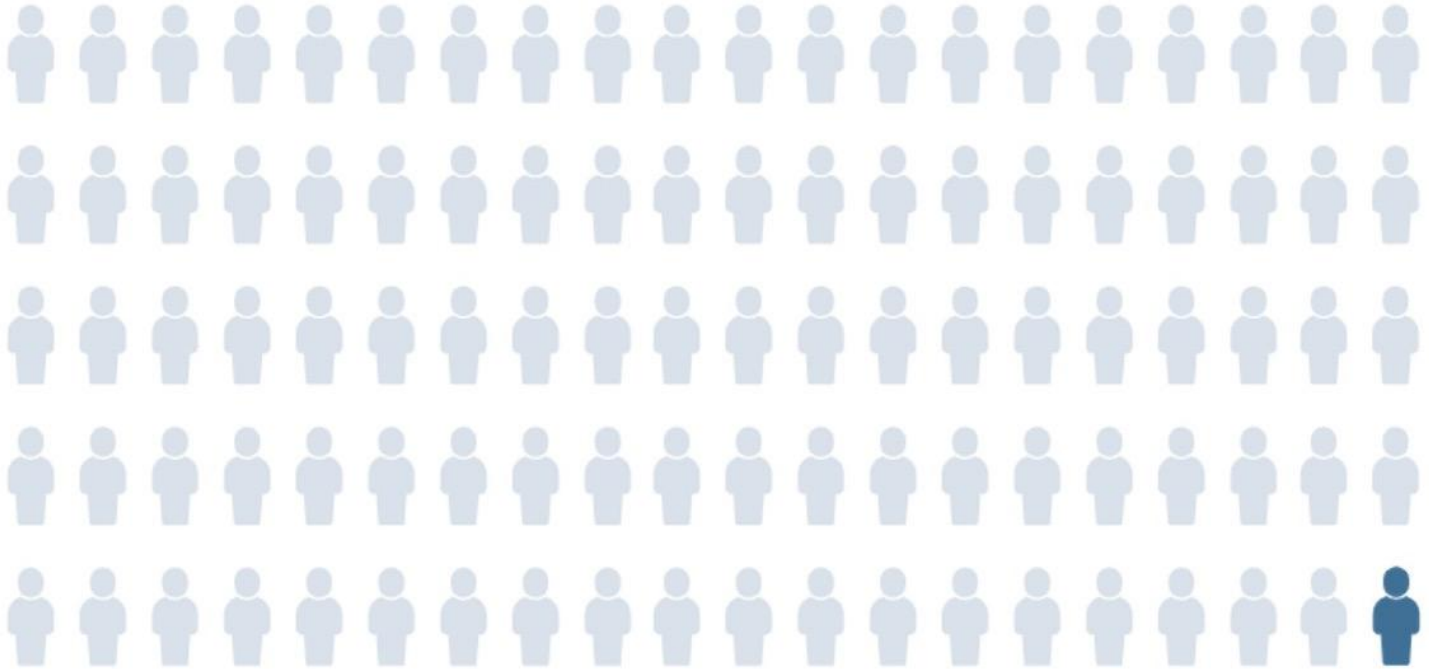
THE UNIVERSITY OF CHICAGO | PEDIATRIC CANCER DATA COMMONS

Transforming the Way Researchers Share Data

LESSONS FROM THE PEDIATRIC CANCER DATA COMMONS

http://sam.am/LHC2020

Pediatric cancer is a rare disease



18 million new cases of cancer worldwide every year

224,000 (**1%**) are children

Sources: Globocan 2018 (adult), Ward 2019 (pediatric)

Current state: Manual data entry

October — William

Sun	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*
	1 Bactrim Cyotxan 20mg	**2** Bactrim Cyotxan 20mg	**3** cycle 2 DTR 9:30am Avastin IV Cytoxan 20mg	**4** DTR 9:30am Cytoxan IV Zometa IV	**5** Cytoxan 20mg	**6** Cytoxan 20mg
7 Cytoxan 20mg	**8** Bactrim Cytoxan 20mg Labs and blood pressure at home	**9** Bactrim Cytoxan 20mg	**10** Cytoxan 20mg	**11** Cytoxan 20mg	**12** Cytoxan 20mg	**13** Cytoxan 20mg
14 Cytoxan 20mg	**15** Bactrim Cytoxan 20mg	**16** Bactrim Cytoxan 20mg	**17** DTR 9:30am Avastin IV Cytoxan 20mg	**18** Cytoxan 20mg	**19** Cytoxan 20mg	**20** Cytoxan 20mg
21 Cytoxan 20mg	**22** Bactrim Cytoxan 20mg SSKI	**23** Bactrim Cytoxan 20mg MIBG, x-ray knees Bone marrow biopsies SSKI	**24** Cytoxan 20mg MIBG CT scan SSKI	**25** Cytoxan 20mg SSKI	**26** Cytoxan 20mg SSKI	**27** Cytoxan 20mg
28 Cytoxan 20mg	**29** Bactrim Cytoxan 20mg	**30** Bactrim Cytoxan 20mg	**31** Cycle 3 DTR 9:30am Avastin IV Cytoxan 20mg	**1** DTR 9:30am Cytoxan IV Zometa IV		

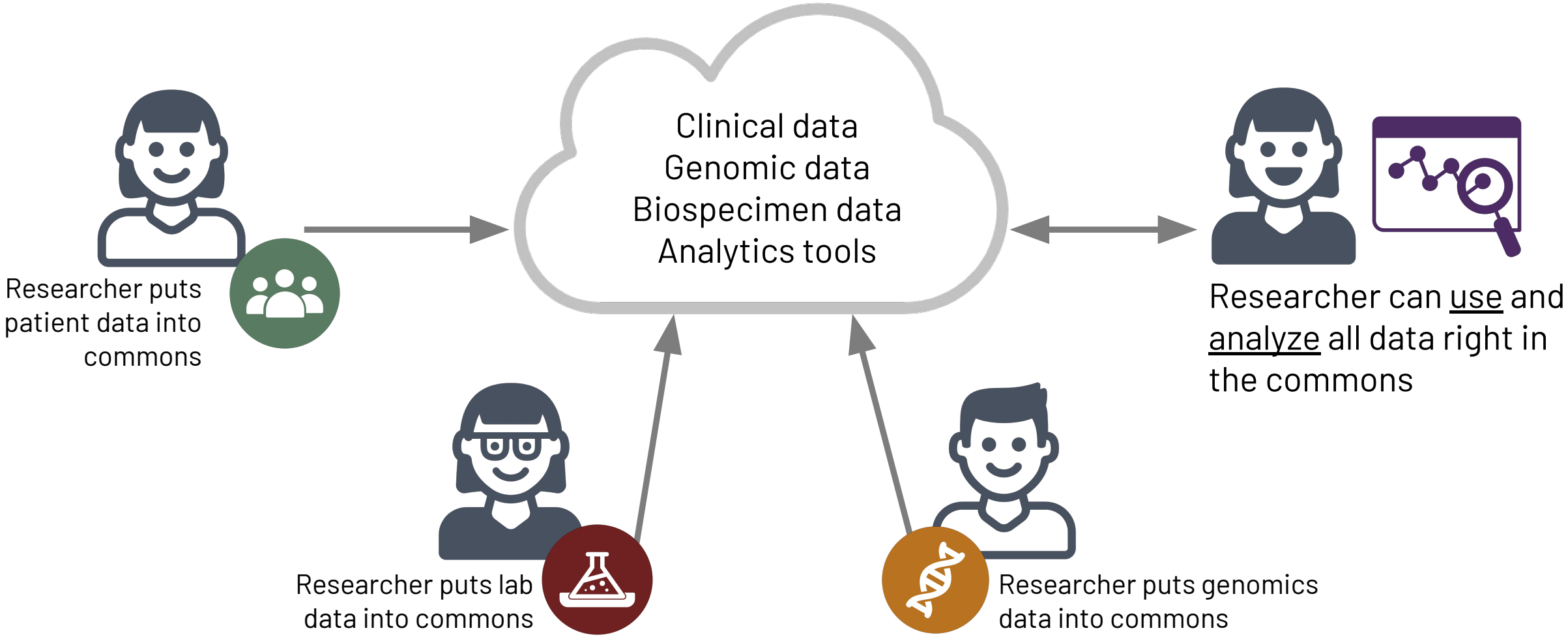
Current state: Lack of standardization

Ways that age is expressed in the Gene Expression Omnibus (GEO)

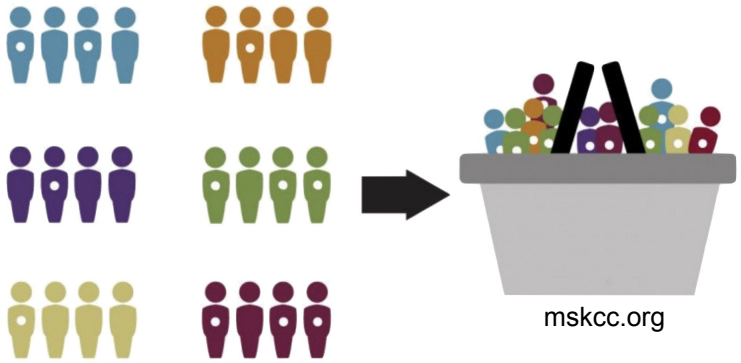
age	age [y]
Age	age [year]
AGE	age [years]
`Age	age in years
age (after birth)	age of patient
age (in years)	Age of patient
age (y)	age of subjects
age (year)	age(years)
age (years)	Age(years)
Age (years)	Age(yrs.)
Age (Years)	Age, year
age (yr)	age, years
age (yr-old)	age, yrs
age (yrs)	age.year
Age (yrs)	age_years

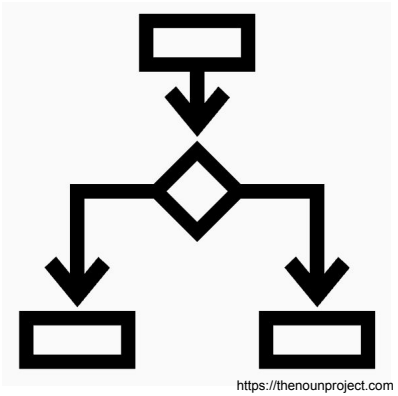
Adapted from Mark A. Musen, M.D., Ph.D.

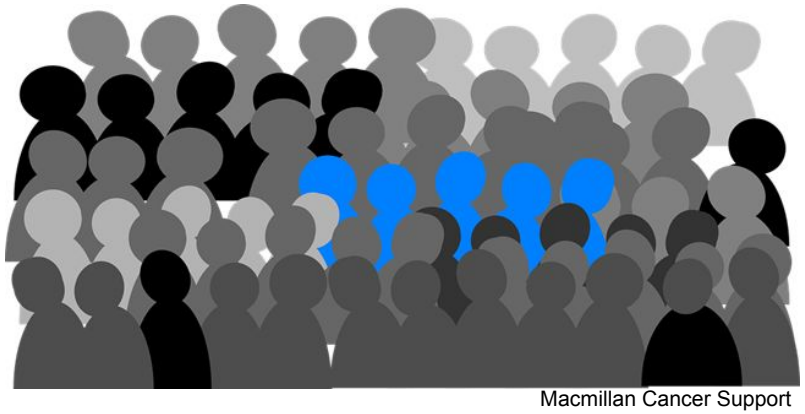
What is a data commons?



What kinds of research can a commons enable?







- Correlating biomarkers with clinical outcomes across trials
- Understanding impact of dose modifications across trials
- Performing patterns of failure analyses
- Examining toxicity prognosticators
- Validating consensus staging definitions across trials
- Validating prognostic scoring systems
- Enhancing risk stratification
- Prognosis of rare subgroups
- Age-related differences in therapeutic response
- Disparities analyses

Essential elements for building a commons

CONSORTIUM
building trust between groups

SCIENTIFIC GOALS
why build it? what data to include?

DATA GOVERNANCE
publication policy, approving data use

DATA DICTIONARY
everyone speaking the same language

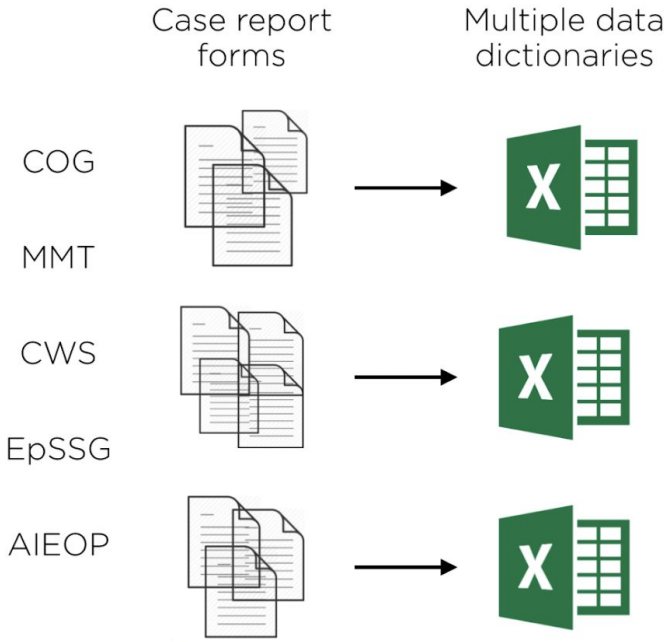
DATA TRANSFORMATION AND AGGREGATION
statisticians and data scientists

FUNDING
building and sustaining the commons

Consensus-based decision making and data sharing

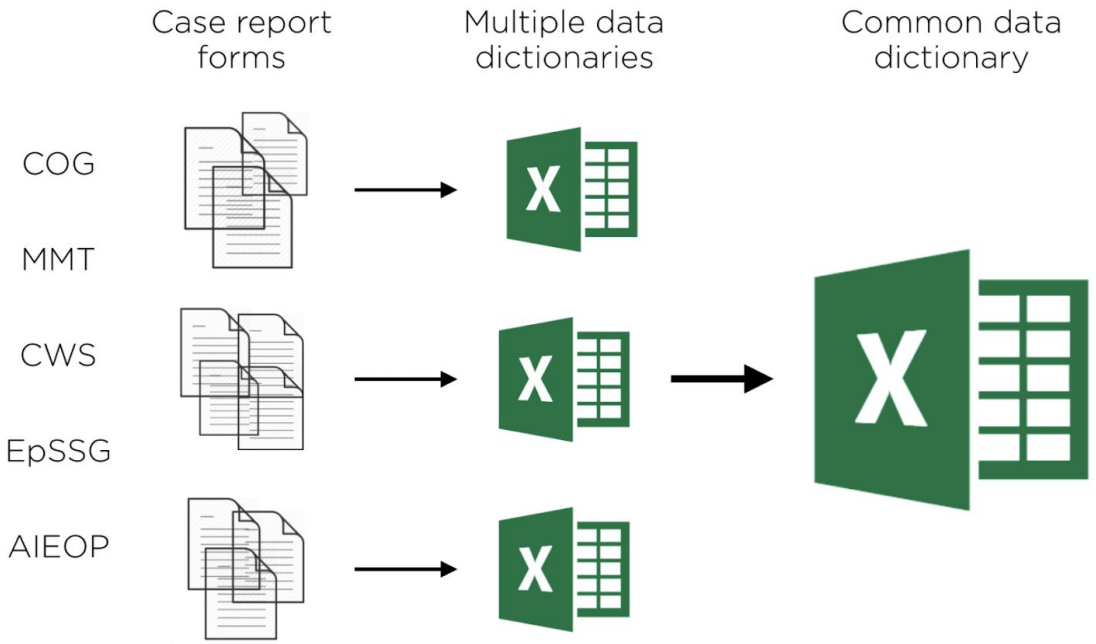
Balloting a consensus data dictionary

Disease consortia



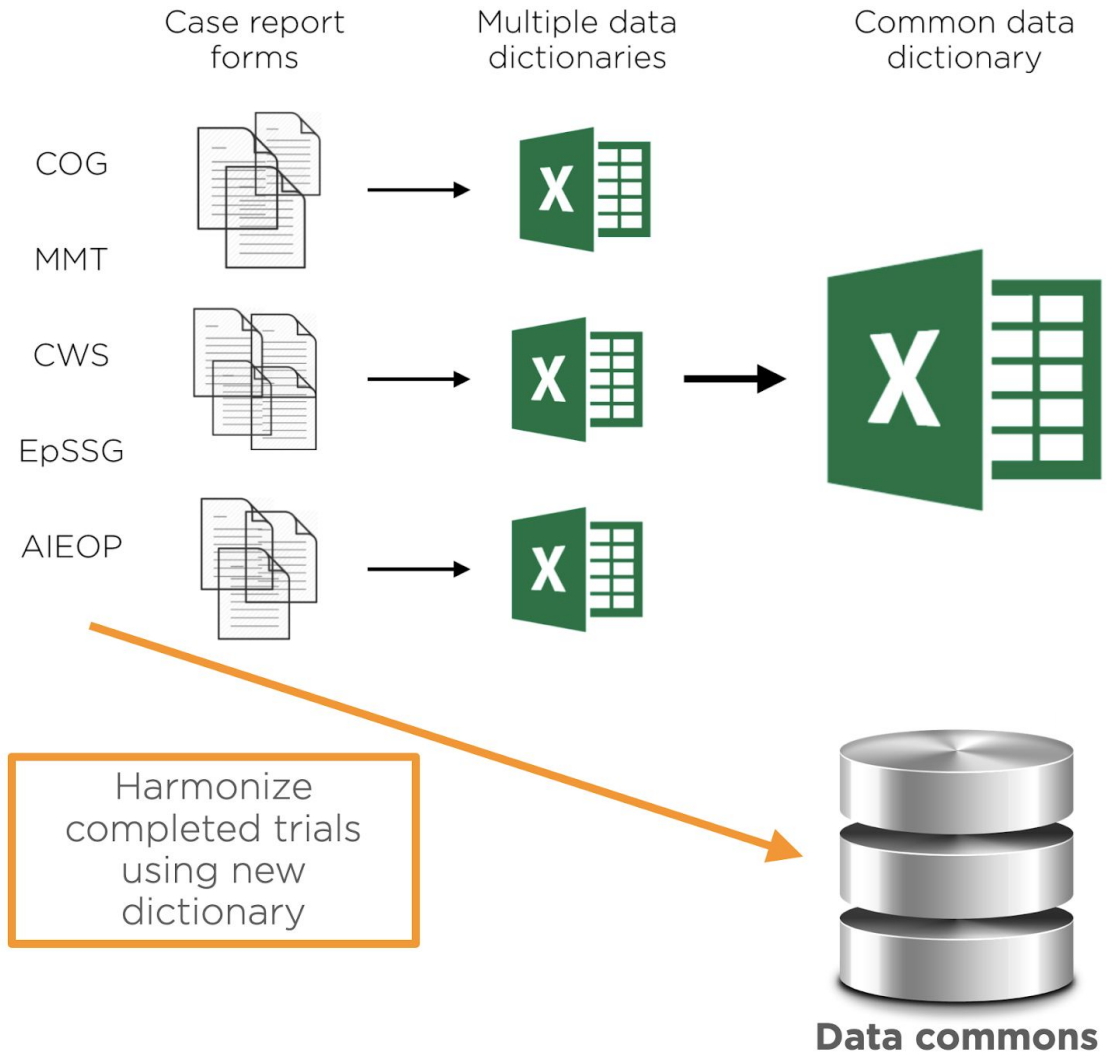
Balloting a consensus data dictionary

Disease consortia



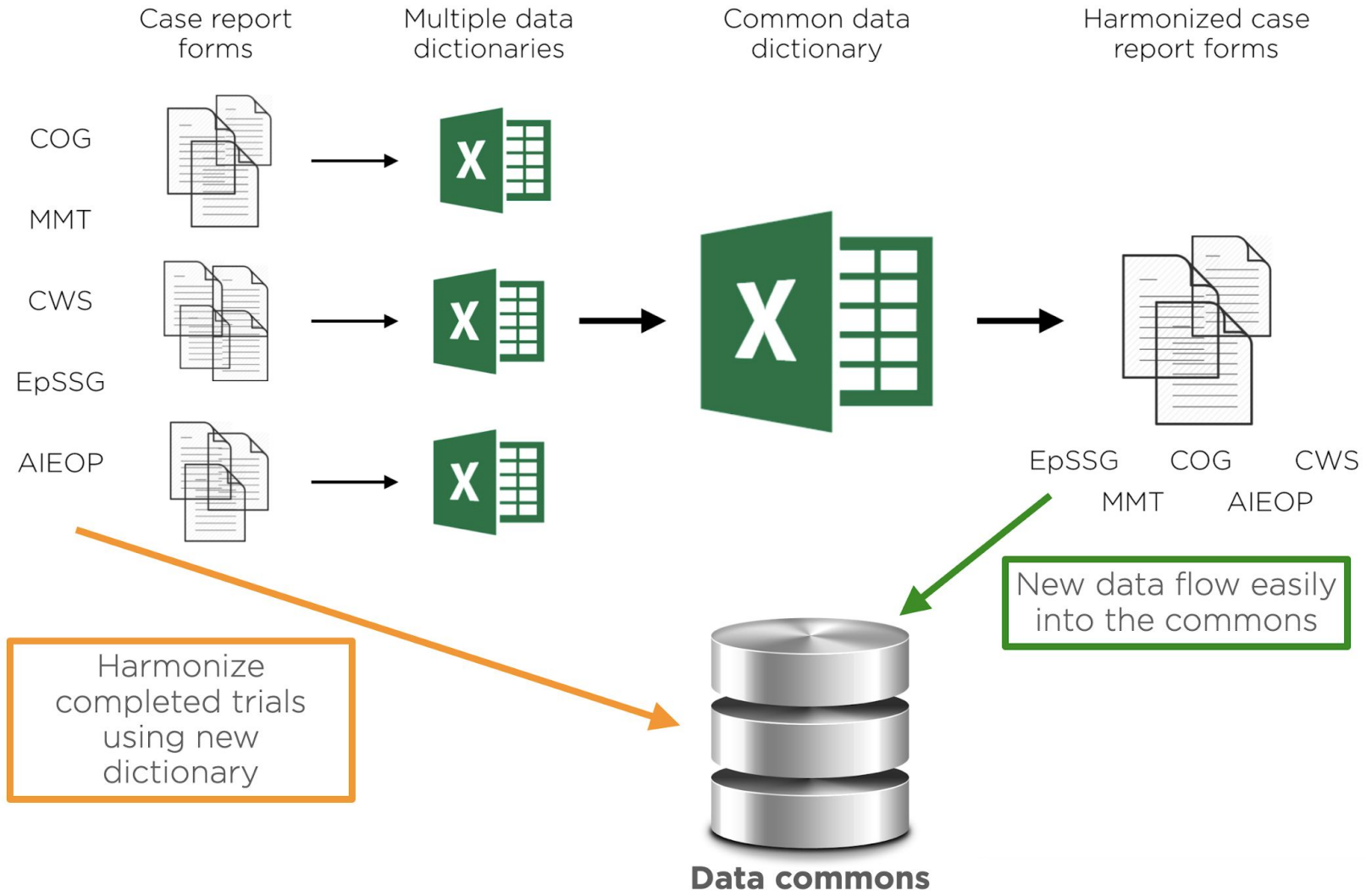
Balloting a consensus data dictionary

Disease consortia

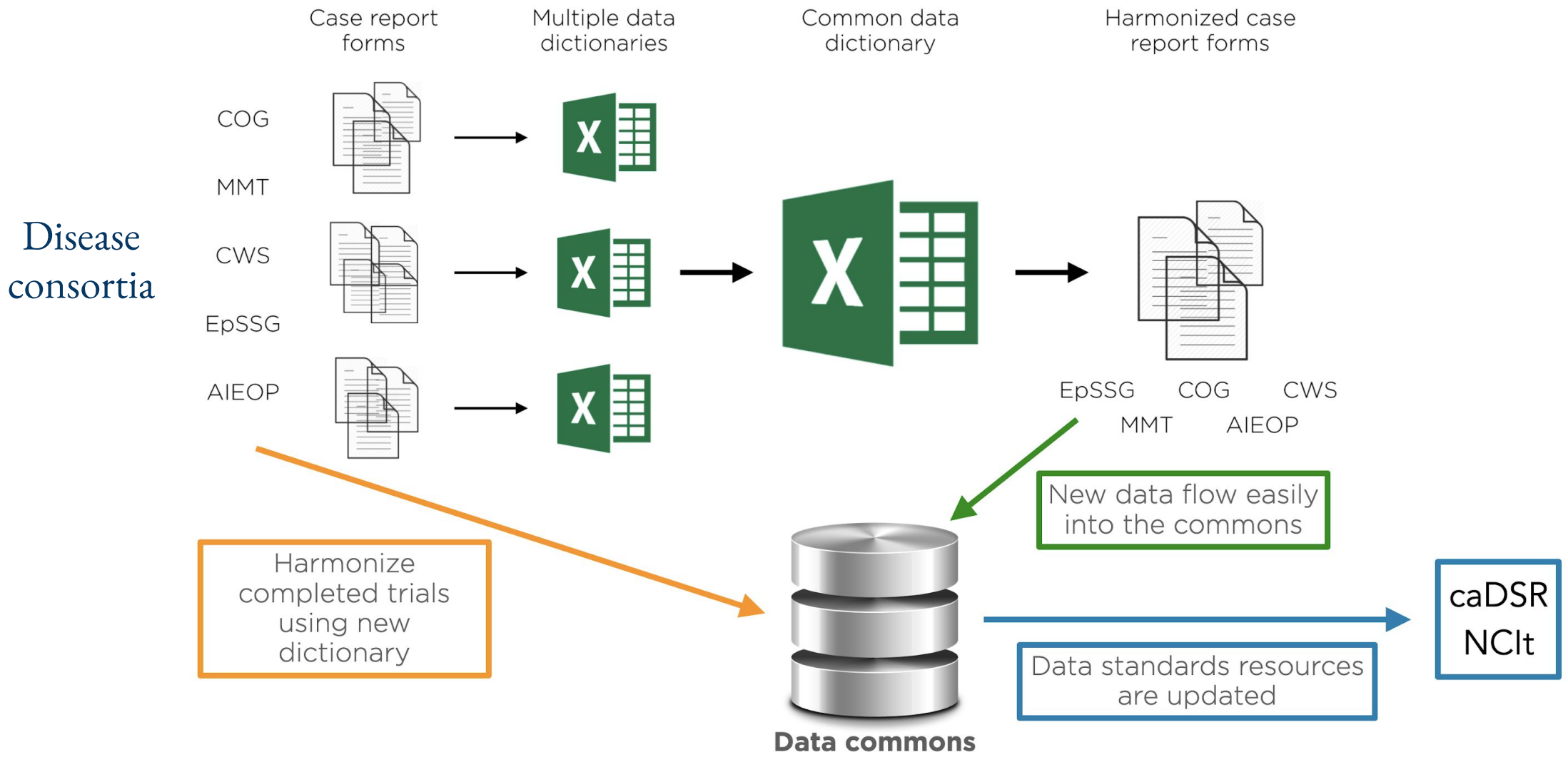


Balloting a consensus data dictionary

Disease consortia



Balloting a consensus data dictionary



Data standards example

Neuroblastoma stage

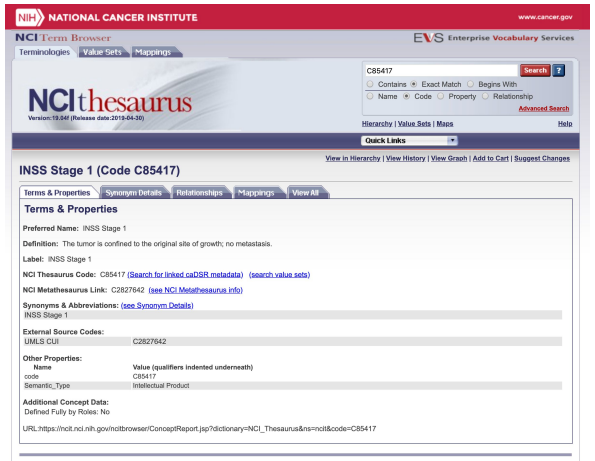
1 = Stage 1	Stage 1
2 = Stage 2a	Stage 2a
3 = Stage 2b	Stage 2b
4 = Stage 3	Stage 3
5 = Stage 4	Stage 4
6 = Stage 4s	Stage 4s
9 = Unknown	Unknown

Data standards example

Neuroblastoma stage

1 = Stage 1	Stage 1	C85417	The tumor is confined to the original site of growth; no
2 = Stage 2a	Stage 2a	C85418	The tumor is unilateral and the resection is grossly incor
3 = Stage 2b	Stage 2b	C85419	The tumor is unilateral and the resection is complete or
4 = Stage 3	Stage 3	C85420	The tumor extends across the midline and the regional l
5 = Stage 4	Stage 4	C85421	Tumor spread to distant lymph nodes, bone marrow, b
6 = Stage 4s	Stage 4s	C85422	Patients are less than one year old with localized primar
9 = Unknown	Unknown	C17998	Not known, not observed, not recorded, or refused.

Data standards example



Neuroblastoma stage

1 = Stage 1	Stage 1	C85417	The tumor is confined to the original site of growth; no
2 = Stage 2a	Stage 2a	C85418	The tumor is unilateral and the resection is grossly incor
3 = Stage 2b	Stage 2b	C85419	The tumor is unilateral and the resection is complete or
4 = Stage 3	Stage 3	C85420	The tumor extends across the midline and the regional l
5 = Stage 4	Stage 4	C85421	Tumor spread to distant lymph nodes, bone marrow, b
6 = Stage 4s	Stage 4s	C85422	Patients are less than one year old with localized primar
9 = Unknown	Unknown	C17998	Not known, not observed, not recorded, or refused.

Building a pan-pediatric cancer data dictionary

Systematic review of all existing data dictionaries	PCDC data model	Template for future data dictionaries
• All PCDC data dictionaries • Comparison across cancer groups • Comparison to other data standards	• Minimum elements vs. overall harmonization • Seven main types: 1. protocol 2. demographics 3. disease attributes 4. tests 5. treatments 6. response 7. events	• Used to inform new dictionaries • Detailed guidance on implementation into data commons

PCDC data model - variables

Demographics: one row per subject					
AML, INRG, EWS, HL, GCT, STS, NRSTS	**SEX**	Code	Subject's biological sex	Male	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Female	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Indeterminate	GCT
				Unknown	AML, INRG, EWS, HL, GCT, STS, NRSTS
AML, INRG, EWS, HL, GCT, STS, NRSTS	**RACE**	Code	Subject's race	American Indian or Alaska Native	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Asian	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Black or African American	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Multiracial	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Native Hawaiian or Other Pacific Islander	AML, INRG, EWS, HL, GCT, STS, NRSTS
				White	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Other	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Unknown	AML, INRG, EWS, HL, GCT, STS, NRSTS
AML, INRG, EWS, HL, GCT, STS, NRSTS	**ETHNICITY**	Code	Subject's ethnicity	Hispanic or Latino	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Not Hispanic or Latino	AML, INRG, EWS, HL, GCT, STS, NRSTS
				Unknown	AML, INRG, EWS, HL, GCT, STS, NRSTS

Choosing the right standard

Considerations:

- Consistency across cancer groups
- Interoperability with non-PCDC data groups (e.g., St. Jude, Dana-Farber)
- Ability to connect clinical data to outside data sources (genomic, imaging, etc.)

Our approach:

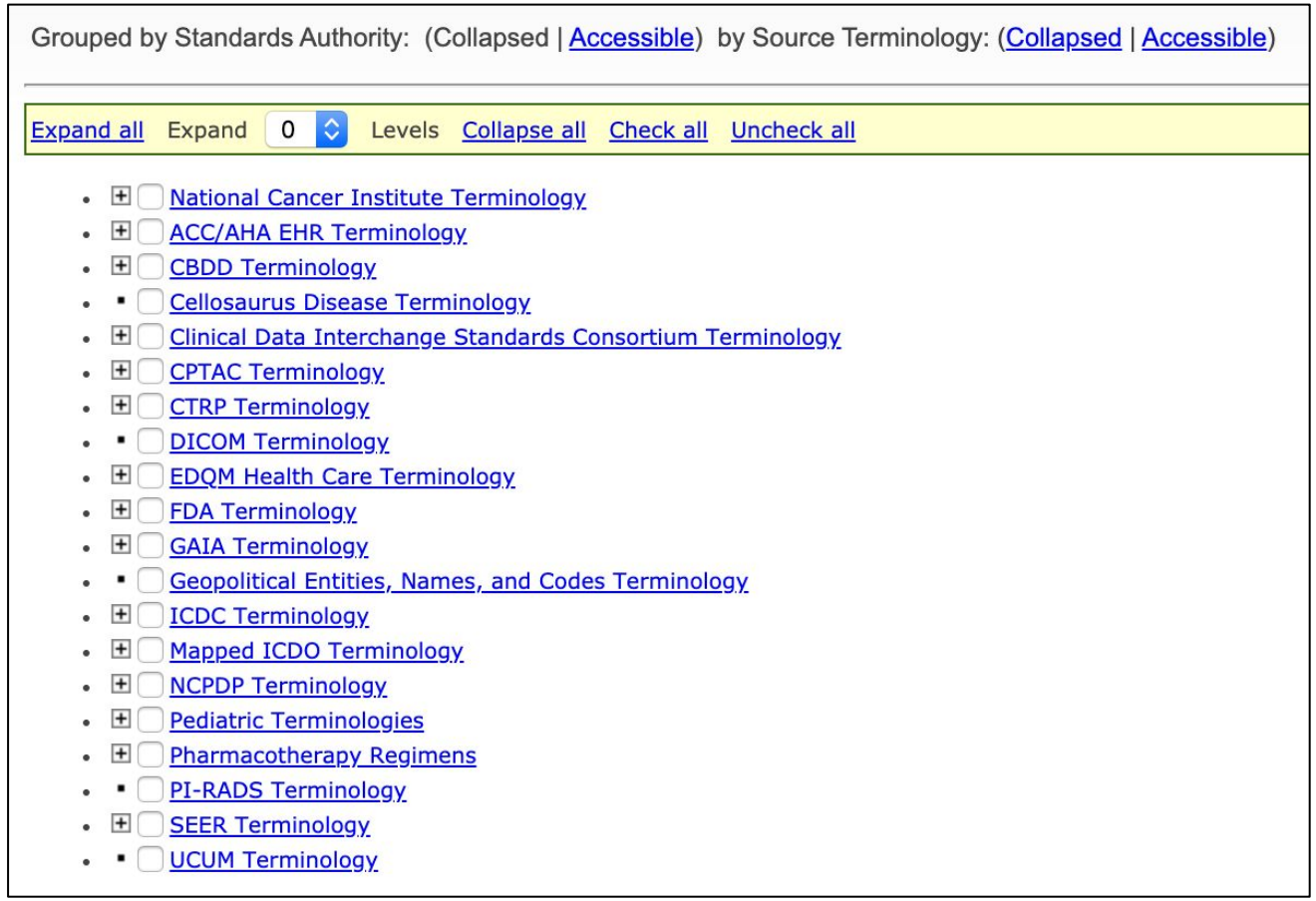
- Working with NCI for standardization
- PCDC Data Dictionary Work Group is composed of international pediatric oncologists, statisticians, and data standards experts

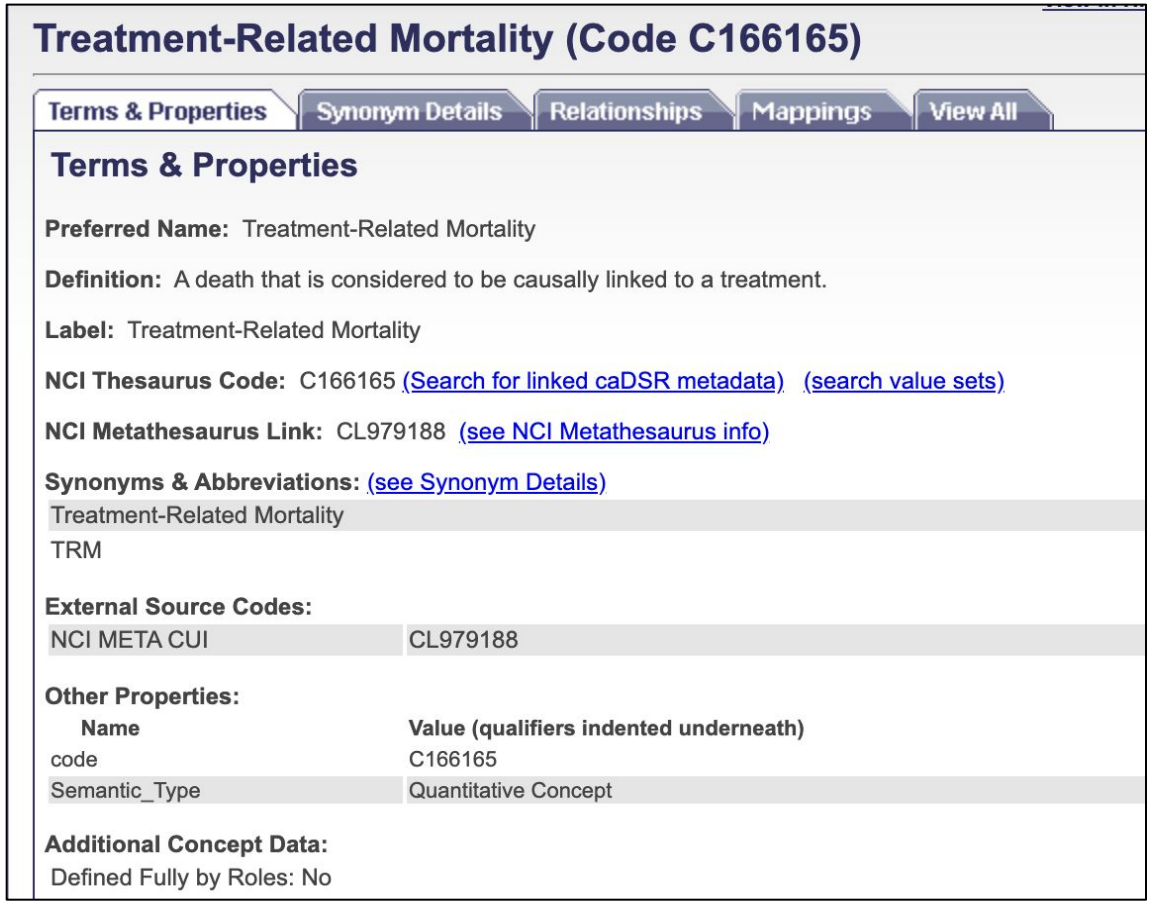
Many different standards

- **International Standards Organization** (ISO) - BRIDG
- **Health Level 7** (HL7) - FHIR, mCODE
- **Clinical Data Interchange Standards Consortium** (CDISC) - SDTM
- **Observational Health Data Sciences and Informatics** (OHDSI) - OMOP
- **Patient-Centered Outcomes Research Institute** (PCORI) - PCORnet CDM
- **Informatics for Integrating Biology and Bedside / Accrual to Clinical Trials** (i2b2/ACT)
- **Sentinel initiative**

Pediatric data standards have not been developed thoroughly.

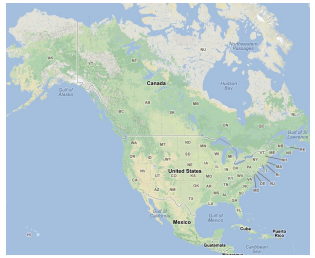
NCI-t


Grouped by Standards Authority: (Collapsed | Accessible) by Source Terminology: (Collapsed | Accessible)
Expand all Expand 0 Levels Collapse all Check all Uncheck all
National Cancer Institute Terminology
ACC/AHA EHR Terminology
CBDD Terminology
Cellosaurus Disease Terminology
Clinical Data Interchange Standards Consortium Terminology
CPTAC Terminology
CTRP Terminology
DICOM Terminology
EDQM Health Care Terminology
FDA Terminology
GAIA Terminology
Geopolitical Entities, Names, and Codes Terminology
ICDC Terminology
Mapped ICDO Terminology
NCPDP Terminology
Pediatric Terminologies
Pharmacotherapy Regimens
PI-RADS Terminology
SEER Terminology
UCUM Terminology


Treatment-Related Mortality (Code C166165)
Terms & Properties
Synonym Details
Relationships
Mappings
View All
Terms & Properties
Preferred Name: Treatment-Related Mortality
Definition: A death that is considered to be causally linked to a treatment.
Label: Treatment-Related Mortality
NCI Thesaurus Code: C166165 (Search for linked caDSR metadata) (search value sets)
NCI Metathesaurus Link: CL979188 (see NCI Metathesaurus info)
Synonyms & Abbreviations: (see Synonym Details)
Treatment-Related Mortality
TRM
External Source Codes:
NCI META CUI CL979188
Other Properties:
Name Value (qualifiers indented underneath)
code C166165
Semantic_Type Quantitative Concept
Additional Concept Data:
Defined Fully by Roles: No

Neuroblastoma data commons

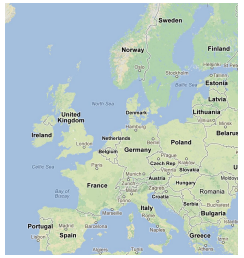
INTERNATIONAL NEUROBLASTOMA RISK GROUP (INRG), 2004




Children's Oncology Group (COG)

Germany

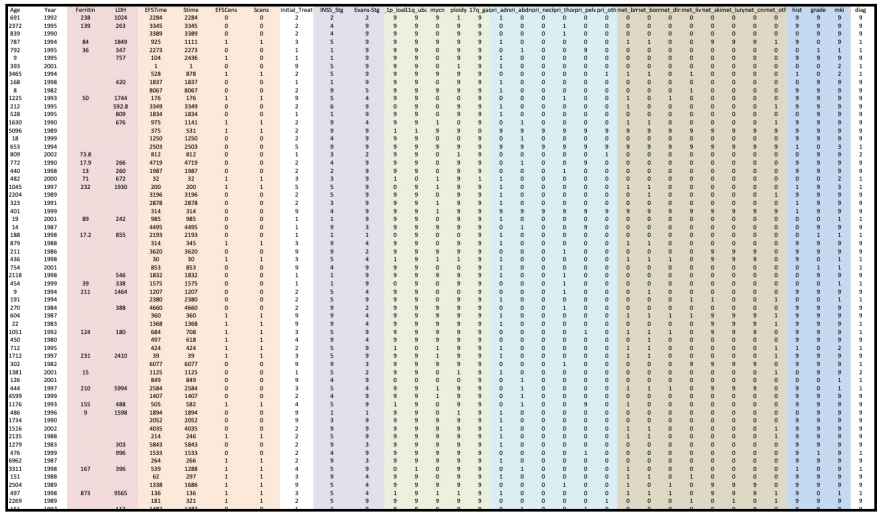
Japan



SIOPEN

Group	Number
COG	4235
Germany	1938
Japan	470
SIOPEN	936
Total	8800

The good news: 8800 patients



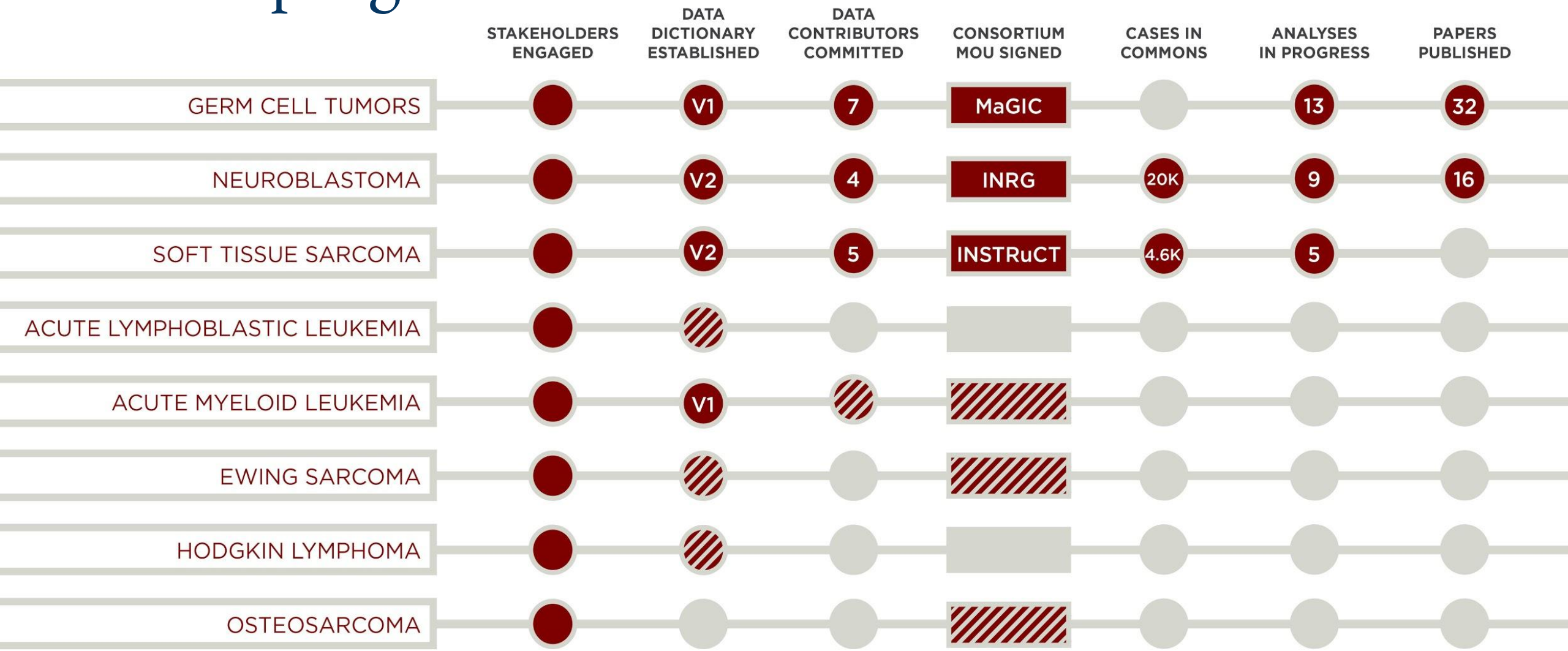
The bad news: data in Excel

Neuroblastoma data commons - Cohort discovery

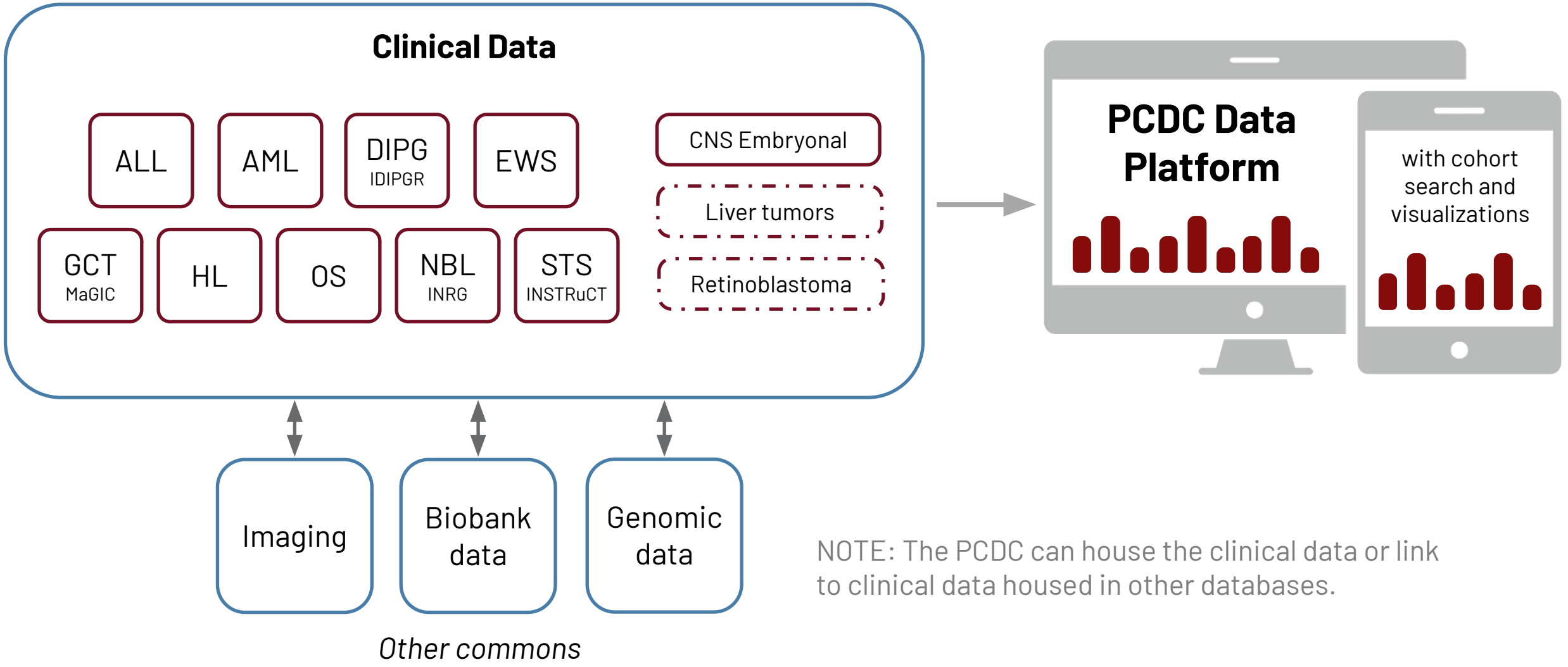
Publicly available at inrgdb.org

External data

PCDC progress to date



An integrated pediatric cancer data commons



NOTE: The PCDC can house the clinical data or link to clinical data housed in other databases.

Guiding principles governing the PCDC

1 **OUR GOAL**

is to lift barriers and connect researchers to data

2 **STAKEHOLDER APPROVAL**

for data release from any disease commons

3 **CONTRIBUTOR APPROVAL**

for data release from the original contributing group

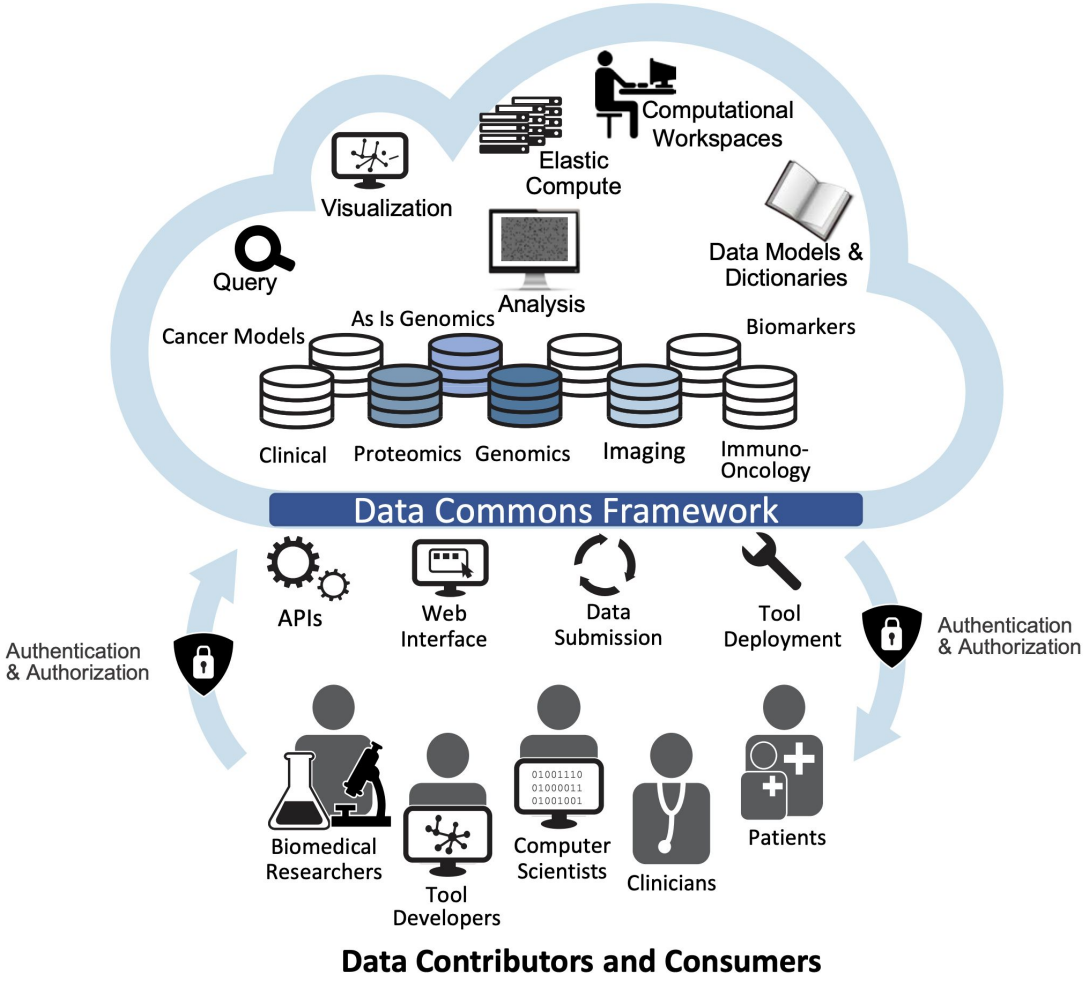
4 **REPRESENTATION**

on the Executive Committee from every disease group

5 **RECOGNIZE REGIONAL DIFFERENCES**

European and US legal regulations are not the same

NCI Cancer Research Data Commons ecosystem

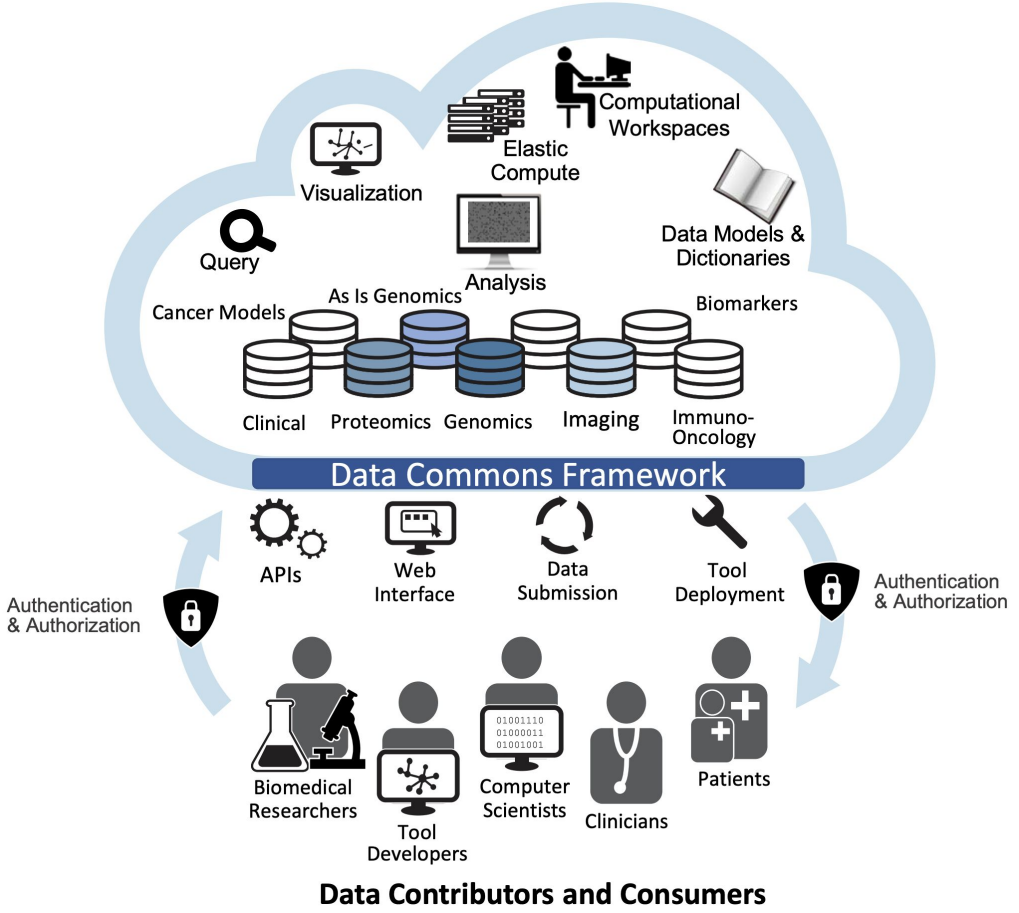


Credit: Allen Dearry

Center for Cancer Data Harmonization (CCDH)

CENTER *for* CANCER DATA HARMONIZATION

- **Facilitate** retrospective and prospective semantic harmonization of data across nodes of the CRDC
- **Coordinate** the community to ensure implementation of standards that will facilitate interoperability of heterogeneous data types and CRDC resources
- **Find agreement** across the communities built around CRDC
  - match and extend data models
  - annotation, harmonization
  - quality assurance

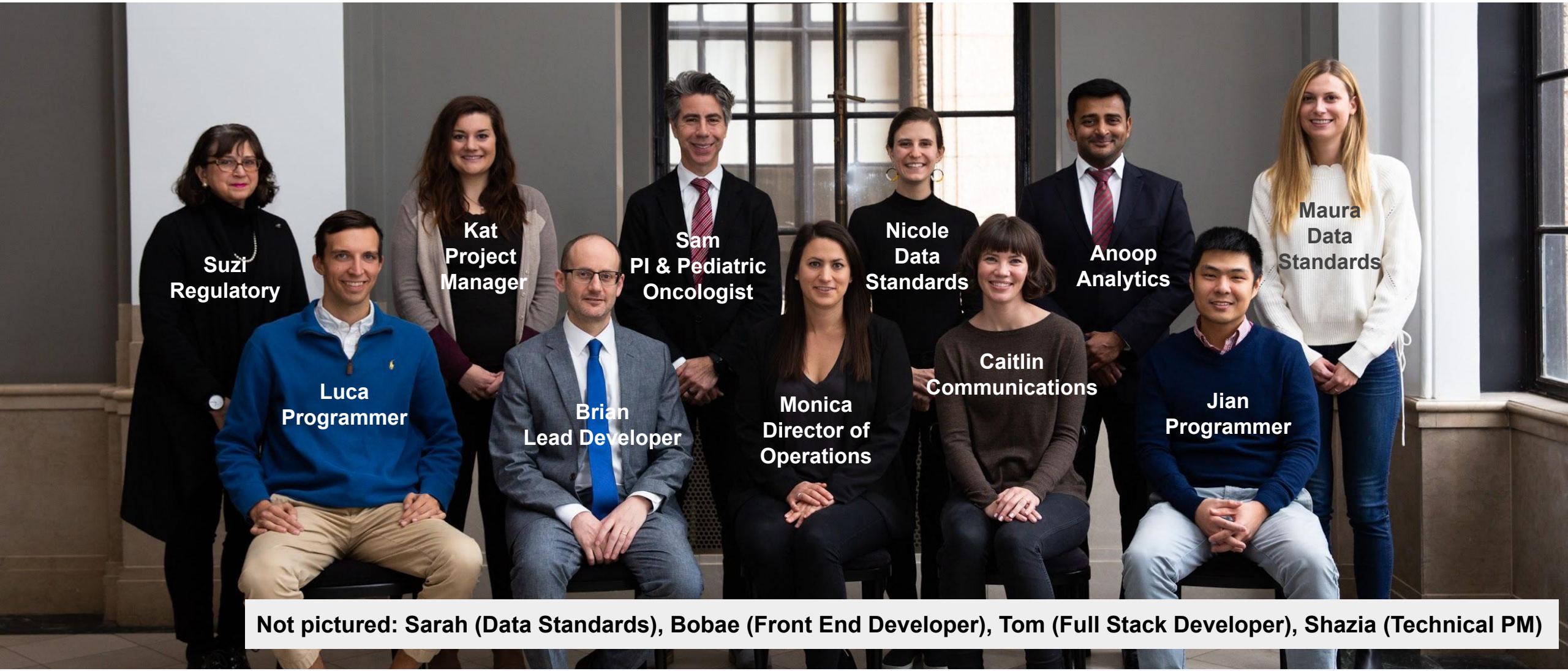


Credit: Allen Dearry

Take home points

- Studying pediatric cancer requires collaboration and sharing
- Data sharing needs to be built on a foundation of trust and consensus
- Connecting disparate data types and sources enriches research
- Consensus data standards are critical for the success of national and international data ecosystems - allowing aggregation across trials and diseases
- Early adoption of data standards and consideration for the lifecycle of the data is critical to accelerating progress in discovery

The Pediatric Cancer Data Commons team



Not pictured: Sarah (Data Standards), Bobae (Front End Developer), Tom (Full Stack Developer), Shazia (Technical PM)

We gratefully acknowledge our funders