Utilizing Data to Make Advancements for Cancer

Jill Barnholtz-Sloan, PhD

Acting Director, CBIIT
Associate Director for Informatics and Data Science, CBIIT
Senior Investigator, DCEG

I do not have any conflicts of interest. I am a full-time paid employee of the NIH/NCI

NIH NATIONAL CANCER INSTITUTE

Everyone has a role!

National Cancer Plan

How the goals interact to improve health outcomes for everyone

Agenda

1. **Open Science -- Importance and Goals of Data Sharing**
2. Cancer Research Data Commons (CRDC) and Childhood Cancer Data Initiative (CCDI)
3. Data Drives Discovery
4. Other “Big Data” Resources
5. Misc Items of Interest

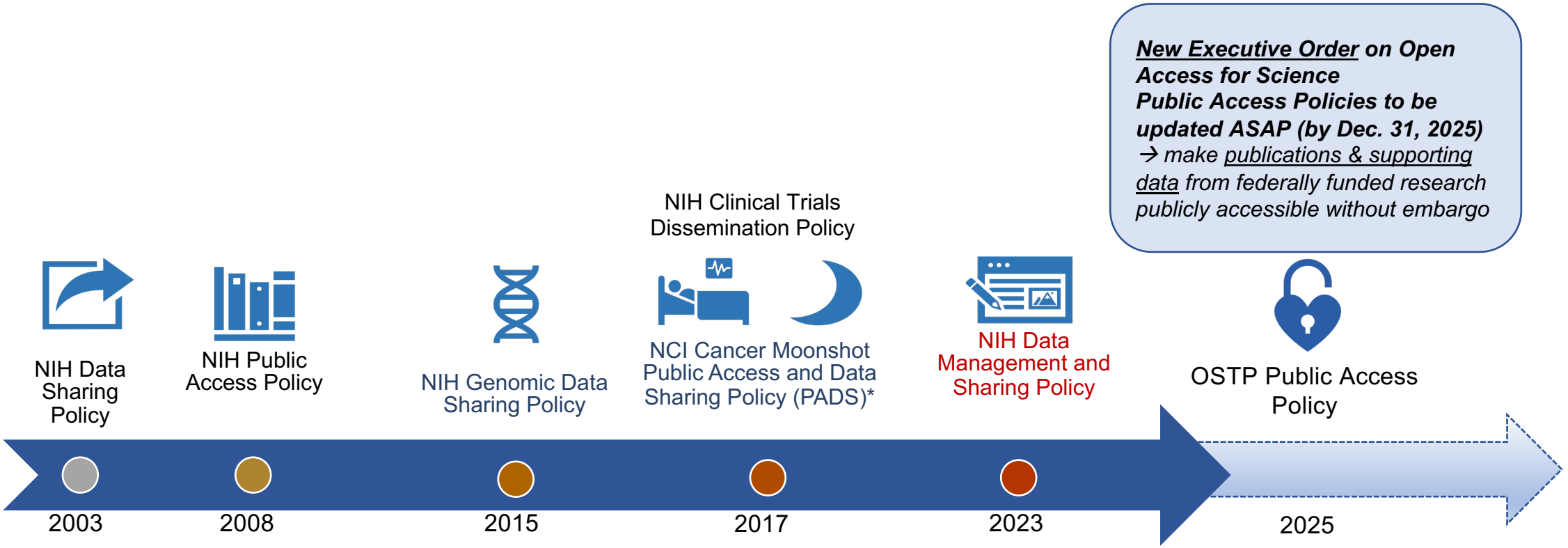
Open Science

- Advancing national open science policy
- Providing access to the results of the nation's taxpayer-supported research
- Accelerating discovery and innovation
- Promoting public trust
- Driving more equitable outcomes

NIH's Data Management and Sharing Policy went into effect on January 25, 2023, fulfilling the memorandum's provisions around public access to scientific data.

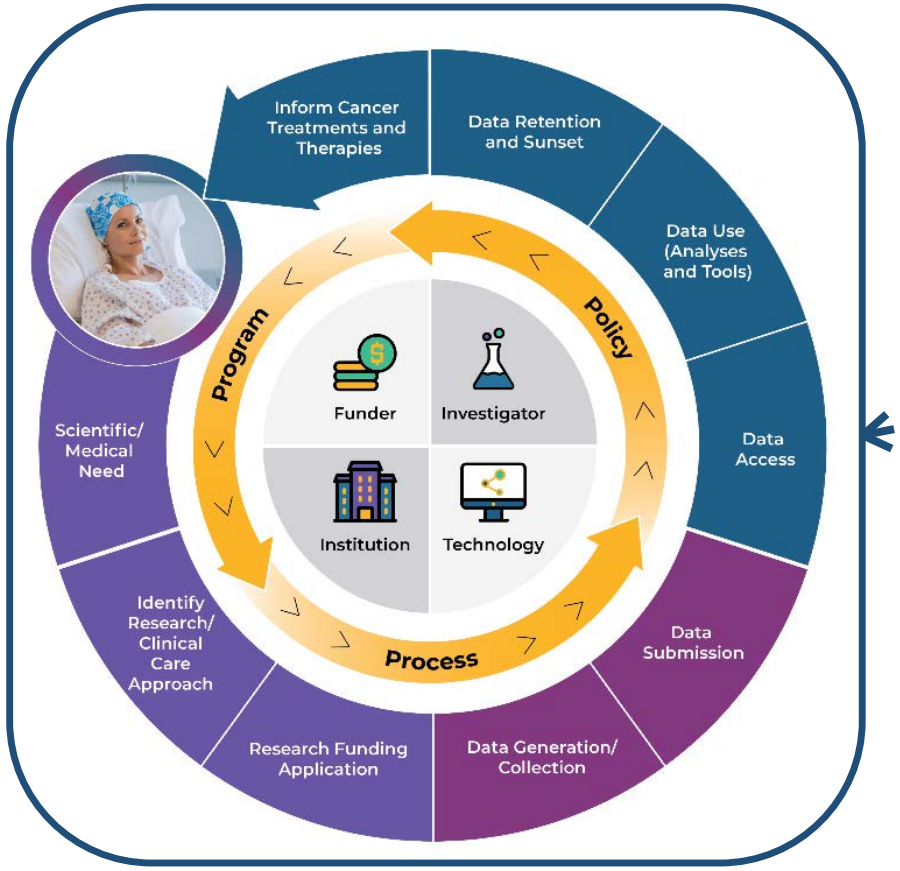
Driving Impactful Data Sharing

Implement data management, sharing and access policies that ensure **rapid, free** and **immediate access** to many types of data



Boja ES, Guidry Auvil JM. Nat Med. 2024 May 23. doi: 10.1038/s41591-024-02950-7.

Keys to Impactful Discovery in Scientific Data Lifecycle



Science: Critical Questions to Answer

Plan research that defines scientific needs and essential gaps to be filled using structured datasets

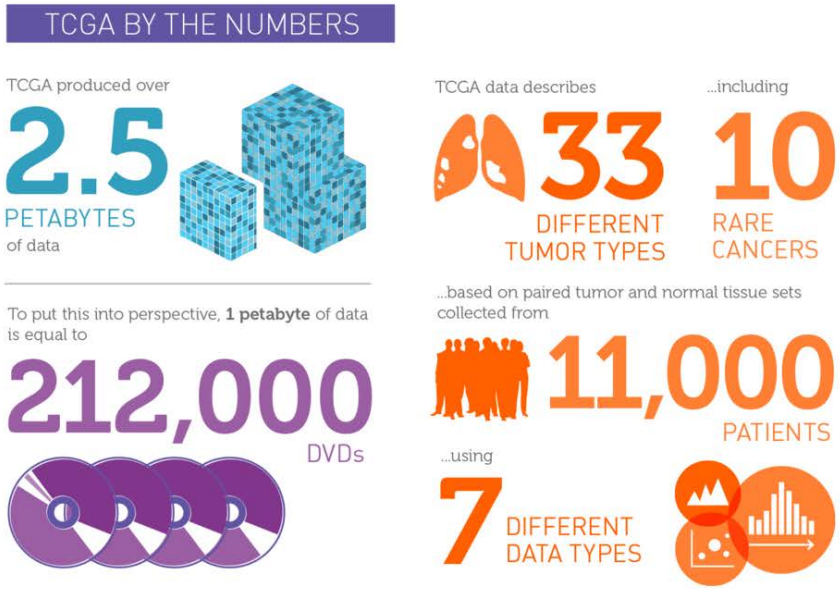
Policy & Process: Promote Broad Use

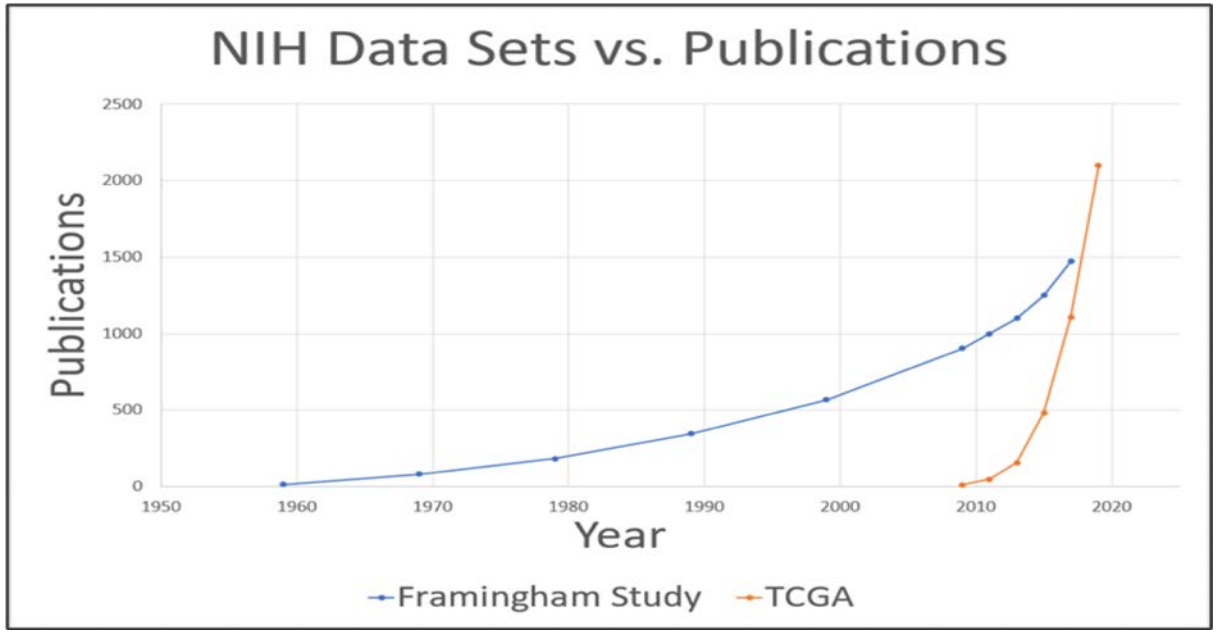
Implement data management, sharing and access policies that ensure rapid, free and immediate access to many types of data

Technology: Support FAIR Principles

Share data via technology platforms and tools that employ standards to make data findable, accessible, interoperable and reusable

The Cancer Genome Atlas: Success in Open Science





	Framingham Heart Study	The Cancer Genome Atlas
Study Length	70 years	12 years
Cases Studied	15,144	11,429
Publications	**3,698**	**3,747**
Data Use	Consortia-based; most data available on publication	Collaborative Teams & Public Use of Data; All data immediately available
Approved Users	715 (Individual Level Data)	3,335 (Individual Level Data)

The Cancer Moonshot: Success in Mission-Driven Science

Take Home Message: Purposeful, broad, early access to data leads to much faster and impactful outcomes

What Policies Promote Broad Sharing and Reuse?

Description of the *data and metadata* and documentation

Related *tools, software, code* to analyze data

Standards for the data/metadata

Data *preservation, access, and associated timelines*

Access, distribution, and reuse considerations

Oversight of data management and sharing

Streamlining Access: Enhancing Data Utility

Support FAIR Principles

- ***Unrestricted-access****: Data are publicly available; includes study protocols, metadata, certain phenotype data
- ***Managed-access***: Investigators must sign a data use agreement and obtain approval from NCI program staff (*scientific review*)
- ***Controlled-access****: Investigators must sign a data access agreement and obtain approval from NIH Data Access Committees (*ethical review*) to use the requested data (e.g., dbGaP); includes *individual-level sequence data and potentially identifiable phenotype and analyzed data*

**Informed consent is the basis for institutions to determine the appropriateness of submitting human data to unrestricted or controlled-access NIH data repositories*

Agenda

1. Open Science -- Importance and Goals of Data Sharing
2. **Cancer Research Data Commons (CRDC) and Childhood Cancer Data Initiative (CCDI)**
3. Data Drives Discovery
4. Other “Big Data” Resources
5. Misc Items of Interest

Vision for the Cancer Data Ecosystem

Stakeholder Focus

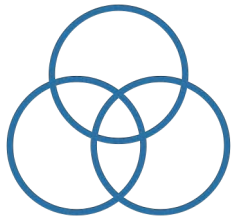
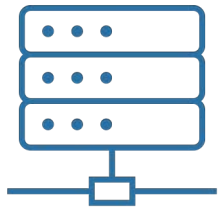
- Include all scientists and clinicians (of all technical abilities)
- Intramural and Extramural scientists and staff

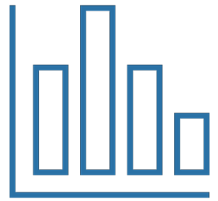
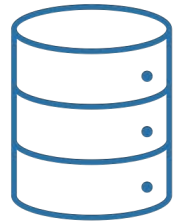
Lower Barriers

- Data submission
- FAIR data access, search, and retrieval
- Integration of data for cross-domain analysis
- Analysis platforms, tools, and workflows

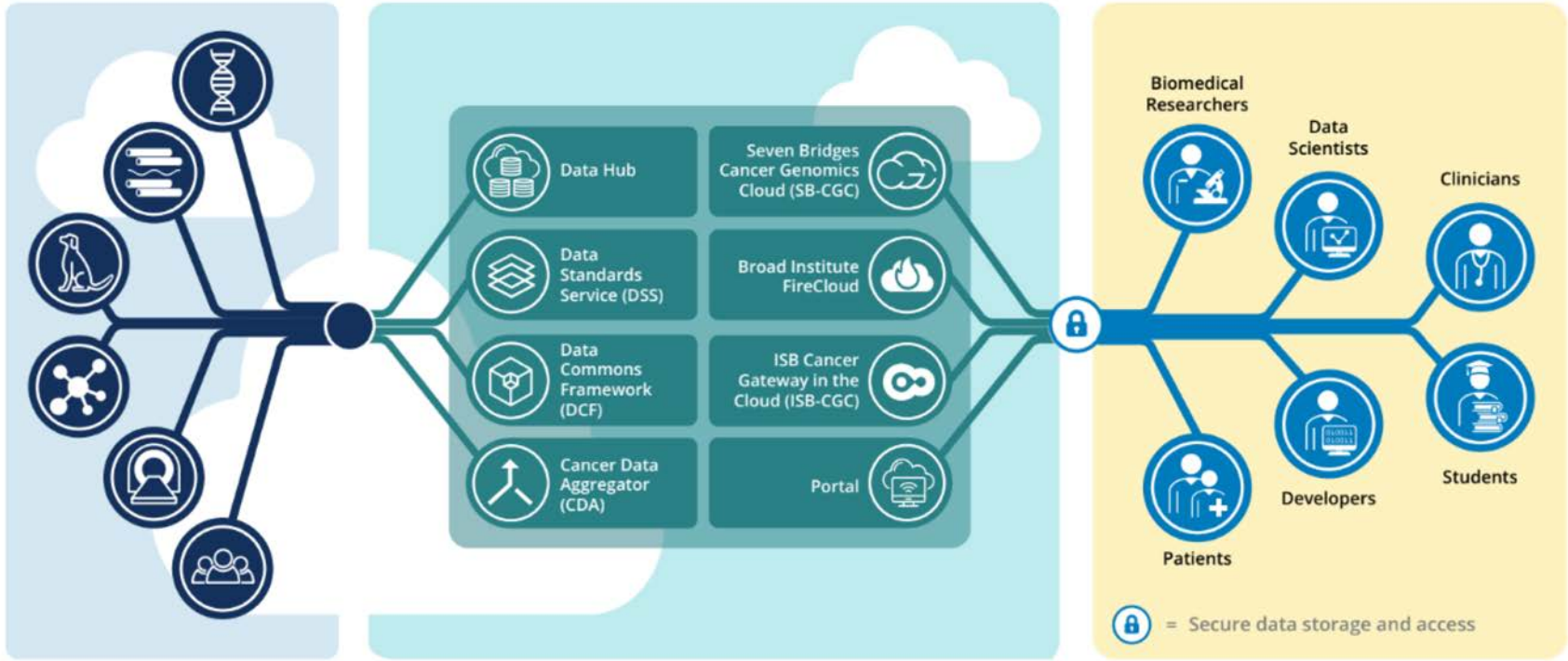
Infrastructure & Sustainability

- Security and appropriate access for sensitive data
- Sustainable, reusable, and uniform architecture
- Comprehensive plan for long term data storage and accessibility to tools

NCI Cancer Research Data Commons (CRDC)



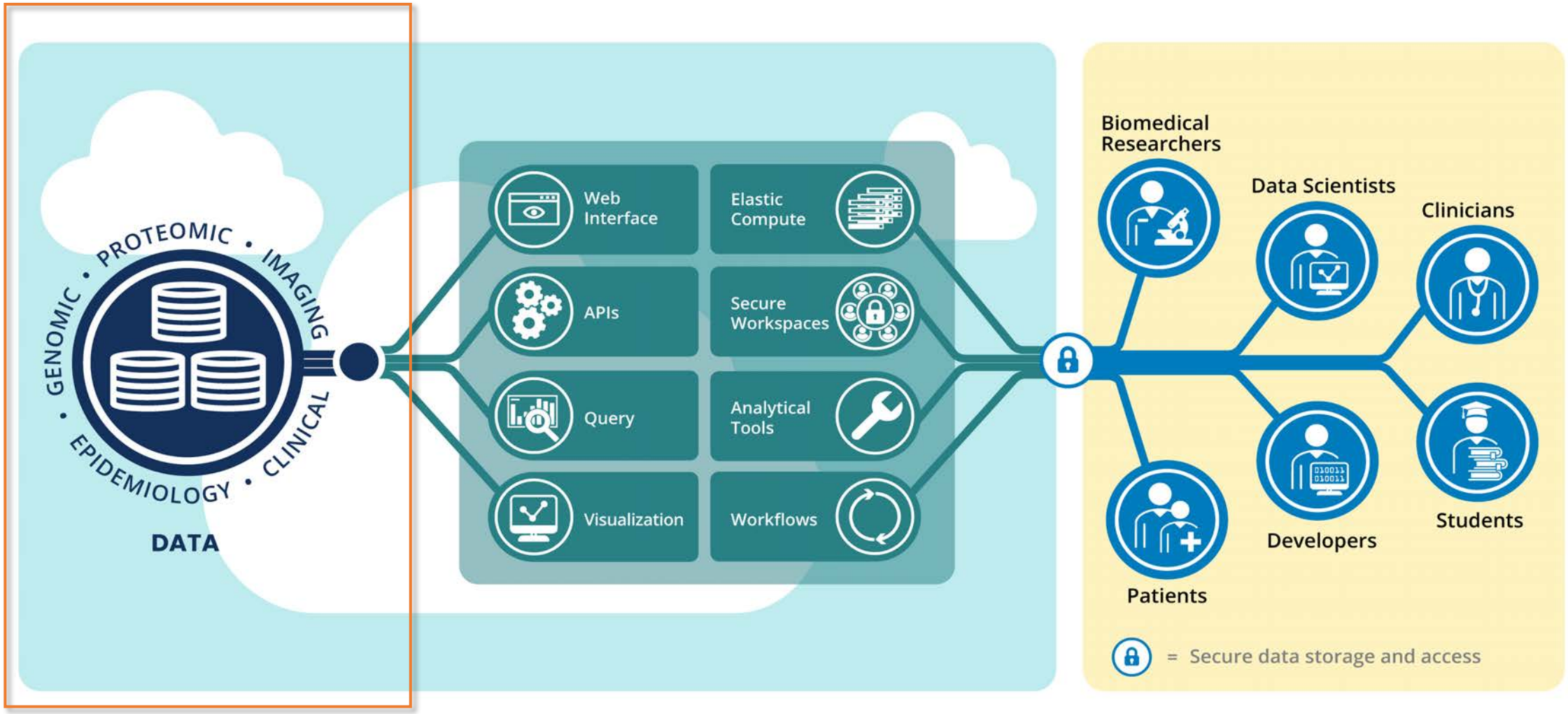
Mission

- The CRDC **empowers researchers** by providing access to a cancer data ecosystem with **state-of-the-art visualization, analysis, and interoperability tools** in a flexible, cloud-based computational environment

Goals

- Preserve **long-term value** of NCI-funded data
- Improve **data submission, access, and interoperability**
- Accelerate cancer research through **integrative analysis** of multi-modal data

CRDC Ecosystem – Data Commons



https://datacommons.cancer.gov/

CRDC Ecosystem -- Data Commons

Data Commons

Genomic Data Commons

Proteomic Data Commons

Imaging Data Commons

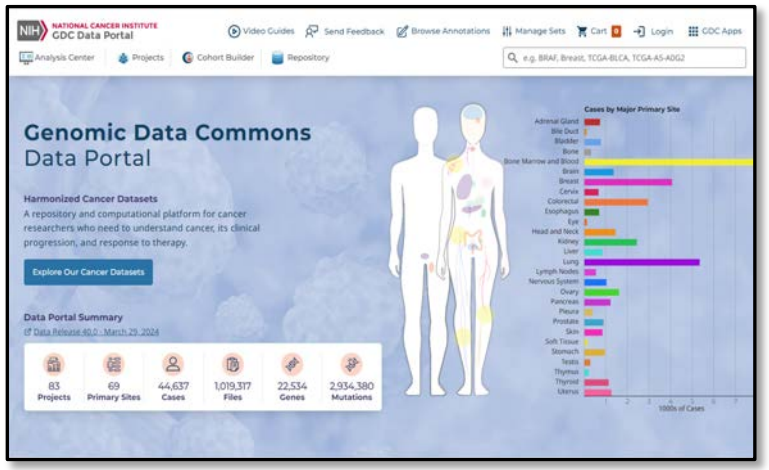
Integrated Canine Data Commons

Cancer Data Service

Clinical and Translational Data Commons – Live Sept 2024

MVP Sept 2024 – Population Science Data Commons

CRDC Data Commons



Genomic Data Commons

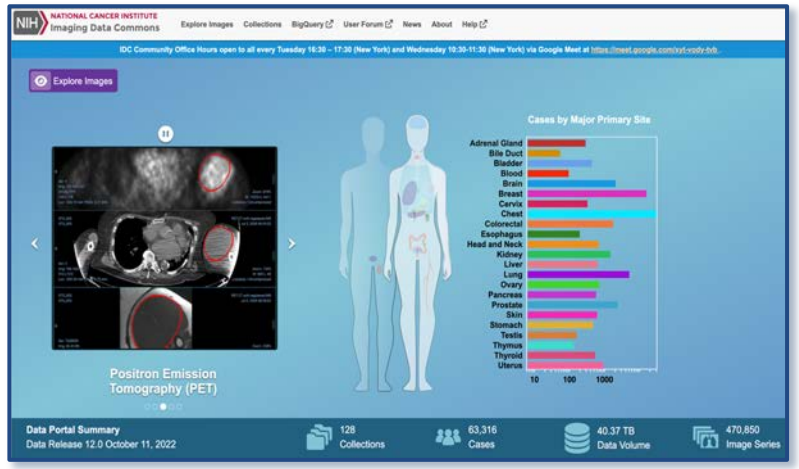
- Share, analyze, and visualize genomic data
- Harmonized to the same genome standard and variant calling pipeline

https://portal.gdc.cancer.gov/

Proteomic Data Commons

- Filter, query, search, visualize and download proteomic data and metadata
- Data harmonization pipeline to uniformly analyze all PDC data

https://pdc.cancer.gov/

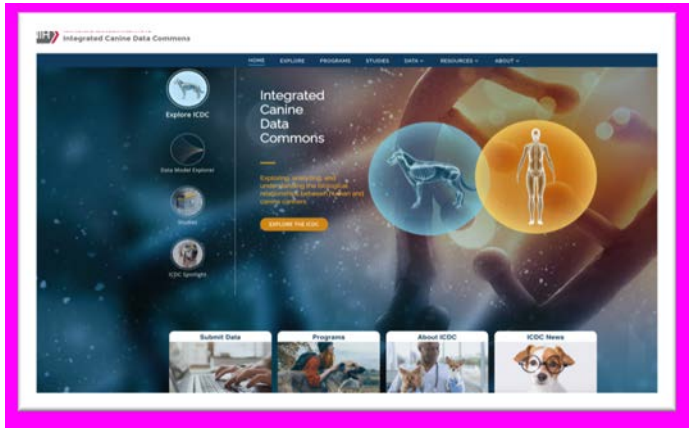


Imaging Data Commons

- Share, analyze, and visualize de-identified multi-modal imaging data, as medical images (MRI, PET, CT)
- Uses DICOM standard

https://imaging.datacommons.cancer.gov/

CRDC Data Commons



Integrated Canine Data Commons

- Share data from canine clinical trials for comparative research
- All data (including raw sequence data) are open access

https://caninecommons.cancer.gov/

Cancer Data Service

- Access NCI-funded data currently not hosted by other CRDC data commons
- All datatypes accepted

https://dataservice.datacommons.cancer.gov/

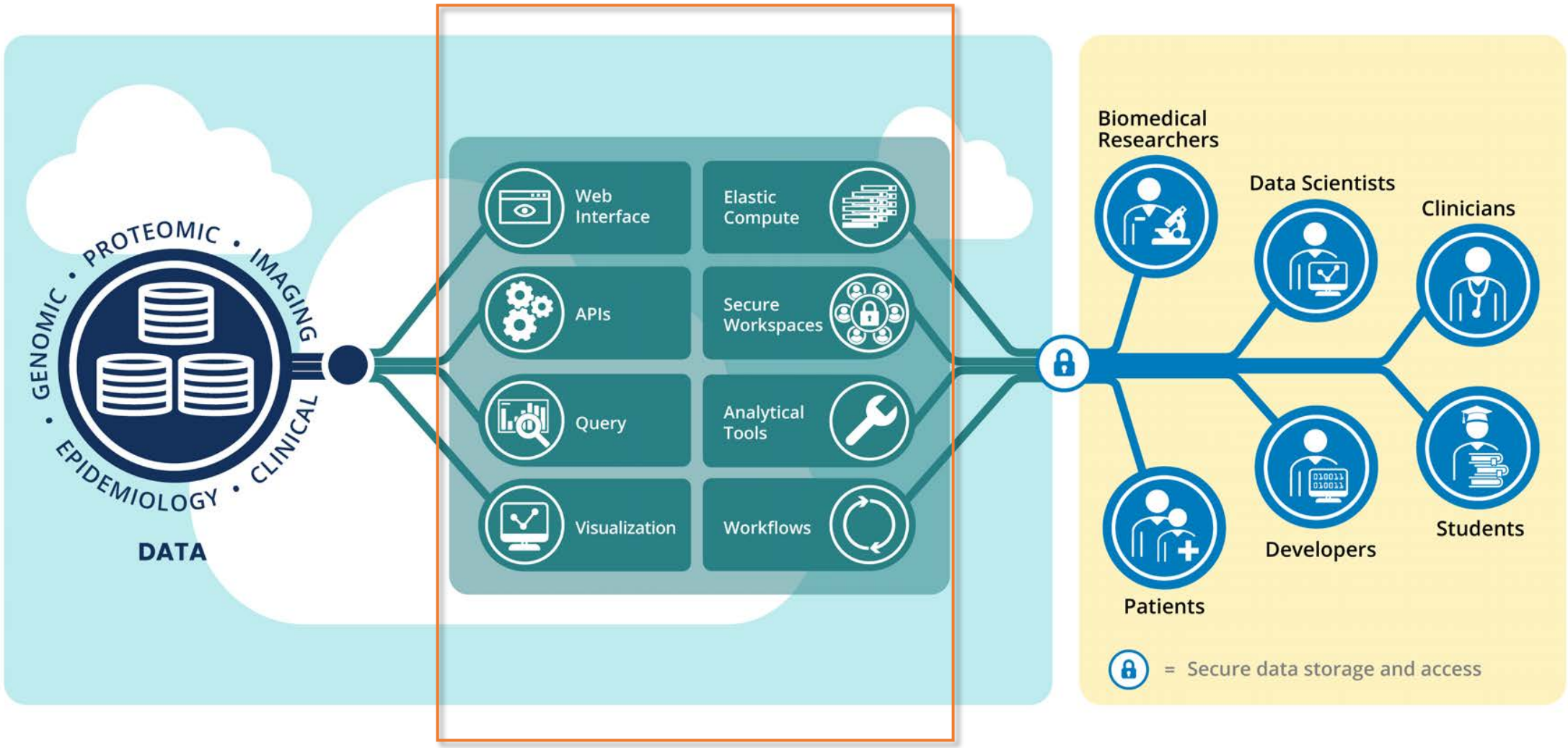
Clinical and Translational Data Commons

- First release: Cancer Moonshot Biobank
- Sept 2024

https://clinical.datacommons.cancer.gov

Data types available across data commons: WGS, WXS, RNAseq, miRNA-seq, scRNAseq, ATAC-seq, DNA methylation, mass spectrometry-based proteomic data, DICOM.

CRDC Ecosystem – Cloud Resources



https://datacommons.cancer.gov/

CRDC Cloud Resources

Broad FireCloud (FC), powered by Terra

- Based on the Google Cloud Platform (GCP)
- Offers extensive repositories of pre-built tools and workflows in the Workflow Definition Language (WDL).

The ISB Cancer Gateway in the Cloud (ISB-CGC)

- Offers Google Cloud Platform (GCP) native tools and Google BigQuery for big data analytics and Google Compute Engine for complex workflow execution.
- Designed for users looking to use derived data.

The Seven Bridges Cancer Genomics Cloud (SB-CGC), powered by Velsera

- Based on the Amazon Web Services (AWS) platform
- Offers a curated library of over 850 tools and workflows optimized for the cloud using the Common Workflow Language (CWL).

CRDC Centralized Data Submission

Submission Request Form: December 2023

Data Submission MVP 2024

Live: June 2024

Data Submission Requestor

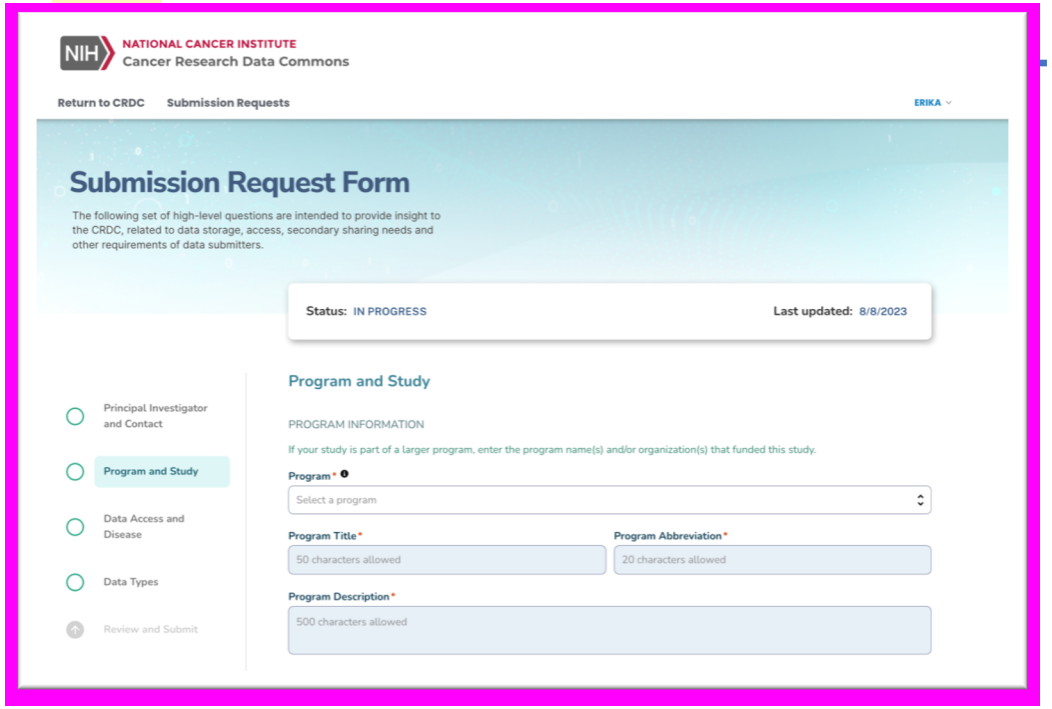
Individual PI/ Program submits Submission Request Form

CRDC Data Submission Portal acknowledge receipt

Request sent to *Submission Review Committee*

No, do not accept → End

Yes, accept



Request sent to CRDC *Data Intake Team* → OK to proceed → Data Upload and Validation → Coordination with target Data Commons

Aggregate Data Across all Data Commons

Cancer Data Aggregator (CDA)

Enables users to query and connect data distributed across the CRDC for integrative analysis.

A fully point and click graphical interface

A single API in a common language

Multiple APIs, or just data in buckets

```
Q('primary_diagnosis_site = “brain"').subject.count.run().to_dataframe()
```

```
Q('primary_diagnosis_site = "brain"').subject.count.run()
Getting results from database

                    Total execution time: 0
                    min 10.813 sec 10813 ms

  files : 4924982

    total : 3015
```

sex	count
None	1378
female	653
male	981
not reported	3

race	count
None	1378
white	1312
american indian or alaska native	4
black or african american	96
not reported	136
Unknown	21
asian	33
not allowed to collect	25
other	9
native hawaiian or other pacific islander	1

ethnicity	count
None	1378
not hispanic or latino	1286
not reported	219
hispanic or latino	85
Unknown	22
not allowed to collect	25

cause_of_death	count
None	2746
Not Reported	199
Cancer Related	48
Unknown	8
Not Cancer Related	9
Infection	3
Surgical Complications	2

subject_identifier_system	count
IDC	2585
PDC	309
GDC	1455

https://cda.readthedocs.io/en/latest/QuickStart/QuickStart/

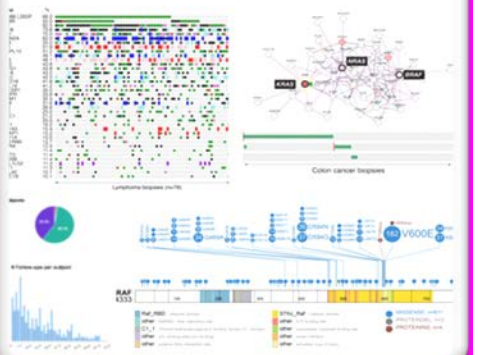
Lessons Learned: Lowering Barriers

Data Submission Dashboard

Data Standardization, Harmonization & Integration

Multi-Modal Cancer Data

Data Discovery Dashboard

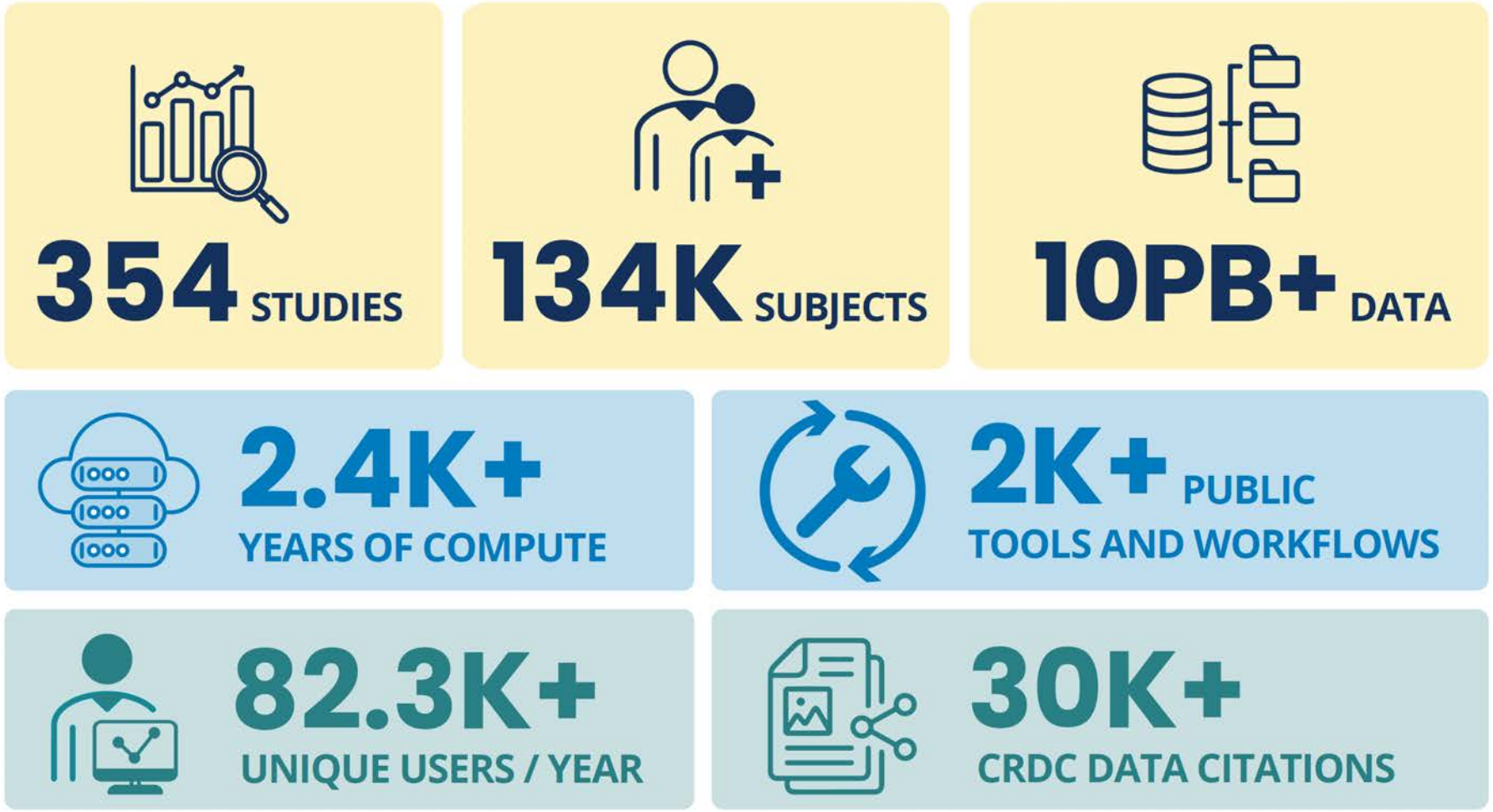


Data Submission Concierge

Data Generators and Users

Data Discovery Concierge

CRDC: Statistics & Impact



datacommons.cancer.gov

Kim et al, Cancer Research

Support for Researchers

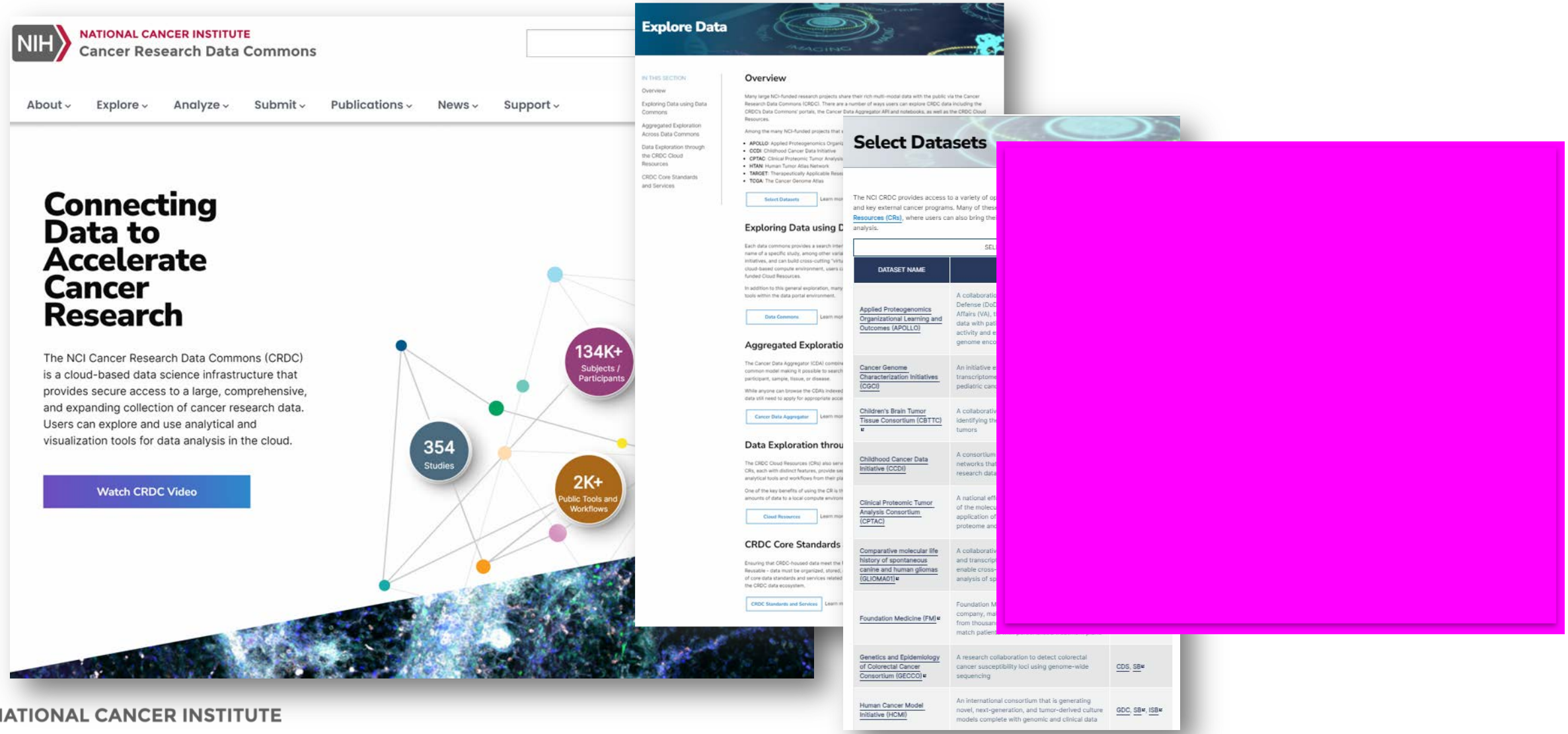
https://datacommons.cancer.gov/support-for-researchers

CRDC COMPONENT	RESOURCE	AVAILABLE SUPPORT
Cloud Resources	Broad FireCloud Powered by Terra (FC)	• How to Set Started on FC • Broad Institute FireCloud Workshop Tutorials • Terra Self-Service Learning Resources • Broad Institute FireCloud FAQs
	ISB Cancer Gateway in the Cloud (ISB-CGC)	• How to Get Started on ISB-CGC • ISB-CGC FAQs The ISB-CGC team offers virtual office hours through Google Meet. Note that the link is different for each of the days. • Tuesdays at 2:00 pm ET; Link: http://meet.google.com/jkg-cxke-yzs • Thursdays at 11:00 am, ET; Link: http://meet.google.com/jai-kgkg-sii Find tutorials and user guides on the ISB Cancer Gateway website.
	Seven Bridges Cancer Genomics Cloud (CGC), powered by Velsera (SB-CGC)	• How to Get Started on SB-CGC • SB-CGC Introduction to the CGC Webinar • SB-CGC Scaling Single-Cell Research • SB-CGC Troubleshooting Tutorial The Seven Bridges team offers virtual office hours through Google Meet at https://meet.google.com/kbs-ojnj-dcg at the following times: • Tuesdays at 10:00 am ET • Thursdays at 2:00 pm ET Learn more about SB-CGC through their user guides, video tutorials, and webinars.

CRDC COMPONENT	RESOURCE	AVAILABLE SUPPORT
Data Commons	Genomic Data Commons (GDC)	• How to get started on GDC • GDC Webinars • GDC FAQs • NGS Studies of Familial Data Using Cloud Computing
	Proteomic Data Commons (PDC)	• PDC FAQs • NCI OCCPR Webinar on PDC
	Imaging Data Commons (IDC)	The IDC offers community office hours every week through Google Meet at https://meet.google.com/xyt-vody-tvb. • Tuesdays, 16:30 – 17:30 (ET/New York) • Wednesdays, 10:30-11:30 (ET/New York) Learn more from the IDC user guide and white papers. If you have questions about the IDC, email the team at support@canceridc.dev or start a thread in their online forum.
	Integrated Canine Data Commons (ICDC)	If you have questions, please email the ICDC team at: ICDCHelpDesk@mail.nih.gov.
	Cancer Data Service (CDS)	If you have questions, please email the CDS team at: CDSHelpDesk@mail.nih.gov.

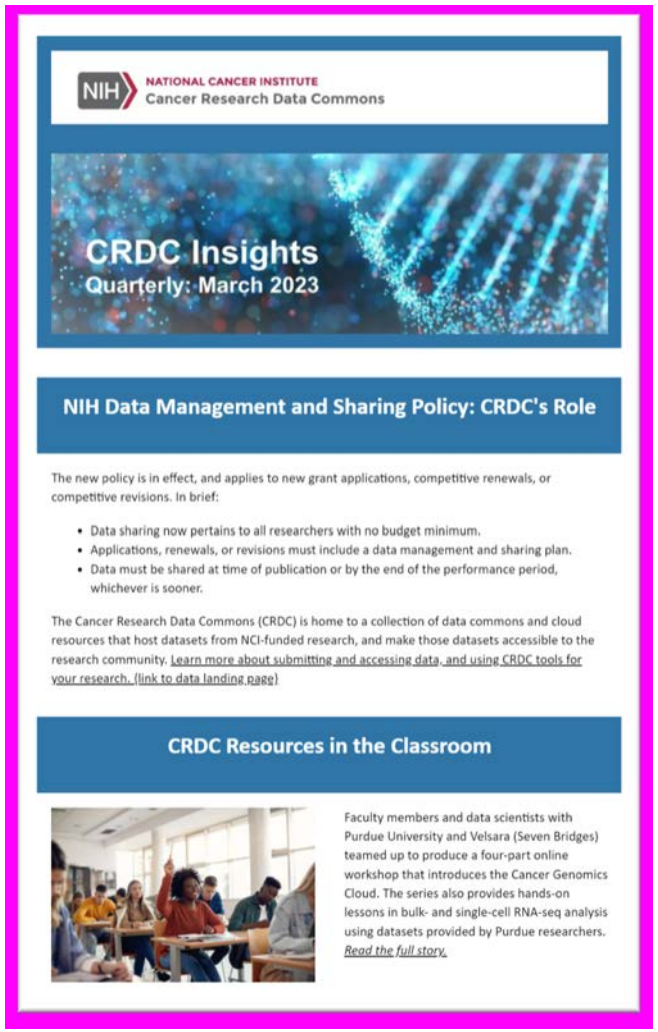
CRDC COMPONENT	RESOURCE	AVAILABLE SUPPORT
Infrastructure	Cancer Data Aggregator (CDA)	If you have questions, contact the CDA team through the CDA Helpdesk

CRDC Website:
https://datacommons.cancer.gov/



CRDC Insights: External Newsletter (quarterly)

https://datacommons.cancer.gov/crdc-insights



NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

CRDC Insights
Quarterly: March 2023

NIH Data Management and Sharing Policy: CRDC's Role

The new policy is in effect, and applies to new grant applications, competitive renewals, or competitive revisions. In brief:

- Data sharing now pertains to all researchers with no budget minimum.
- Applications, renewals, or revisions must include a data management and sharing plan.
- Data must be shared at time of publication or by the end of the performance period, whichever is sooner.

The Cancer Research Data Commons (CRDC) is home to a collection of data commons and cloud resources that host datasets from NCI-funded research, and make those datasets accessible to the research community. Learn more about submitting and accessing data, and using CRDC tools for your research. (link to data landing page)

CRDC Resources in the Classroom

Faculty members and data scientists with Purdue University and Velsara (Seven Bridges) teamed up to produce a four-part online workshop that introduces the Cancer Genomics Cloud. The series also provides hands-on lessons in bulk- and single-cell RNA-seq analysis using datasets provided by Purdue researchers. *Read the full story.*

HTAN: Methods Workshop at AACR 2023 Annual Meeting

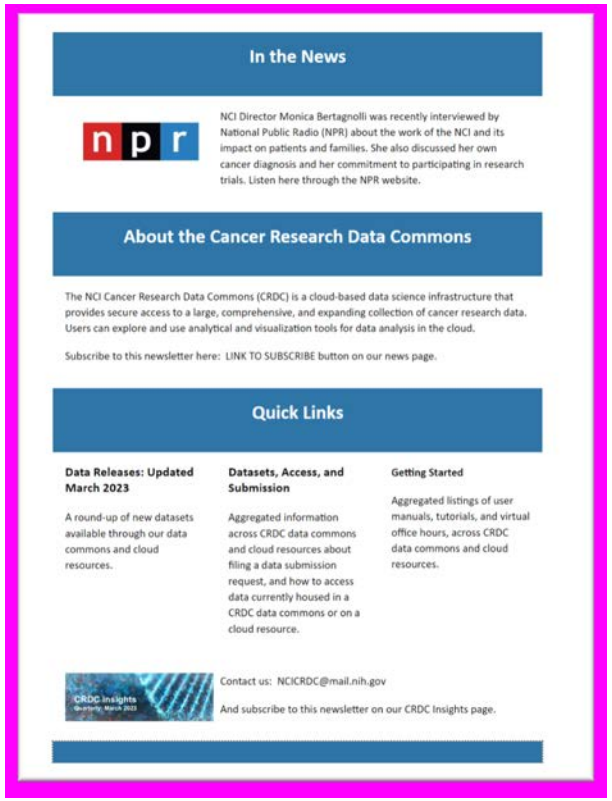
The Human Tumor Atlas Network (HTAN) is working closely with CRDC to ensure long-term legacy and reuse of HTAN data, and to share data through NCI's Cloud Resources. This methods workshop will demonstrate how to access, query, use data within the cloud environment, and visualize HTAN data derived from a variety of assay types. Read more about this workshop on the AACR Annual Meeting website.

Announcement: Funding Opportunities

The Office of XYZ has released a RFP/grant solicitation regarding data interoperability. Find more information on their Interoperability Initiative page.

CRDC: Empowering the Scientific Community to Make New Discoveries

A new infographic illustrates how CRDC supports the work of cancer researchers. This is available for use in presentations. Contact our general email box below.



In the News

npr

NCI Director Monica Bertagnolli was recently interviewed by National Public Radio (NPR) about the work of the NCI and its impact on patients and families. She also discussed her own cancer diagnosis and her commitment to participating in research trials. Listen here through the NPR website.

About the Cancer Research Data Commons

The NCI Cancer Research Data Commons (CRDC) is a cloud-based data science infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data. Users can explore and use analytical and visualization tools for data analysis in the cloud.

Subscribe to this newsletter here: LINK TO SUBSCRIBE button on our news page.

Quick Links

Data Releases: Updated March 2023

A round-up of new datasets available through our data commons and cloud resources.

Datasets, Access, and Submission

Aggregated information across CRDC data commons and cloud resources about filing a data submission request, and how to access data currently housed in a CRDC data commons or on a cloud resource.

Getting Started

Aggregated listings of user manuals, tutorials, and virtual office hours, across CRDC data commons and cloud resources.

Contact us: NCICRDC@mail.nih.gov

And subscribe to this newsletter on our CRDC Insights page.

AACR Cancer Research Series

A four-part invited series published online in March 2024 highlighting the CRDC's accomplishments from the past 10 years.

- LESSONS LEARNED AND FUTURE STATE
- RESOURCES TO SHARE KEY CANCER DATA
- CLOUD-BASED ANALYTICAL RESOURCES
- CORE STANDARDS AND SERVICES

Learn more about the series on the CRDC Website.

CRDC 2024 Fall Symposium: October 16-17, 2024

A one-and-a-half-day event highlighting the 10th Anniversary of the CRDC and its future initiatives

Save the Date

NIH MASUR AUDITORIUM, BETHESDA MD (10/16)
NCI CAMPUS, ROCKVILLE MD (10/17)

PRE-REGISTRATION REQUIRED
Register & More Information at:
DATACOMMONS.CANCER.GOV

CRDC and ODS Collaboration Session
Wednesday, October 16 @ 1:30 PM ET

- Data Sharing & Access within CRDC
- CRDC Symposium Kick-Off

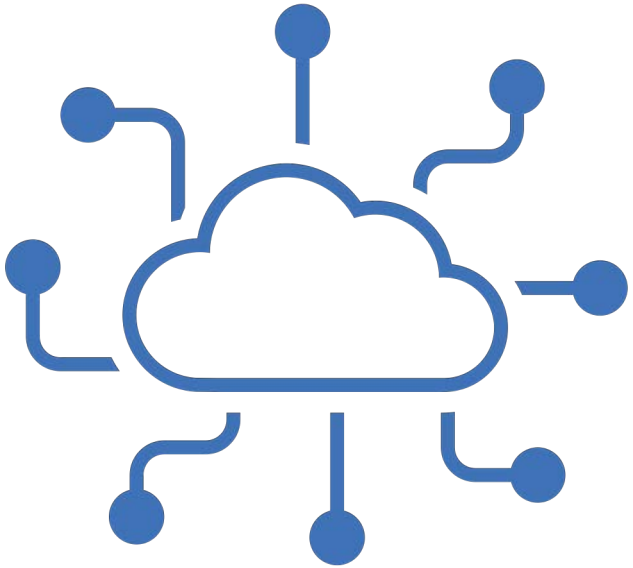
Immediately following NCI Office of Data Sharing Symposium (separate event registration)

CRDC Session
Thursday, October 17 @ 9:00 AM ET

- CRDC History & Current State
- Success Stories & Impactful Programs
- Future Spotlight
- Fireside Chat

CRDC: Interoperability Needs for Cancer Data

Challenge: Access comprehensive datasets like TCGA and CPTAC from multiple repositories for integrative analysis

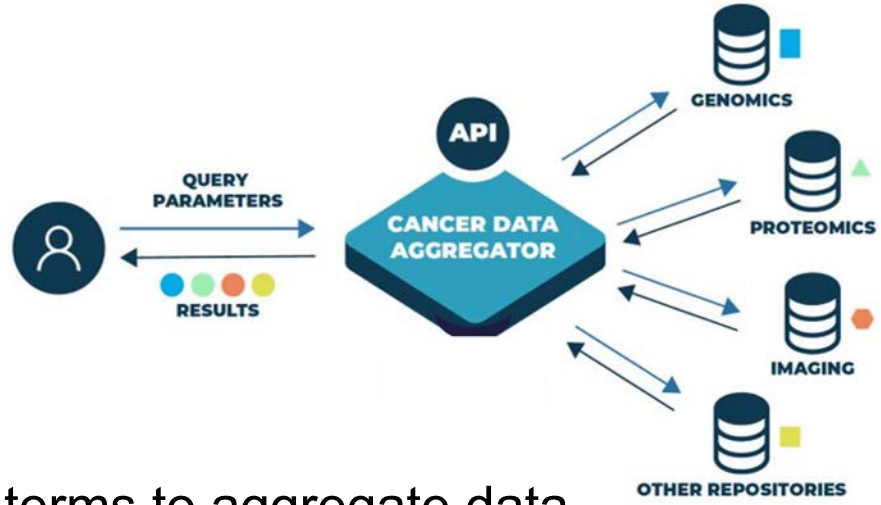


- Discover relevant datasets *across multiple resources using common standards*
- *Aggregate and analyze data* housed in separate data repositories using latest analytical tools

CRDC: Internal Interoperability Projects

CRDC Data Standards Services (DSS)

- Semantic harmonization across CRDC datasets
- Shared data models for submission & search
- Leverage existing standards, eg NCIt

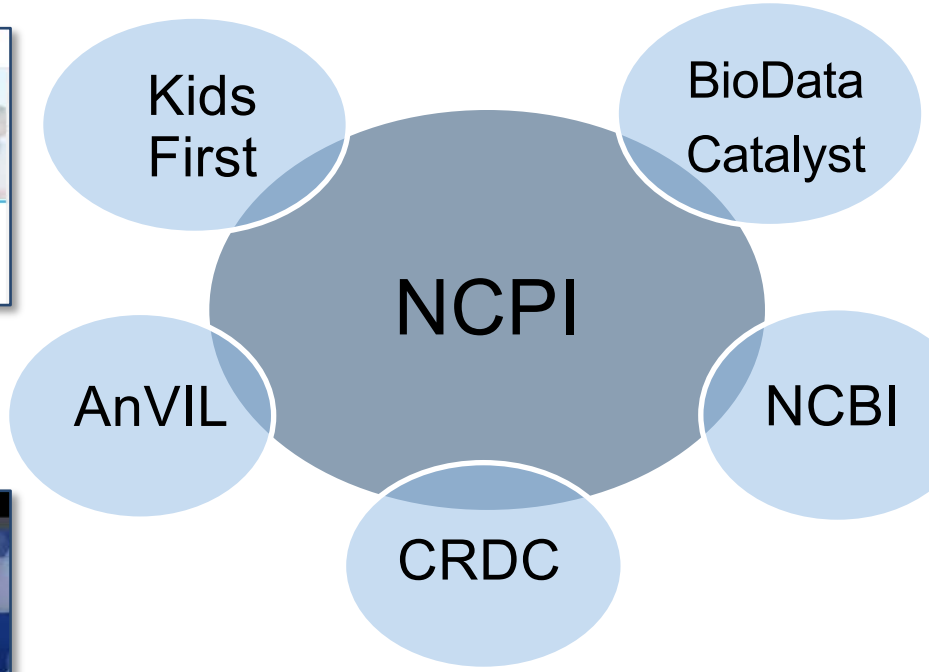


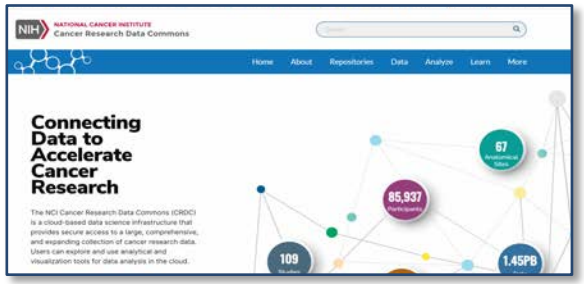
CRDC Cancer Data Aggregator (CDA)

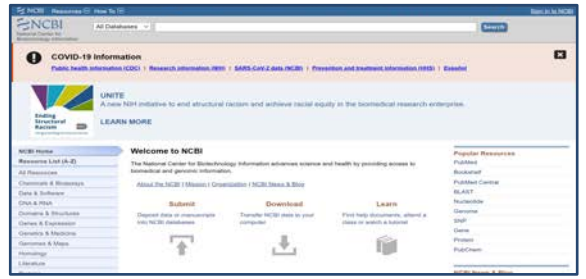
- Search metadata by harmonized, common language terms to aggregate data distributed across CRDC repositories
- Get information about subjects, files or specimens in a standard tsv format that can be opened in Excel, integrated into a pipeline or uploaded to a cloud resource
- cdapython available via interactive browser, notebooks or local install

NCPI: NIH Cloud Platform Interoperability

Connecting with a Greater Data Ecosystem







Childhood Cancer Data Initiative (CCDI) Exemplar for Building a Health Learning System

- Community of researchers, advocates, hospitals and networks committed to sharing pediatric cancer data to accelerate research on childhood cancers.
- Federated Pediatric Cancer Data Ecosystem
  - Childhood cancer data and resources from across the nation
    - Research repositories
    - Patient registries
    - Hospitals

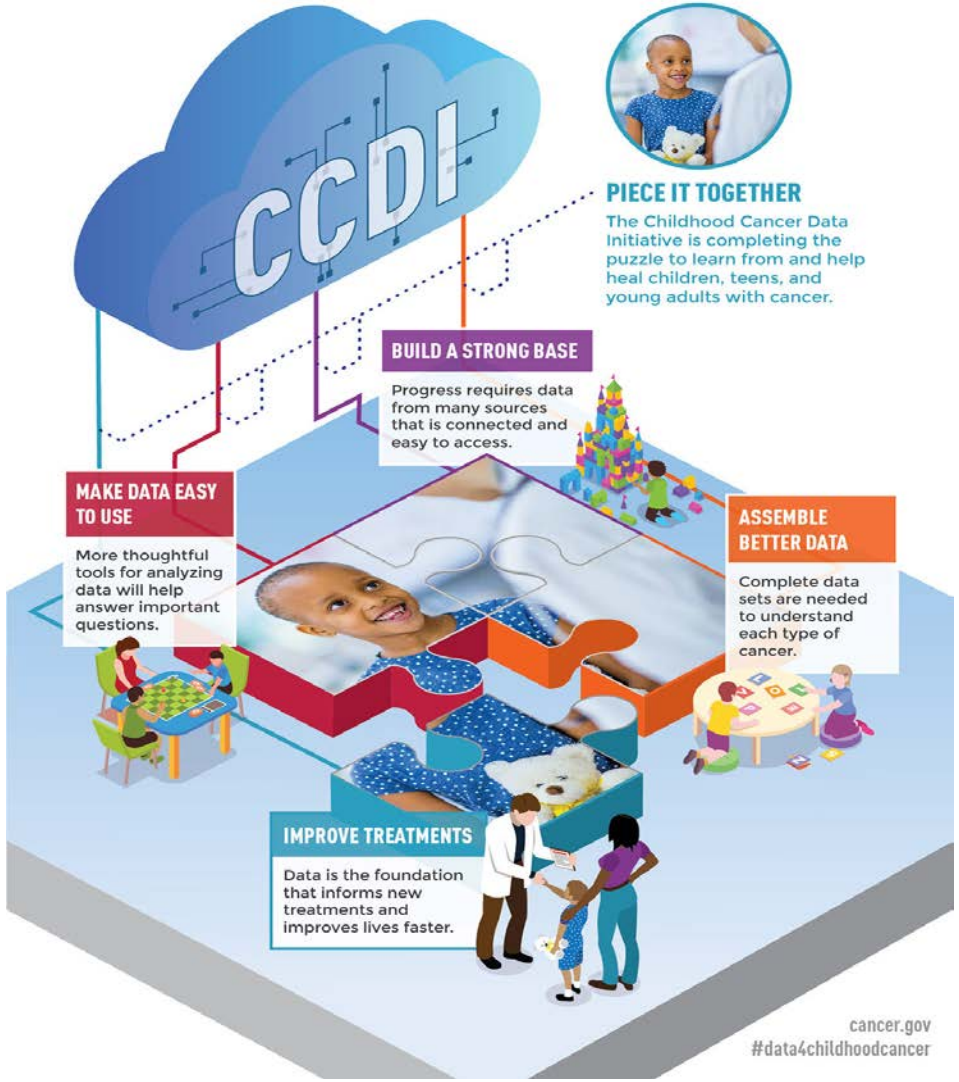
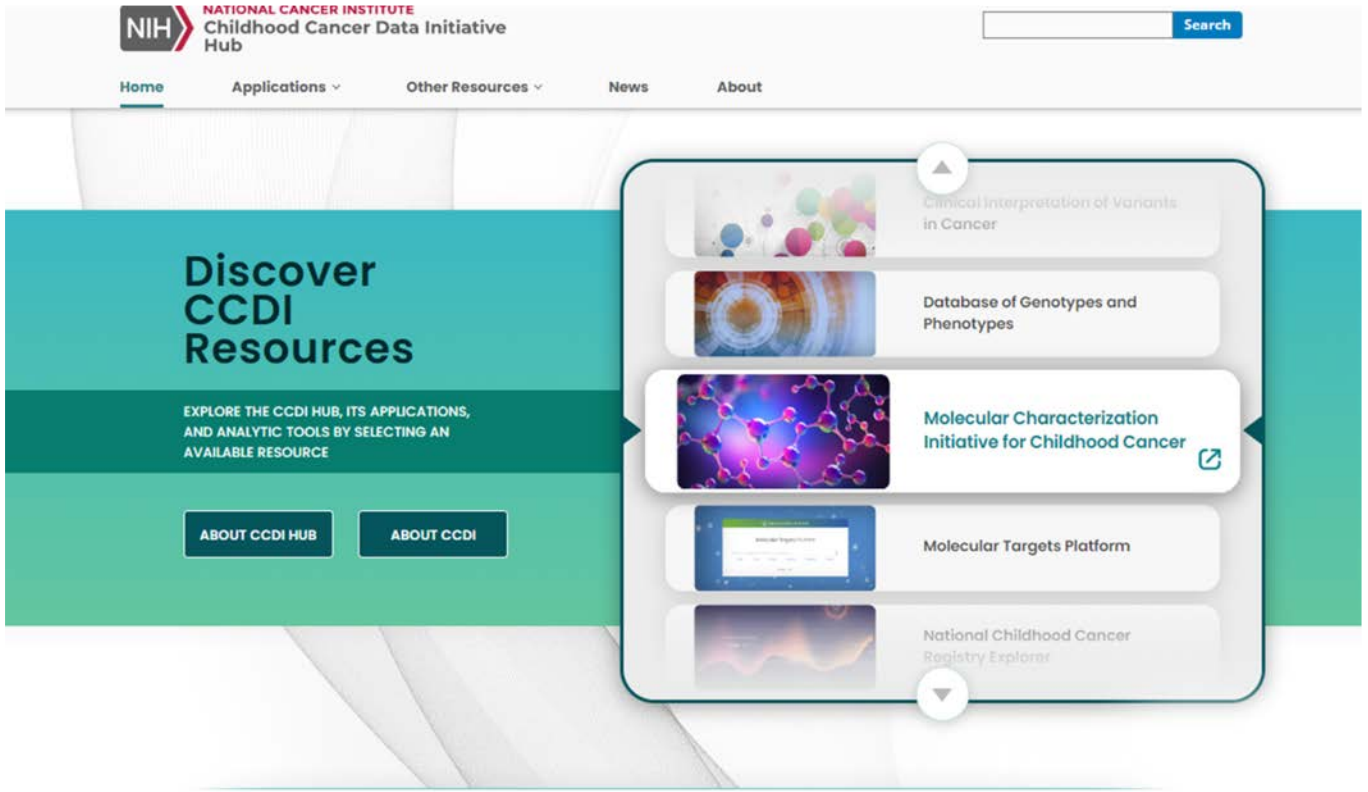
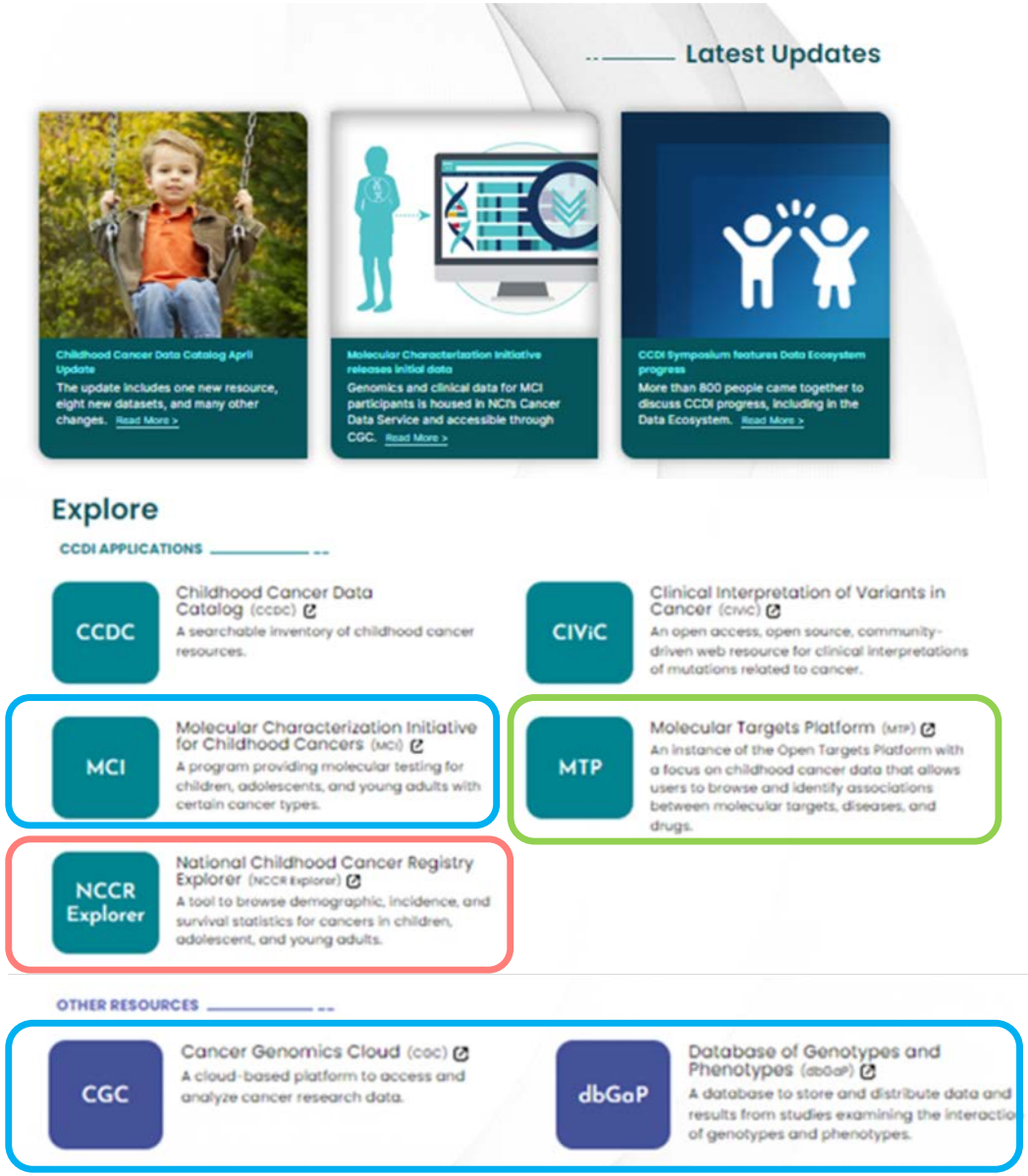


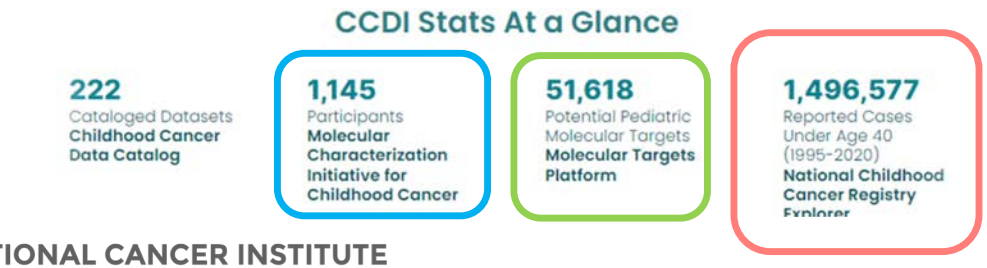
Image: Childhood Cancer Data Initiative. National Cancer Institute.
https://www.cancer.gov/research/areas/childhood/childhood-cancer-data-initiative/about

Access to CCDI Data & Tools

- CCDI Data Hub provides links to data, knowledge bases and tools (*https://ccdi.cancer.gov*)

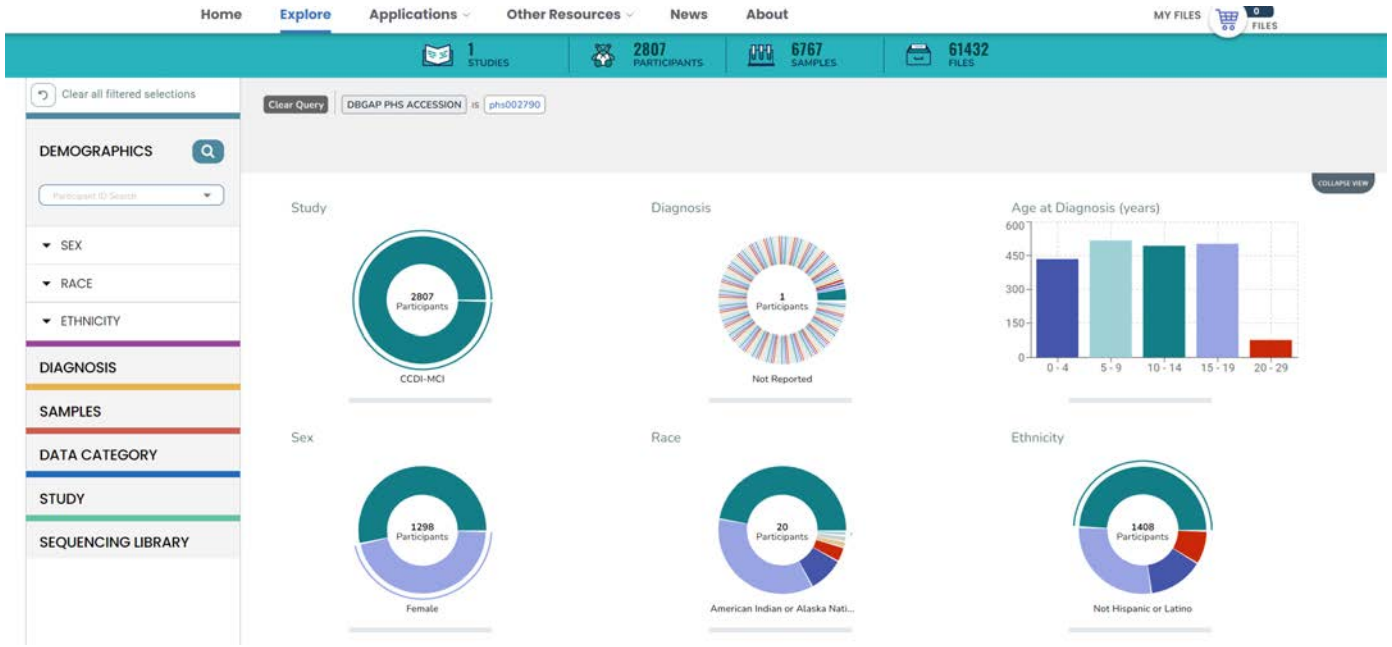






CCDI Available Datasets Via CRDC

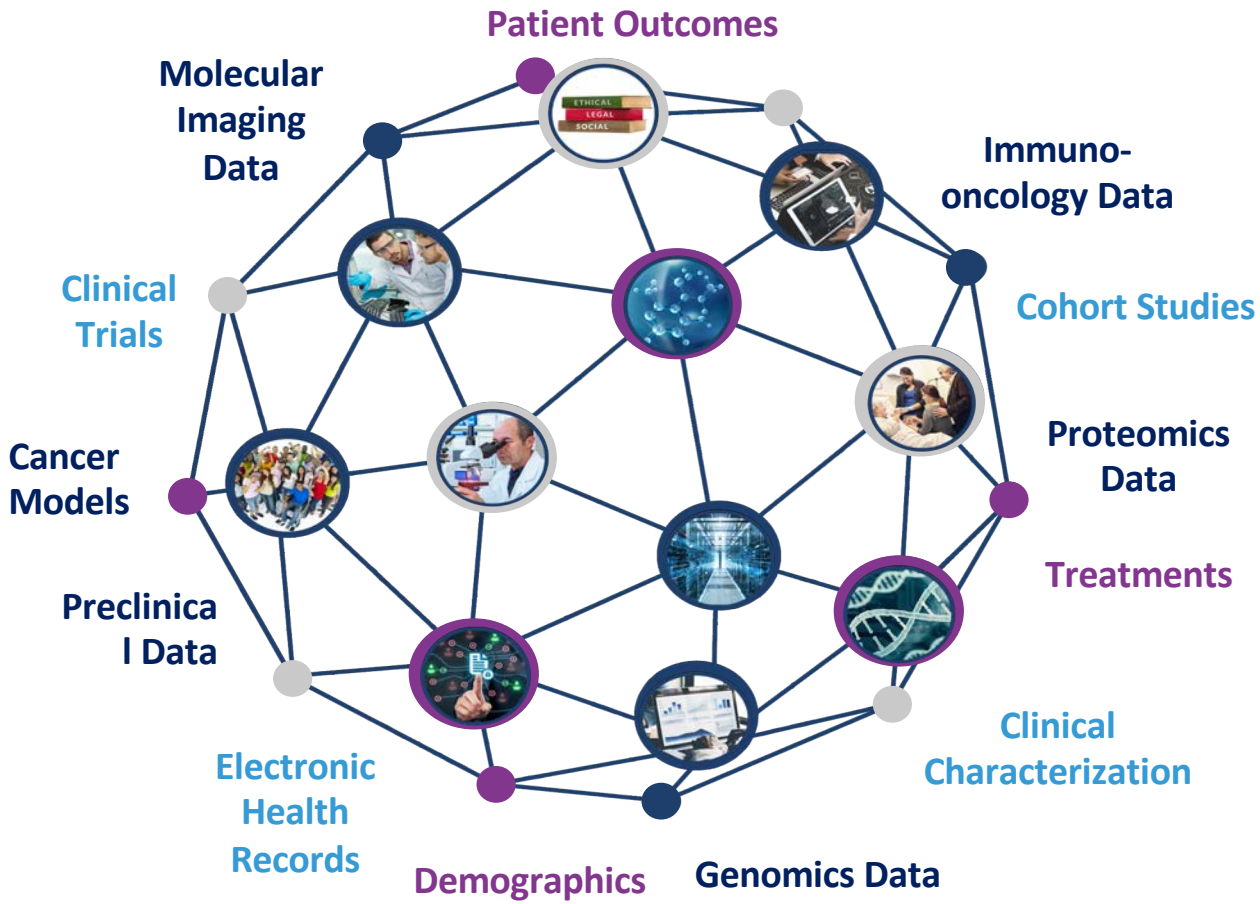
- Cancer Data Service (CDS) and Imaging Data Commons (IDC)
  - CCDI data from 1400+ participants
  - 70,000+ files
    - Genetic testing data
    - WGS and WXS
    - Full transcriptome sequencing
    - Single cell analysis
    - Imaging data



- CCDI Data Hub
  - Bento framework
    - GUI for Data Exploration
  - Authentication & Authorization controls
    - Integration of dbGaP A&A workflows for controlled data access

Establishing a *Federated* Pediatric Cancer Data Ecosystem

- Underlying data infrastructure
- Enhanced cloud-computing
- Services linking various data types (clinical, image, & molecular data)
- Standards & tools for data interoperability
- Data repositories (e.g. ***Childhood Cancer Clinical Data Commons***)
- Linked data (***National Childhood Cancer Registry***)



Discovery Science → Clinical Studies/Care → Surveillance

cancer.gov/CCDI

#data4childhoodcancer

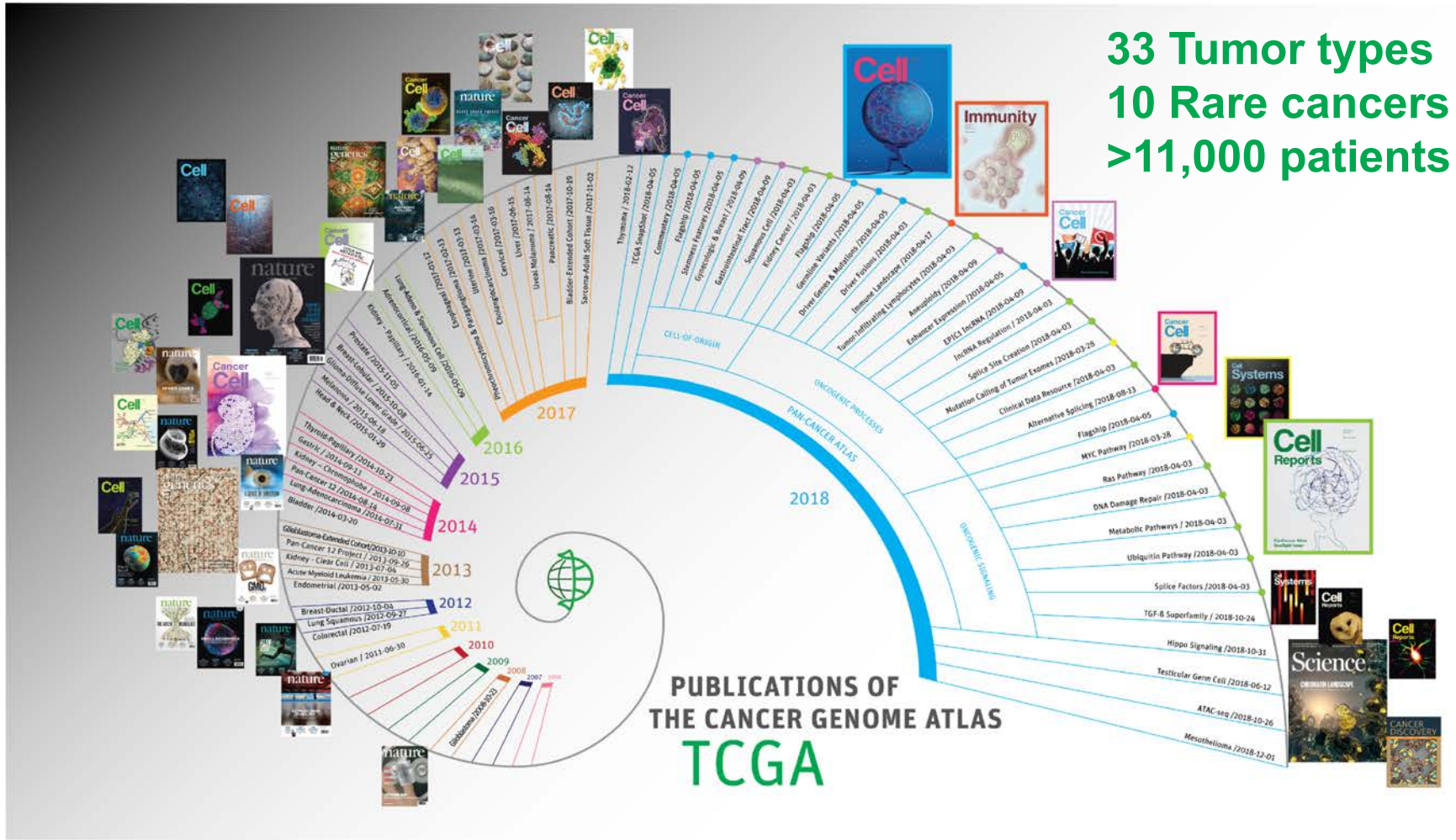
Agenda

1. Open Science -- Importance and Goals of Data Sharing
2. Cancer Research Data Commons (CRDC) and Childhood Cancer Data Initiative (CCDI)
3. **Data Drives Discovery**
4. Other “Big Data” Resources
5. Misc Items of Interest

The Cancer Genome Atlas (TCGA)

NIH NATIONAL CANCER INSTITUTE

NIH National Human Genome Research Institute



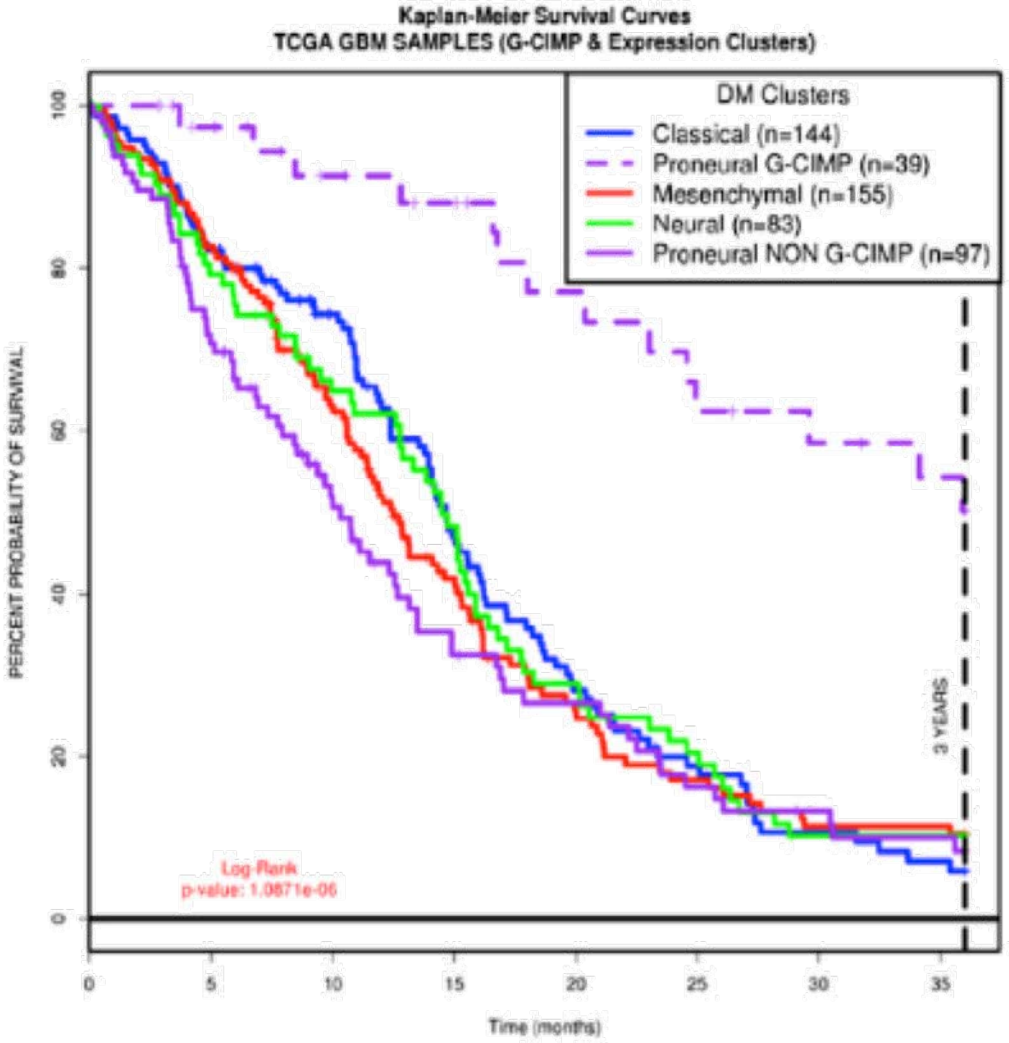
2.5 Petabytes of data

Genomic Data Commons
Proteomic Data Commons
Imaging Data Commons

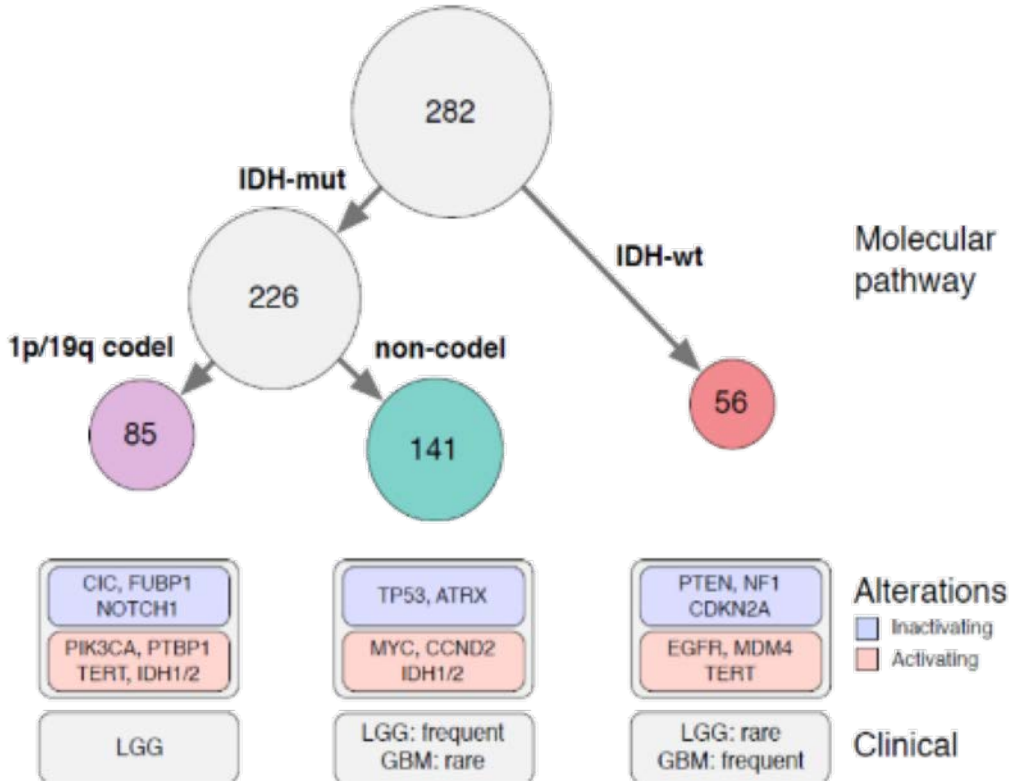
- ✓ Genomic
- ✓ Transcriptomic
- ✓ Epigenomic
- ✓ Imaging
- ✓ Clinical
- ✓ More…

NIH NATIONAL CANCER INSTITUTE

TCGA Impact on Brain Tumors



Brennan et al, Cell, 2013



Led to changes in the WHO classification and treatments for all gliomas

Brat et al, NEJM, 2015

New 2016 WHO Classification - Biomarker Expansion with 2021 WHO Classification

WHO classification of tumours of the central nervous system

Diffuse astrocytic and oligodendroglial tumours
Diffuse astrocytoma, IDH-mutant 9400/3
Gemistocytic astrocytoma, IDH-mutant 9411/3
Diffuse astrocytoma, IDH-wildtype *9400/3*
Diffuse astrocytoma, NOS 9400/3

Anaplastic astrocytoma, IDH-mutant 9401/3
Anaplastic astrocytoma, IDH-wildtype *9401/3*
Anaplastic astrocytoma, NOS 9401/3

Glioblastoma, IDH-wildtype 9440/3
Giant cell glioblastoma 9441/3
Gliosarcoma 9442/3
Epithelioid glioblastoma *9440/3*
Glioblastoma, IDH-mutant 9445/3*
Glioblastoma, NOS 9440/3

Diffuse midline glioma, H3 K27M–mutant 9385/3*

Oligodendroglioma, IDH-mutant and 1p/19q-codeleted 9450/3
Oligodendroglioma, NOS 9450/3

Anaplastic oligodendroglioma, IDH-mutant and 1p/19q-codeleted 9451/3
Anaplastic oligodendroglioma, NOS *9451/3*

Oligoastrocytoma, NOS *9382/3*
Anaplastic oligoastrocytoma, NOS *9382/3*

Other astrocytic tumours
Pilocytic astrocytoma 9421/1
Pilomyxoid astrocytoma 9425/3
Subependymal giant cell astrocytoma 9384/1
Pleomorphic xanthoastrocytoma 9424/3
Anaplastic pleomorphic xanthoastrocytoma 9424/3

Ependymal tumours
Subependymoma 9383/1
Myxopapillary ependymoma 9394/1
Ependymoma 9391/3
Papillary ependymoma 9393/3
Clear cell ependymoma 9391/3
Tanycytic ependymoma 9391/3
Ependymoma, *RELA* fusion–positive 9396/3*
Anaplastic ependymoma 9392/3

Other gliomas
Chordoid glioma of the third ventricle 9444/1
Angiocentric glioma 9431/1
Astroblastoma 9430/3

Choroid plexus tumours
Choroid plexus papilloma 9390/0
Atypical choroid plexus papilloma 9390/1
Choroid plexus carcinoma 9390/3

Neuronal and mixed neuronal-glial tumours
Dysembryoplastic neuroepithelial tumour 9413/0
Gangliocytoma 9492/0
Ganglioglioma 9505/1
Anaplastic ganglioglioma 9505/3
Dysplastic cerebellar gangliocytoma (Lhermitte–Duclos disease) 9493/0
Desmoplastic infantile astrocytoma and ganglioglioma 9412/1
Papillary glioneuronal tumour 9509/1
Rosette-forming glioneuronal tumour 9509/1
Diffuse leptomeningeal glioneuronal tumour
Central neurocytoma 9506/1
Extraventricular neurocytoma 9506/1
Cerebellar liponeurocytoma 9506/1
Paraganglioma 8693/1

Tumours of the pineal region
Pineocytoma 9361/1
Pineal parenchymal tumour of intermediate differentiation 9362/3
Pineoblastoma 9362/3
Papillary tumour of the pineal region 9395/3

Embryonal tumours
Medulloblastomas, genetically defined
Medulloblastoma, WNT-activated 9475/3*
Medulloblastoma, SHH-activated and *TP53*-mutant 9476/3*
Medulloblastoma, SHH-activated and *TP53*-wildtype 9471/3
Medulloblastoma, non-WNT/non-SHH 9477/3*
Medulloblastoma, group 3
Medulloblastoma, group 4
Medulloblastomas, histologically defined
Medulloblastoma, classic 9470/3
Medulloblastoma, desmoplastic/nodular 9471/3
Medulloblastoma with extensive nodularity 9471/3
Medulloblastoma, large cell / anaplastic 9474/3
Medulloblastoma, NOS 9470/3

Embryonal tumour with multilayered rosettes, C19MC-altered 9478/3*
Embryonal tumour with multilayered rosettes, NOS *9478/3*
Medulloepithelioma 9501/3
CNS neuroblastoma 9500/3
CNS ganglioneuroblastoma 9490/3
CNS embryonal tumour, NOS 9473/3
Atypical teratoid/rhabdoid tumour 9508/3
CNS embryonal tumour with rhabdoid features *9508/3*

Tumours of the cranial and paraspinal nerves
Schwannoma 9560/0
Cellular schwannoma 9560/0
Plexiform schwannoma 9560/0

Table 2 Key Diagnostic Genes, Molecules, Pathways, and/or Combinations in Major Primary CNS Tumors

Tumor Type	Genes/Molecular Profiles Characteristically Altered[a]
Astrocytoma, IDH-mutant	*IDH1, IDH2, ATRX, TP53, CDKN2A/B*
Oligodendroglioma, IDH-mutant, and 1p/19q-codeleted	*IDH1, IDH2*, 1p/19q, *TERT* promoter, *CIC, FUBP1, NOTCH1*
Glioblastoma, IDH-wildtype	IDH-wildtype, *TERT* promoter, chromosomes 7/10, *EGFR*
Diffuse astrocytoma, *MYB*- or *MYBL1*-altered	*MYB, MYBL1*
Angiocentric glioma	*MYB*
Polymorphous low-grade neuroepithelial tumor of the young	*BRAF*, FGFR family
Diffuse low-grade glioma, MAPK pathway-altered	*FGFR1, BRAF*
Diffuse midline glioma, H3 K27-altered	H3 K27, *TP53, ACVR1, PDGFRA, EGFR, EZHIP*
Diffuse hemispheric glioma, H3 G34-mutant	H3 G34, *TP53, ATRX*
Diffuse pediatric-type high-grade glioma, H3-wildtype, and IDH-wildtype	IDH-wildtype, H3-wildtype, *PDGFRA, MYCN, EGFR* (methylome)
Infant-type hemispheric glioma	NTRK family, *ALK, ROS, MET*
Pilocytic astrocytoma	*KIAA1549-BRAF, BRAF, NF1*
High-grade astrocytoma with piloid features	*BRAF, NF1, ATRX, CDKN2A/B* (methylome)
Pleomorphic xanthoastrocytoma	*BRAF, CDKN2A/B*
Subependymal giant cell astrocytoma	*TSC1, TSC2*
Chordoid glioma	*PRKCA*
Astroblastoma, *MN1*-altered	*MN1*

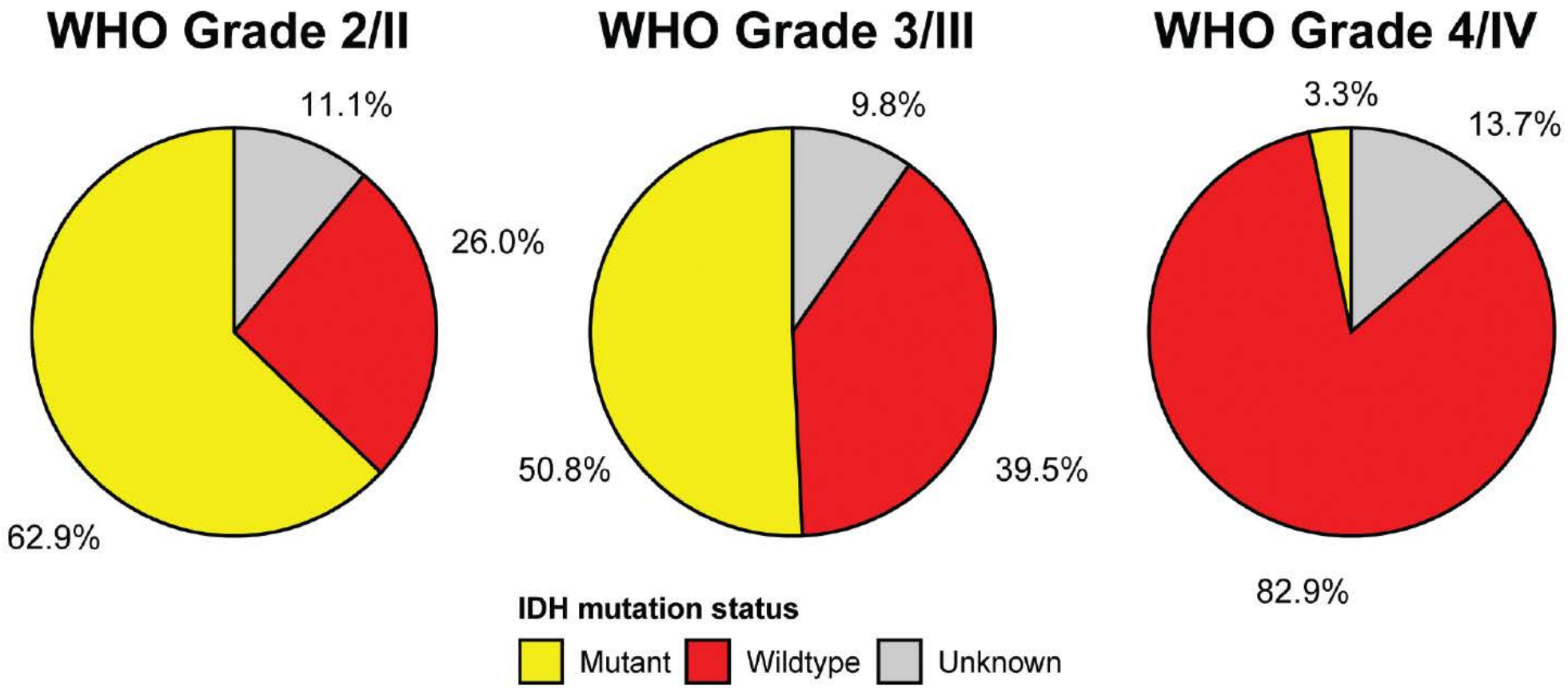
Louis, et al., Acta Neuropathol, 2016
Louis, et al., Neuro Oncology, 2021

Collecting Molecular Marker Data on Brain Tumors in the US WHO 2016 Alignment: New Proposed Adjustments for WHO 2021

Code	Description
01	Diffuse astrocytoma, IDH-mutant (9400/3)
02	Diffuse astrocytoma, IDH-wildtype (9400/3)
03	Anaplastic astrocytoma, IDH-mutant (9401/3)
04	Anaplastic astrocytoma, IDH-wildtype (9401/3)
05	Glioblastoma, IDH-wildtype (9440/3)
06	Oligodendroglioma, IDH-mutant and 1 p/19q co-deleted (9450/3)
07	Anaplastic oligodendroglioma, IDH-mutant and 1p/19q co-deleted (9451/3)
08	Medulloblastoma, SHH-activated and TP53-wildtype (9471/3)
09	Embryonal tumor with multilayered rosettes, C19MC-altered (9478/3)
85	Not applicable: Histology not 9400/3, 9401/3, 9440/3, 9450/3, 9451/3, 9471/3, 9478/3
86	Benign or borderline tumor
87	Test ordered, results not in chart
88	Not applicable: Information not collected for this case;
99	Not documented in patient record; No microscopic confirmation; Brain molecular markers not assessed or unknown if assessed

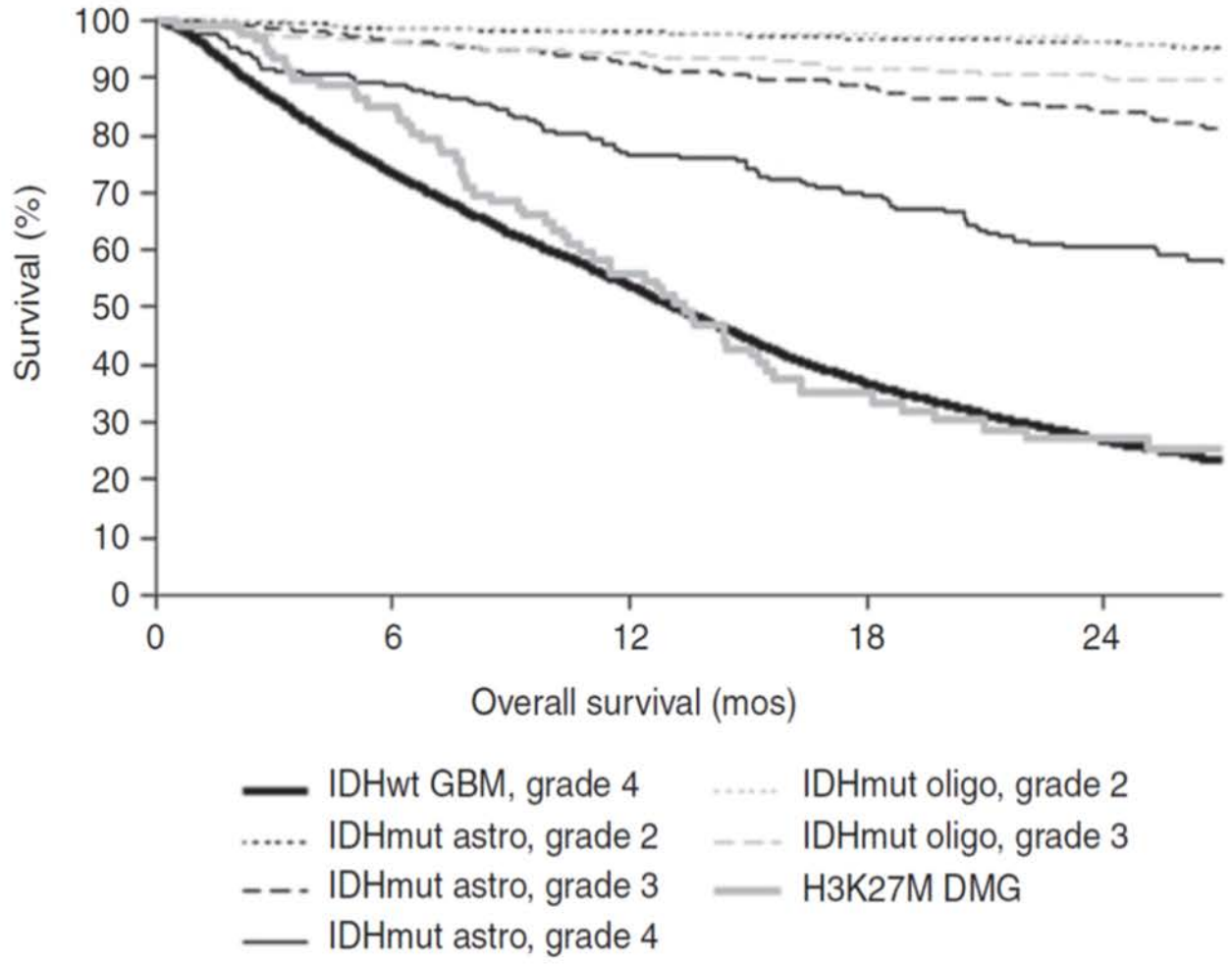
https://apps.naaccr.org/ssdi/naaccr/item/3816/1.7

Molecular Markers for Brain Tumors at the Population-Level



Ostrom, et al. Neuro Oncology, 2023

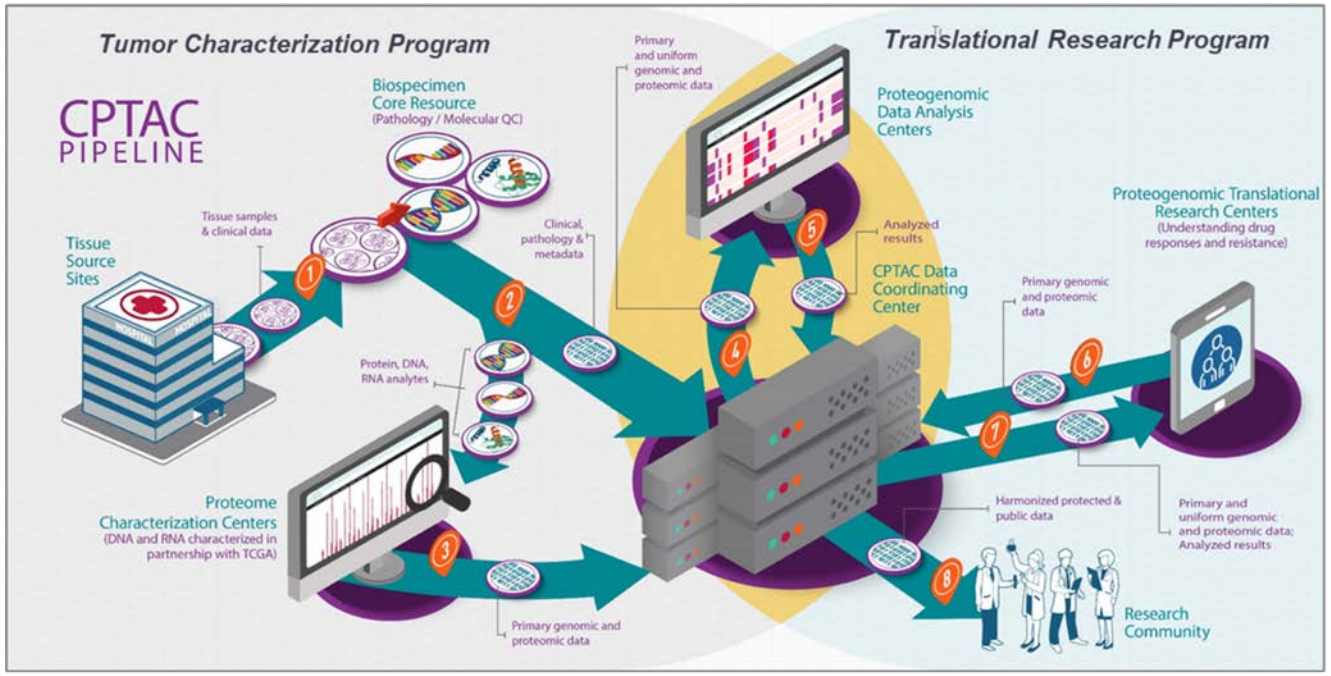
Overall Survival by Molecularly Defined Glioma Type



Ostrom, et al. Neuro Oncology, 2022

Clinical Proteomics Tumor Analysis Consortium (CPTAC)

multidisciplinary (team science)



Goals

1. Accelerate understanding of cancer biology (genotype-to-phenotype) by **comprehensively characterizing tumors** (proteomics and genomics).
2. Produce **public resources** (molecular and clinical data, assays, images, informatic tools) that fuel hypothesis-driven science.
3. Support **clinically relevant research projects** that address mechanisms of treatment response, resistance, or toxicity.

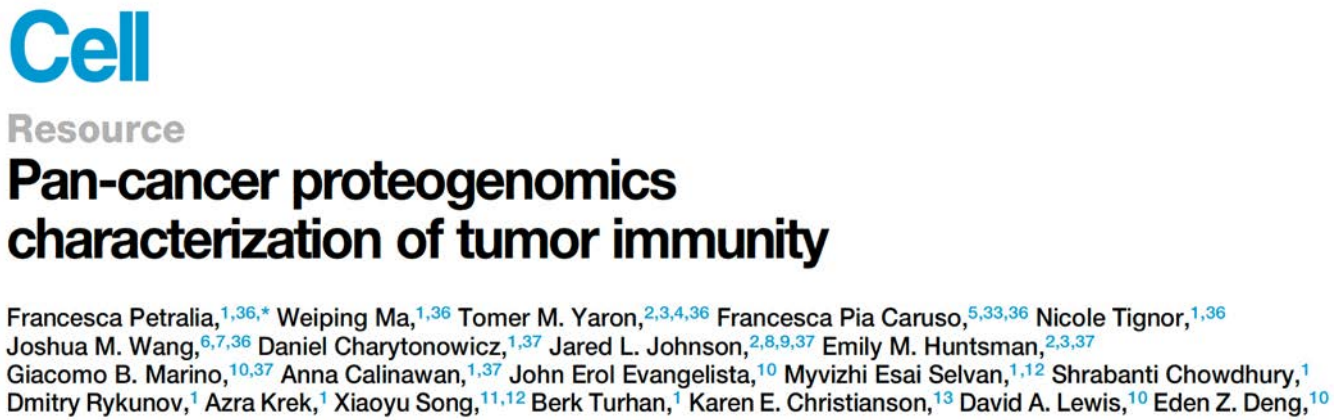
Leveraging CRDC to Decipher the Pan-Cancer Immune Landscape

Problem:

- Immunotherapy only successful in a small proportion of cancer cases

Goal:

- Develop comprehensive understanding TME across cancers
- Reveal immune cell surveillance and tumor immune evasion mechanisms



Petralia, F., et al. (2024) Cell, 187:1255-1277, 2024

Leveraging CRDC to Decipher the Pan-Cancer Immune Landscape

Approach:

- Combine CPTAC data from across CRDC
- Analyze genomic, epigenetic, transcriptomic, and proteomic alterations across tumors
- 1,056 tumor samples,10 cancers
  - Classify tumors into immune subtypes
  - Correlate with clinical outcomes

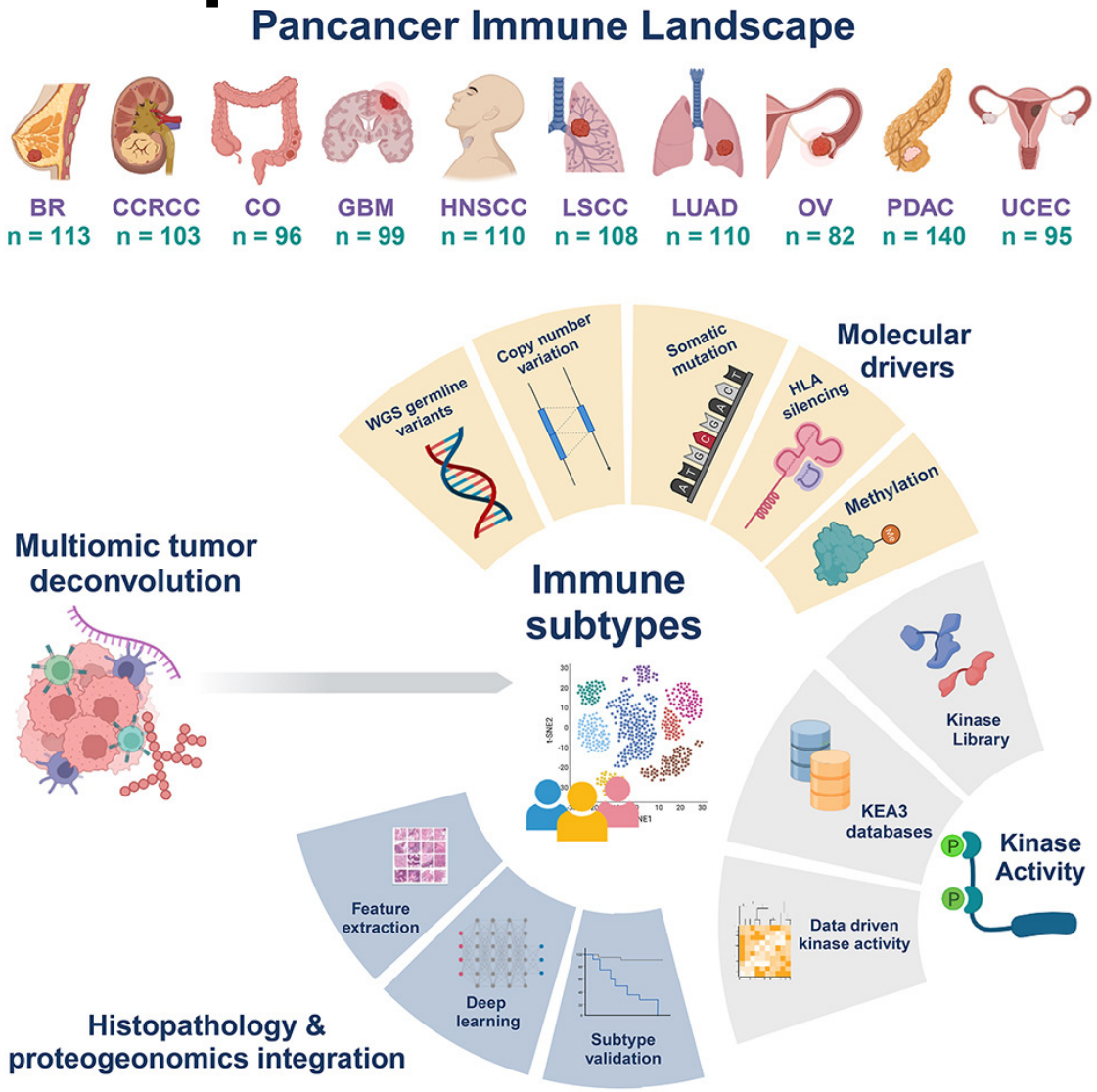


Image: Pan-cancer Immune Landscape Infographic. Derived from Petralia, F., et al., Cell 187, 2024

Leveraging CRDC to Decipher the Pan-Cancer Immune Landscape

Key findings:

- 7 distinct immune subtypes
  - Common immune reactions, evasion mechanisms independent of cancer type
- Correlations between PFS and immune subtypes, TME immune cell load
- Specific kinases activated in subtypes
  - Immune evasion, pathogenesis, and host immunity

Impact:

- Multi-dimensional view of tumor biology
- Novel patient stratification, therapeutic strategies
- New interactive web portals: PhosNet Vis, ProKap
  - Leverage PDC's CPTAC pan cancer kinase and transcription factor activity score data to explore relationships with immune subtypes
  - New avenues for research and target discovery

Agenda

1. Open Science -- Importance and Goals of Data Sharing
2. Cancer Research Data Commons (CRDC) and Childhood Cancer Data Initiative (CCDI)
3. Data Drives Discovery
4. **Other “Big Data” Resources**
5. Misc Items of Interest

How To Choose a Dataset for Your Research

- Have a clear research question
  - Perform a literature review
- Identify which data elements and clinical outcomes you want to study
- Understand the target population coverage for the dataset you choose
- Perform a feasibility analysis to ensure you can robustly test your hypothesis
- Write a detailed analysis plan
- Organize and present results clearly -- Figures vs Tables
- Be clear on the limitations and strengths
  - Data quality and data limitations

There is No Perfect Dataset!

- Registry data
- Administrative claims
- Electronic health record (EHR)/ Electronic medical record (EMR)
- Health care data aggregators
- Networks/Companies with clinical and genomic cancer data

Strengths and Limitations – Registry Data

Strengths

- Large sample size for the population-based registries – rare cancers!
- Longitudinal for trends analysis
- Easy to use and analyze

Limitations

- Limited data elements available, especially outcomes, treatment, genomics
- Delays in reporting

Adapted from Penberthy et al, CA Can J Clin, 2022

Comparing the Cancer Registry Based Datasets

Data Characteristic	NPCR -- USCS	SEER	NCDB
Publicly Available?	YES	YES	Yes, but with approval from your institution
Coverage of the US	100%	SEER 18 ~35% -- Current ~50%	~70%
Basic individual level demographics	YES	YES	YES
Treatment information	YES, surgery only	YES, surgery and radiation only – special request to obtain chemo information	YES, surgery, radiation and chemo
Outcomes	NO, separate request to obtain access to the survival file	Overall survival	Overall survival 30-day and 90-day mortality

NPCR/USCS – https://www.cdc.gov/cancer/uscs/public-use/;
SEER – https://seer.cancer.gov/data/access.html; NCDB -- https://www.facs.org/quality-programs/cancer/ncdb/puf

Strengths and Limitations – Administrative Claims

Strengths

- Medicare/Medicaid
- State and National levels
- Focus on enrollment, demographics, dates of service, diagnosis and procedure codes, vital status, pharmacy
- Clear definition of population
- Longitudinal data
- Sometimes linked with other types of data

Limitations

- Collected for billing purposes not research
- Knowledge of coding schemes is needed
- Preferences for provider and patient not included

Adapted from Penberthy et al, CA Can J Clin, 2022

SEER-Medicare

Linkage data between cancer registry (SEER) and insurance claims data (Medicare). Individuals have unique identifiers

SEER:

- US population data
- Cancer incidence, death
- Tumor information
- Demographics

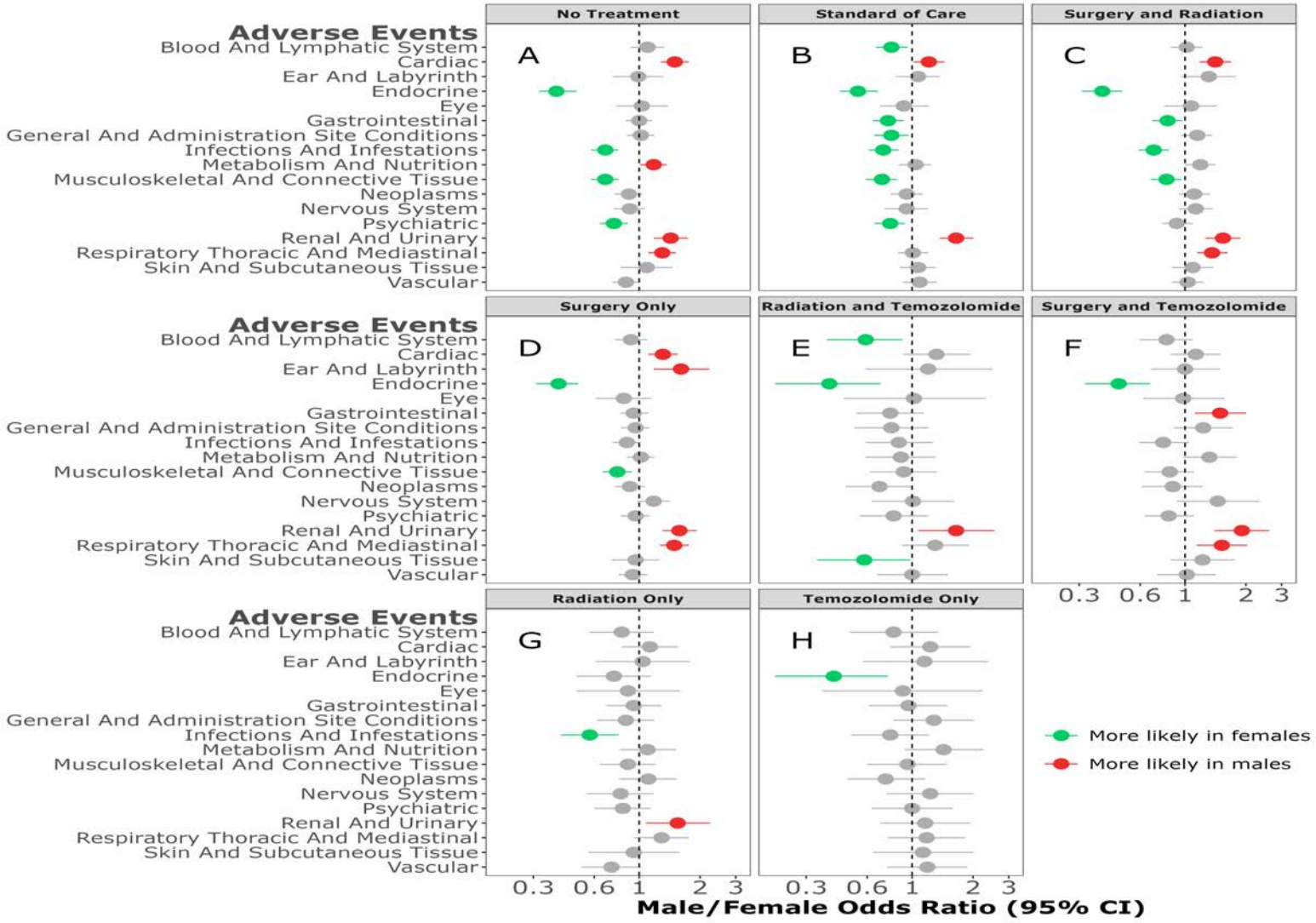
Medicare:

- Insurance claims data (Diagnosis, procedures hospitalization, prescription drugs)
- >90% of 65+ year-olds in US
- Enrollment data
- Data derived from bills submitted by providers and processed by Medicare

https://healthcaredelivery.cancer.gov/seermedicare/

- Formal proposal approval required
- SEER-Medicare approval of manuscripts before submission required
- Analysis developed and performed by research team

Adverse Events in Individuals >= 66 Years of Age with Glioblastoma



Dmukauskas et al; JNO 2024

*adverse coding per CTCAE guidelines

Strengths and Limitations – Electronic Health Record (EHR) Data

Strengths

- Clear definition of population
- Longitudinal data
- More detailed picture of patient care journey with detailed clinical data

Limitations

- System specific so may not be generalizable
- Care received outside the system not captured
- Limited genomic data availability

Adapted from Penberthy et al, CA Can J Clin, 2022

Flatiron

NEW!

Real-World Evidence Services
Generate evidence and insights by partnering with Flatiron's team of experts

Enhanced Datamart
Data subscriptions that cover 22 tumor types

Custom Data
Tailored data curated to answer client-specific research questions

Clinico-Genomic Database
Integrated EHR-derived clinical data with genomic profiling data

Imaging-Linked EHR Data
Link real-world imaging data with Flatiron's EHR-derived data

Claims-Linked EHR Data
Link real-world claims data with Flatiron's EHR-derived data

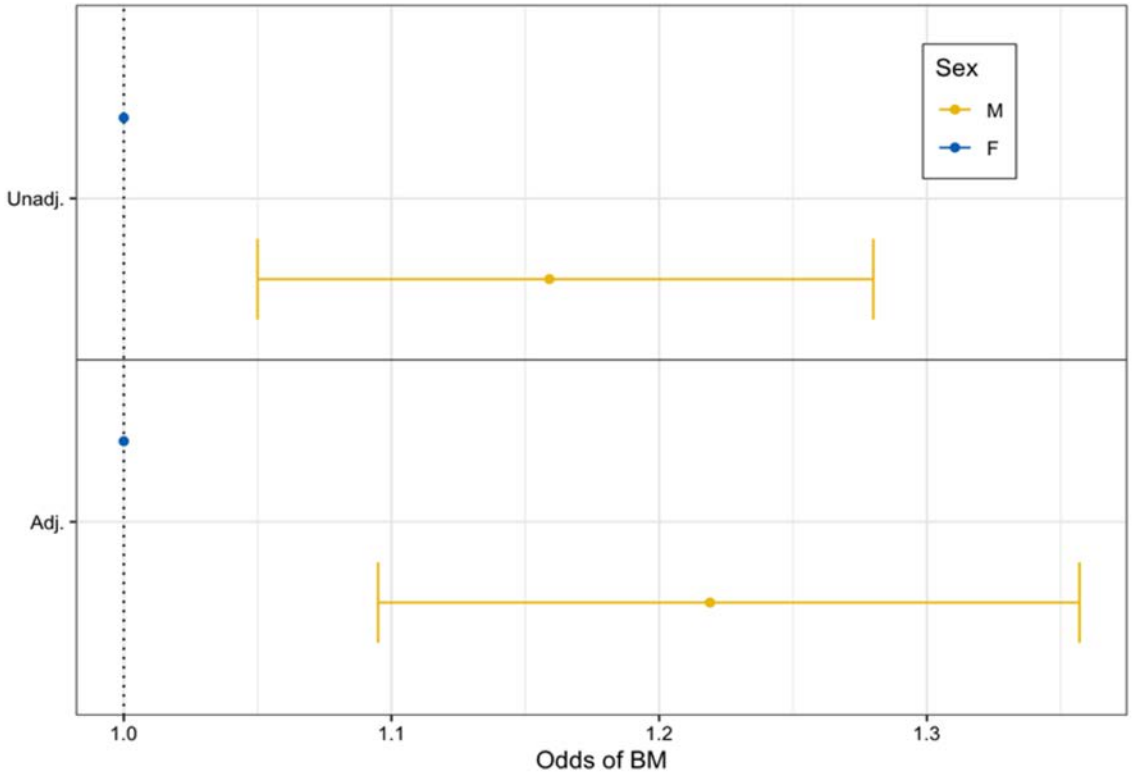
- No access to data - analysis run by internal research staff within Flatiron
- PI's team develops analytic plan and requests analysis provide

OncoEMR

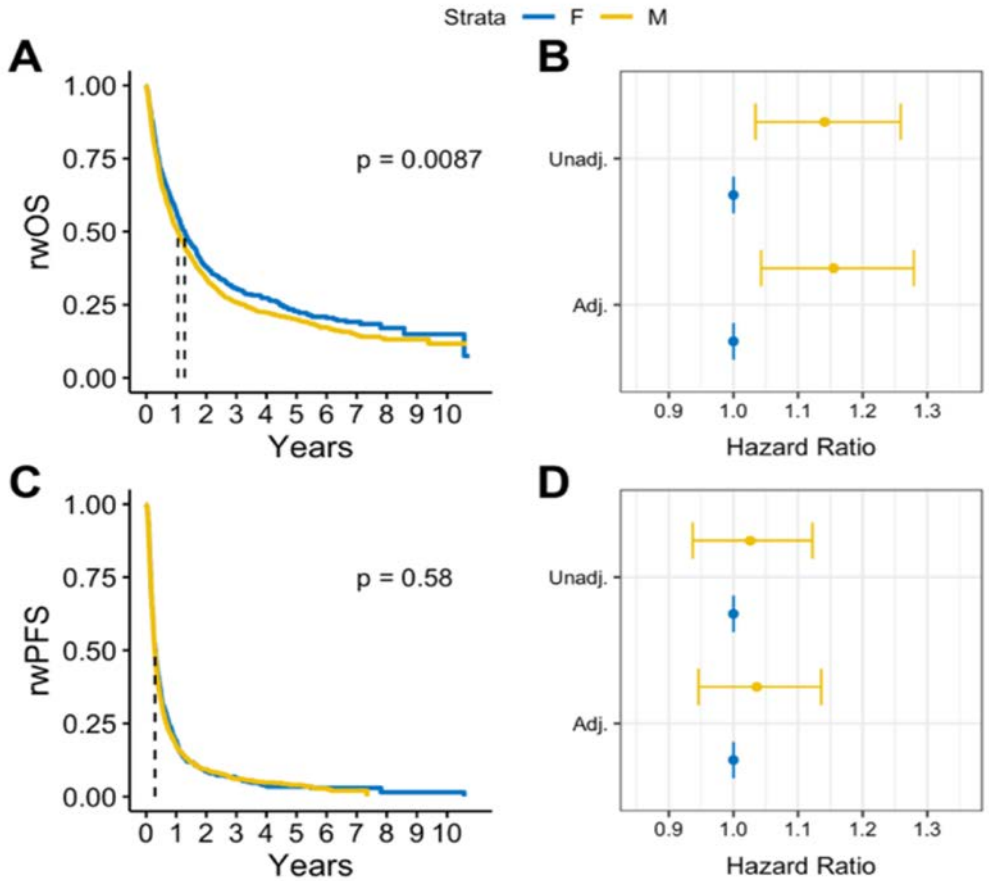
Integrated tools and content

- **Embedded clinical content**, including AJCC staging content, NCCN Templates® and First Databank drug information and patient education resources
- **Integrated clinical decision support** for treatment selection with Flatiron Assist™ backed by NCCN® guidelines, and Appropriate Use Criteria with National Decision Support Company
- **Genomic Profiling Integration with Foundation Medicine** to place, track and view results of comprehensive genomic profiling orders directly within OncoEMR®
- **State Prescription Drug Monitoring Program (PDMP) integration** for ease of e-prescribing narcotics and compliance with state and federal requirements
- **1,800+ live interfaces with 150+ unique healthcare vendors and products**, including the leading inventory management providers (Nucleus, LynxMobile and Cardinal) to seamlessly capture drug charges and waste for billing
- **ONC 2015 certified** with integrated solutions for OCM and MIPS reporting
- **Fully cloud-based solution** with secure access 24/7, no hardware costs or upgrade fees

Sex Differences in Outcomes of Brain Metastasis After Advanced Melanoma Diagnosis (Dataset: Flatiron)

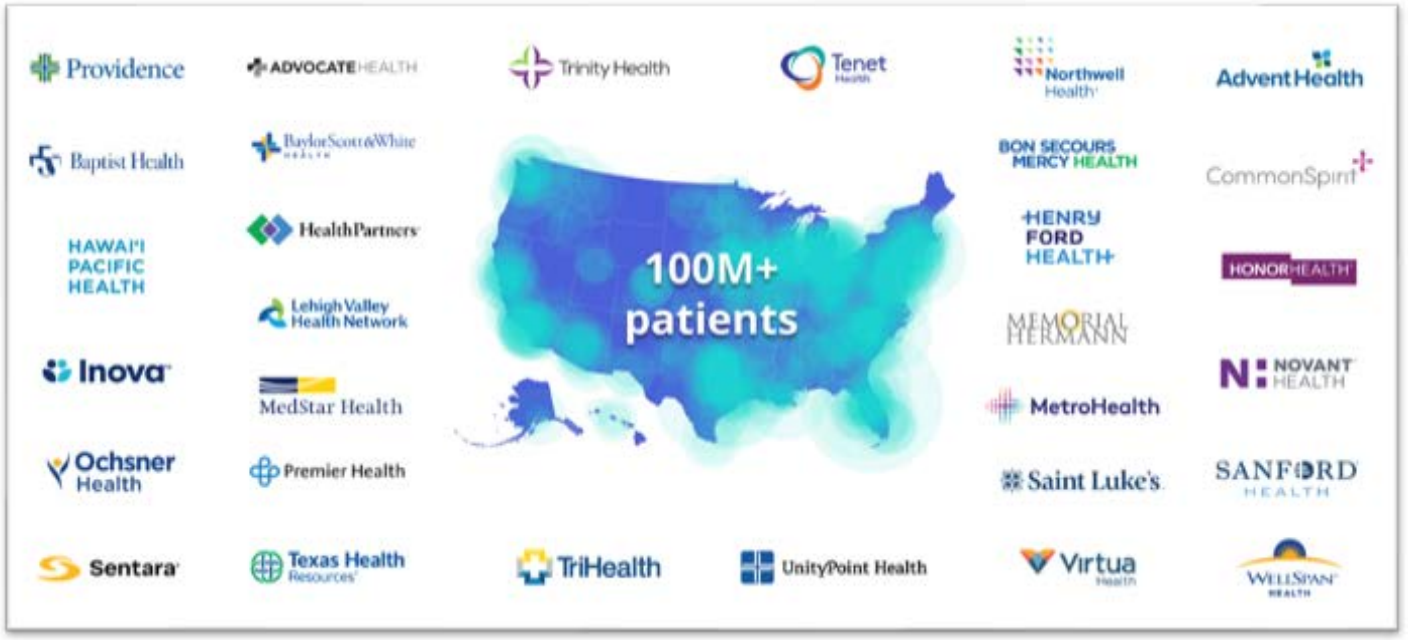


Males have higher odds of developing a brain metastases compared to females



Males have lower hazards of overall survival

Cioffi et al, Cancers 2024



Semi-structured data

- Diagnosis (SNOMED, ICD)
- Procedure (CPT,HCPCS)
- Medication (RXNORM,NDC)
- Labs (LOINC,UCUM)
- Immunizations (CVX)
- Implanted/Explanted Device (UDI)
- In-patient / out-patient care settings
- Pharmacy
- Mother-child data

Unstructured data

- Notes
- Medical images
- Genomics

Linked data

- SDOH
- Mortality
- Claims

5B+	85M+	1M+	8+	72M	188K+
Clinical notes	Imaging studies	Mother-child pairs	Years longitudinal patient history	Patients with SDOH attributes	Unique devices
MORE DETAILS	MORE DETAILS	MORE DETAILS	MORE DETAILS	MORE DETAILS	MORE DETAILS

TRUVETA

Saving Lives with Data

- No formal proposal required
- Data snapshots provide up-to-date data
- Access to data via the Truveta notebook type platform
- Data analysis using R or Python
- High level technical support from Truveta computer scientists
- Linkage with Lexis-Nexis Social determinants of health (SDOH).

NIH NATIONAL CANCER INSTITUTE

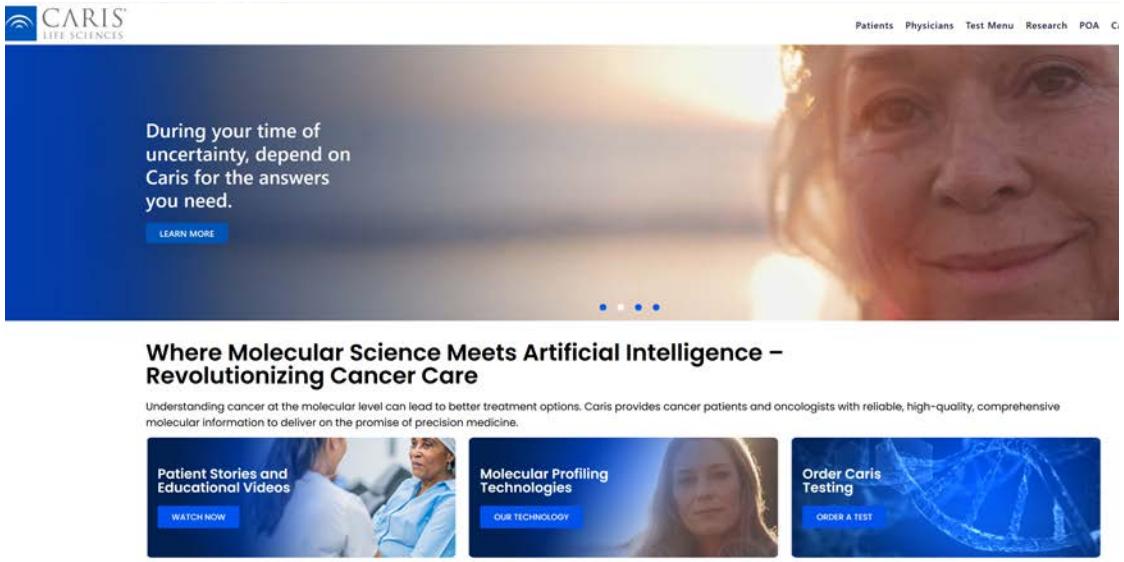
Strengths and Limitations

Networks/Companies with Clinical and Genomic Cancer Data

- **Strengths**
- Access to large amounts of detailed data – including clinical and genomic data

- **Limitations**
- Variable information by source
- Groups assembled for a purpose – genomic analysis
- Access may be limited

Caris



Precision Oncology Alliance

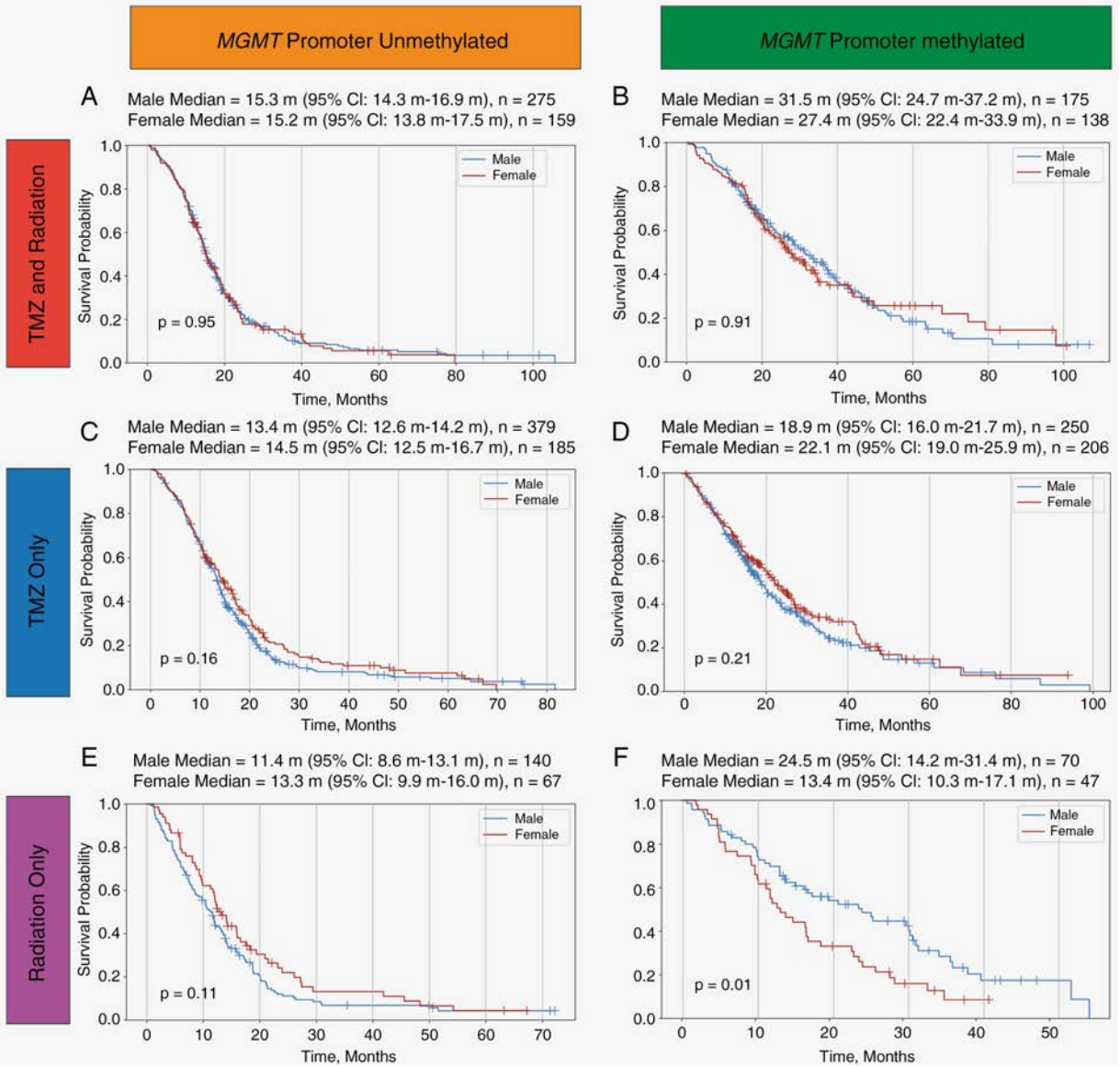
- LOI and presentation required
- No access to data - analysis run by internal research staff within Flatiron
- PI's team develops analytic plan and requests analysis provide

https://www.carislifesciences.com/

Sex Differences in Response to Treatment: Impact of MGMT Methylation (Dataset: CARIS)

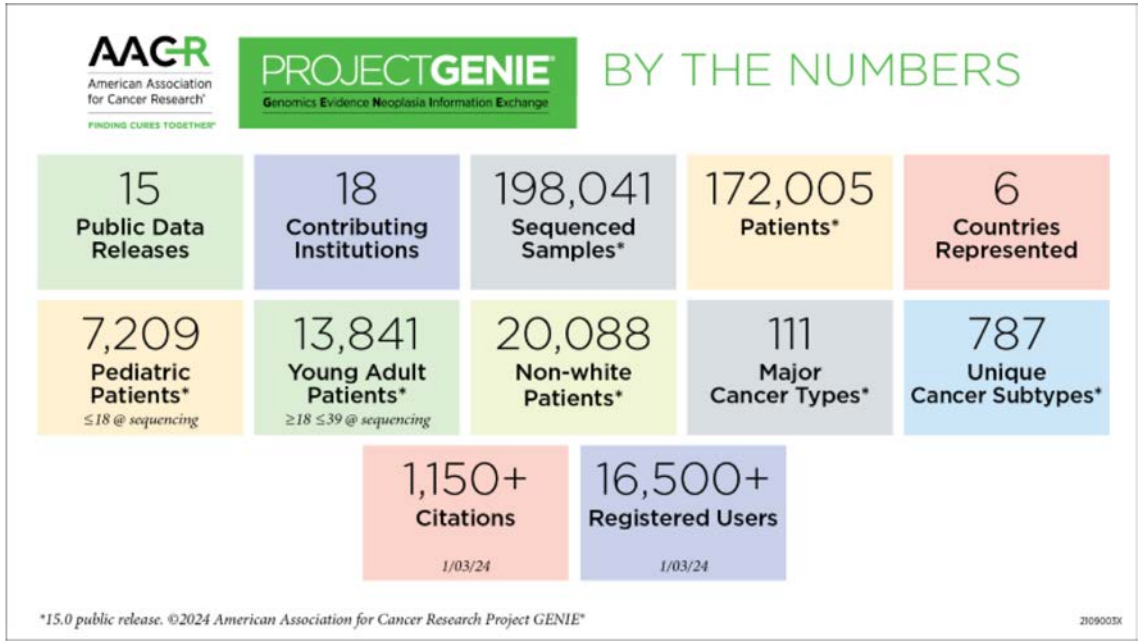
Median survival higher in MGMT methylated compared to unmethylated

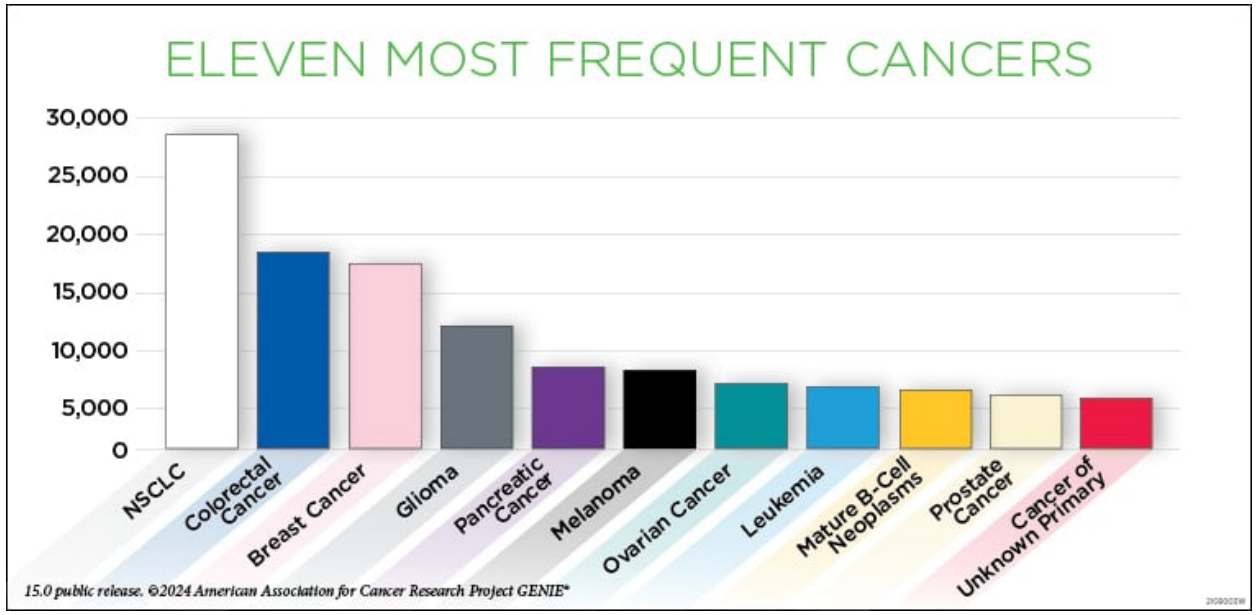
Radiation Only Treatment: increased survival in males with MGMT promoter methylation



Cioffi et al, Neuro Oncol Advances, 2024

AACR Genie
(Genomics Evidence Neoplasia Information Exchange)





Q. How do I get access to the data?
A. Go to www.cbioportal.org/genie/ and request access.

Tempus

5M+
de-identified* research records
to power large real-world evidence studies

950K+
records
with imaging data

500K+
records
with matched clinical and genomic data to understand driver mutations and outcomes associations

*Numbers
data as of

HOW WE PARTNER

Tempus View gives you near real time access to power your research initiatives. Interact with Tempus multi-modal data on-demand through an optimized cloud environment to help get the answers you need, when you need them.

Tempus Explore delivers speed, cost efficiency and deep domain expertise to help meet your research objectives. Derive insights from the full Tempus oncology database by leveraging Tempus bioinformatics experts.

Tempus Download offers a broad, typically multi-year relationship that supports numerous iterative research and development objectives that can span the lifecycle of your product. Get enterprise-wide access to Tempus multi-modal records by stakeholders across your organization in the indication(s) of your choice with this custom partnership model.

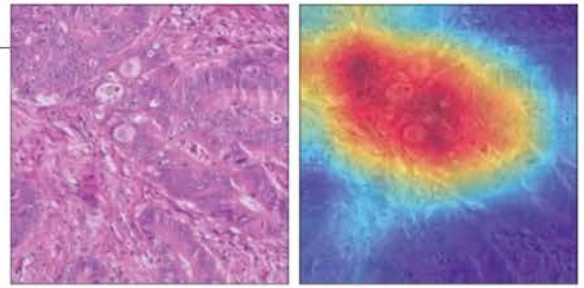
PARTNER WITH OUR AI EXPERTS

Inquire about collaboration opportunities with Tempus **AI and Analytics Experts** to support your drug development programs. We've established machine learning partnerships with numerous biopharmaceutical companies focused on areas such as:

- Risk and response predictions with multi-modal input
- Histogenomics – Image-based molecular biomarker predictions
- Tumor microenvironment characterization

We've partnered with:

15 of the top 20 oncology pharma companies*
70+ biopharma companies



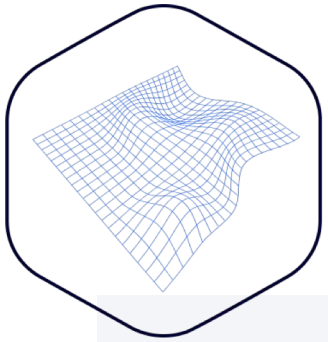
Original Image | Cam on Image

https://www.tempus.com/life-sciences/data-collaborations/

Agenda

1. Open Science Drives Discovery -- Importance and Goals of Data Sharing
2. Cancer Research Data Commons (CRDC) and Childhood Cancer Data Initiative (CCDI)
3. Data Drives Discovery
4. Other “Big Data” Resources
5. **Misc Items of Interest**

ARPA-H Biomedical Data Fabric* Toolbox



NCI in partnership with ARPA-H will advance the next-generation of tools to synthesize and speed use of health research data, starting with cancer

Make biomedical research data easier to use

Reduce effort for data integration

Develop new data fabric capabilities & tools

Build health data science models that can be applied across disciplines

* A data fabric provides a unified, consistent layer of data services that can work across many different systems and environments.

BDF Toolbox - Technical Areas

TA1: Automated Data Collection

Lower barriers to high-fidelity, timely, and automated data collection of research data across labs and health record systems

TA2: Machine-Assisted Curation

Prepare, connect, and harmonize multi-source data for analysis at scale

TA3: Intuitive Exploration

Enable advanced, human-centered data exploration and dashboards for use by diverse stakeholders and decision-makers

TA4: User Engagement

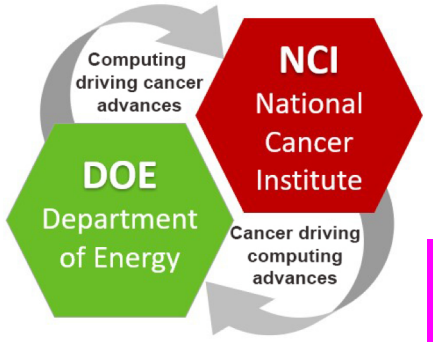
Evaluate data fabric tools across researchers, clinicians, and patients to create tools that will be enthusiastically adopted.

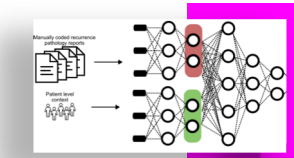
TA5: Cross-Domain Generalization

Leverage tools and platforms to generalize data across biomedical domains and disease types.

National Cancer Institute-Department of Energy Collaboration

A partnership to simultaneously accelerate advances in precision oncology and advanced scientific computing



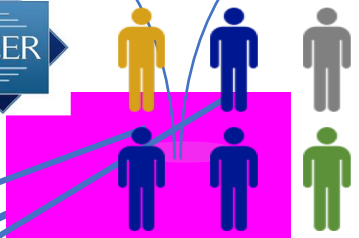


MOSSAIC: Modeling Outcomes using Surveillance data and Scalable Artificial Intelligence for Cancer

Clinical Domain

IMPROVE: Innovative Methodologies and New Data for Predictive Oncology Model Evaluation

Pre-clinical Domain

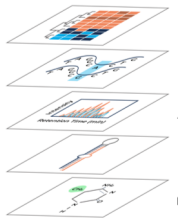


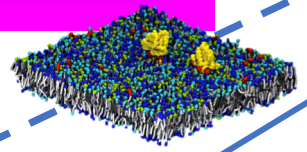
Population cancer models

ADMIRRAL: AI-Driven Multi-scale Investigation of RAS/RAF Activation Lifecycle

Molecular Domain

Multi-scale cancer models





me
models

computational.cancer.gov

NIH NATIONAL CANCER INSTITUTE

NIH Cloud Lab | Experiment in the Cloud

Funded and sponsored by the NIH Office of Data Science Strategy and managed by the Center for Information Technology, Cloud Lab is a no-cost, 90-day program for NIH intra- and extramural researchers to try commercial cloud services in an NIH-approved environment. Cloud Lab provides training and guardrails to protect against financial and security risks.

How It Works

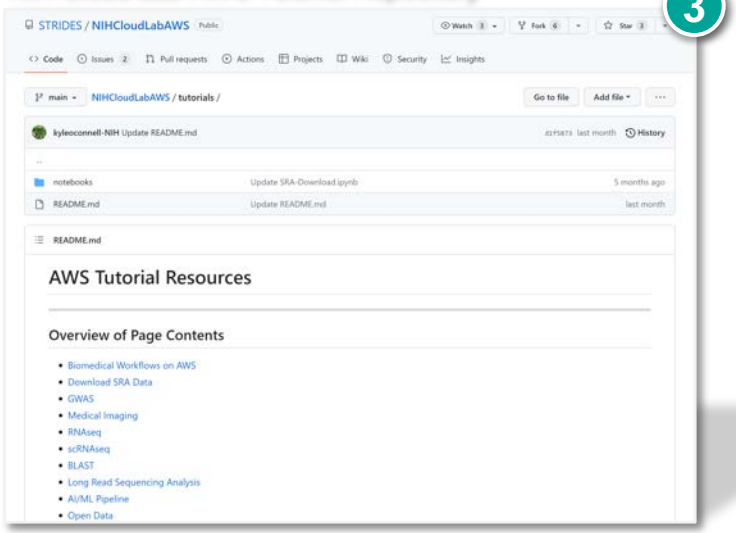
1. **Fill out** request form
2. **Get** account and $500 of credits
3. **Access** tailored cloud trainings
4. **Practice and learn** for 90 days

NIH Cloud Lab Sign Up Page

1

NIH Cloud Lab AWS Tutorial Repository

3



4

Why Use Cloud Lab?

Try Before You Buy

Evaluate if cloud is a good fit for your project without making a long-term commitment

Learn New Skills

Access tutorials that demonstrate how to run realistic bioinformatic, data science, and AI workflows

Develop New Tools

Prototype new architectures and evaluate software and hardware combinations

Explore Generative AI

Run GenAI tutorials to learn how to use this powerful new technology for your research

Get $500 of credits to use:

aws Azure Google Cloud

NIH NATIONAL CANCER INSTITUTE

Train at the NCI

NIH NATIONAL CANCER INSTITUTE

Search

About Cancer | Cancer Types | Research | Grants & Training | News & Events | About NCI

Home > Grants & Training > Training

Training

- Cancer Training at NCI
- Resources for Trainees
- Funding for Cancer Training
- Building a Diverse Workforce
- About Center for Cancer Training (CCT)
- Inside Cancer Careers Podcast
- CCT Staff & Contact

Training

The Center for Cancer Training (CCT) supports NCI's goal of training cancer researchers for the 21st century. CCT provides funding to support training and career development for cancer researchers working at institutions nationwide and also manages intramural training programs offered at NCI laboratories and offices in Maryland. Learn more about CCT's mission.

Find Funding Chatbot

Utilize the "Find Funding" chatbot to easily connect you to Funding for Cancer Training Opportunities at NCI!

New Inside Cancer Careers Podcast

In the first episode, Dr. Mary Grace Katusiime and Dr. Camille Lange discuss the importance of building a strong professional network and having a personal professional motto throughout

https://www.cancer.gov/grants-training/training

Acknowledgments

CBIIT Leadership Team – datascience.cancer.gov

Jeff Shilling, CIO
Jaime Auvil-Guidry, PhD
Marcos Munoz

All IDS Program federal staff and contractors

All ODS Program federal staff and contractors

All CBIIT staff and contractors

Barnholtz-Sloan DCEG Research Team – dceg.cancer.gov

CBTRUS Team – www.cbtrus.org

All partners throughout NCI, NIH, HHS and externally

@NCIJBSloan
jill.barnholtz-sloan@nih.gov