDI

Learn about CCDI and subscribe to our monthly newsletter.
cancer.gov/CCDI

Questions? Email us.
NCIChildhoodCancerDataInitiative@mail.nih.gov

Q&A

CCDI March Community Forum

Monday, March 18, 12:00 – 1:00 pm ET

This session aims to gather community input on the Coordinated Pediatric and Young Adult Rare Cancer Initiative.

Topics include:

- Outcomes from February's Genomic Harmonization Task Force meeting
- Expansion of the Molecular Characterization Initiative
- Update on the Coordinated Pediatric and Young Adult Rare Cancer Initiative

Register Here: https://events.cancer.gov/ccdi/webinar/registration

Thank you!

cancer.gov

cancer.gov/espanol