NCI Cancer Policy & Infrastructure: Driving Impactful Data Sharing

Jaime Guidry Auvil, Ph.D.
Office of Data Sharing

NIH NATIONAL CANCER INSTITUTE

January 25, 2023

NCI Push to Support "Open Science"

A "movement" to make **scientific research** (including ***publications, DATA, physical samples, and software***) and dissemination **accessible** to all levels of society, amateur or professional.

- Open science is transparent and accessible knowledge that is shared and developed through **collaborative networks**.
- It encompasses practices such as:
  - publishing open research & campaigning for **open access**,
  - encouraging scientists to practice open-notebook science (such as **openly sharing data and code**),
  - broader dissemination and engagement in science, and
  - generally making it easier to publish, access and communicate scientific knowledge.
- Usage of the term varies substantially across disciplines, with a notable prevalence in the STEM disciplines.

Benefits of Broad Data Sharing

Collaborator Sharing

- Between investigator to investigator (e.g., sharing upon publication and request to the author)

Consortium Sharing

- Within large collaborative groups (e.g., sharing between investigators within a consortium/ network)

Broad Sharing

- Ensures fair and equitable access and secondary use of data by the wider research community (e.g., NIH Genomic Data Sharing Policy)
- Has the most impact on *driving scientific innovation and discovery and ensuring replication of results*
- Broad sharing ≠ Open access data

Framingham Heart Study: Success in Data Collection Over Time

BY THE NUMBERS: Uncovering the Mysteries of the Heart

By American Heart Association News

70 Years the Framingham Heart Study continues to break new ground on cardiovascular disease

5,209 Initial volunteer participants

15,447 Participants over the past 70 years

3 Generations who have participated in the study

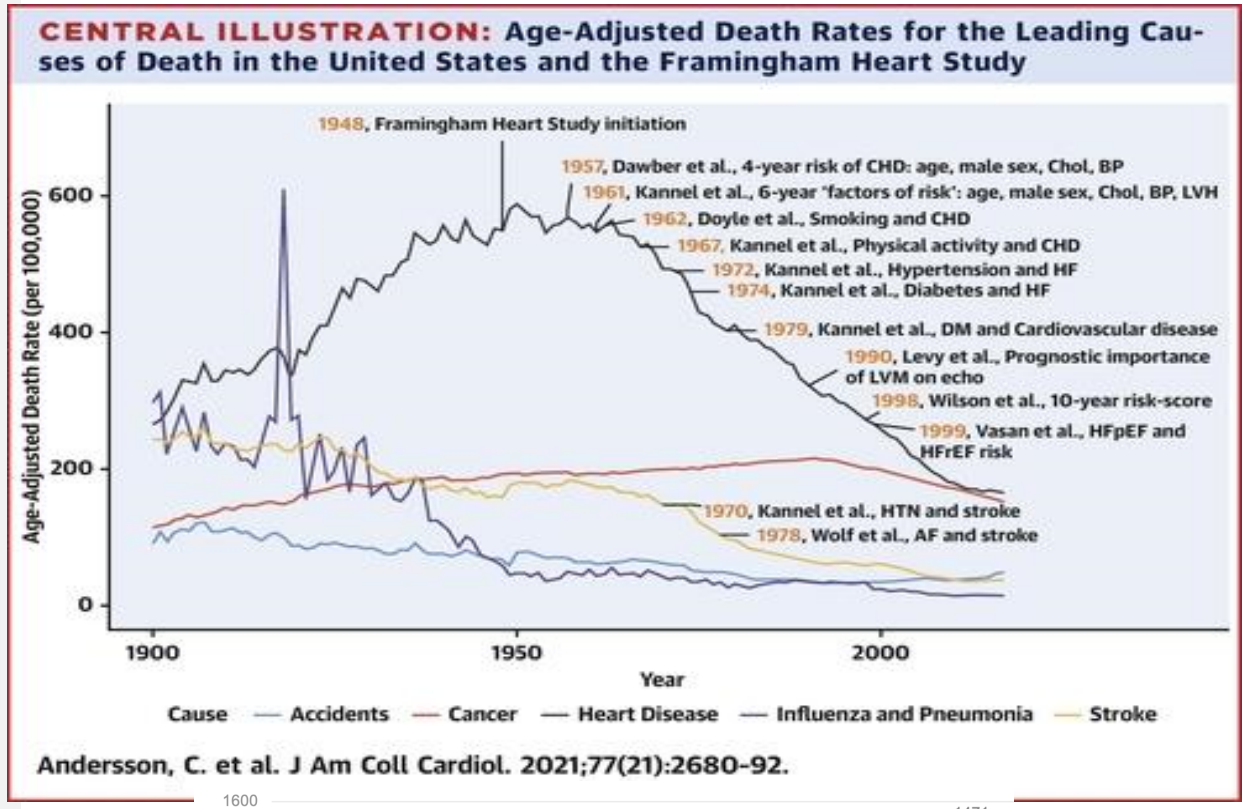
3,698 Published journal articles based on Framingham Heart Study data

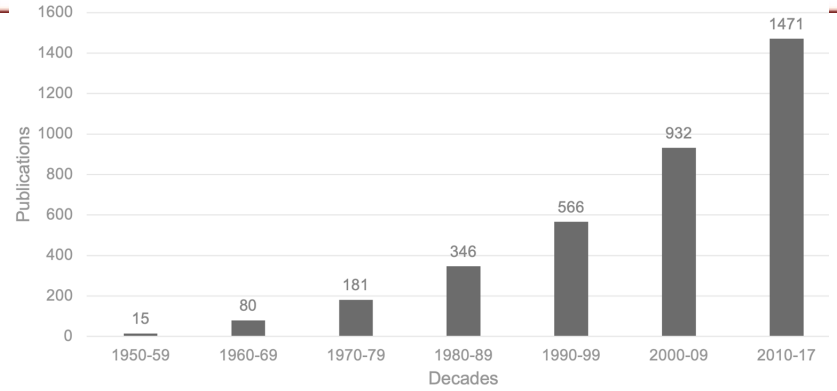
1960 Year the study pinned cigarette smoking as a risk factor for heart disease

802 Participants who have donated or registered to donate their brain for further study

Sources: Framingham Heart Study, Boston University

Published Oct. 10, 2018



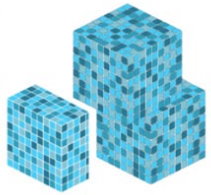


Total articles published through November 2017 = 3,561

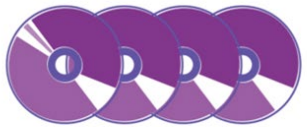
The Cancer Genome Atlas: Success in Open Team Science

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data



To put this into perspective, **1 petabyte** of data is equal to **212,000** DVDs



TCGA data describes **33** DIFFERENT TUMOR TYPES ...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

THE TEAM

20 COLLABORATING INSTITUTIONS across the United States and Canada

WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.

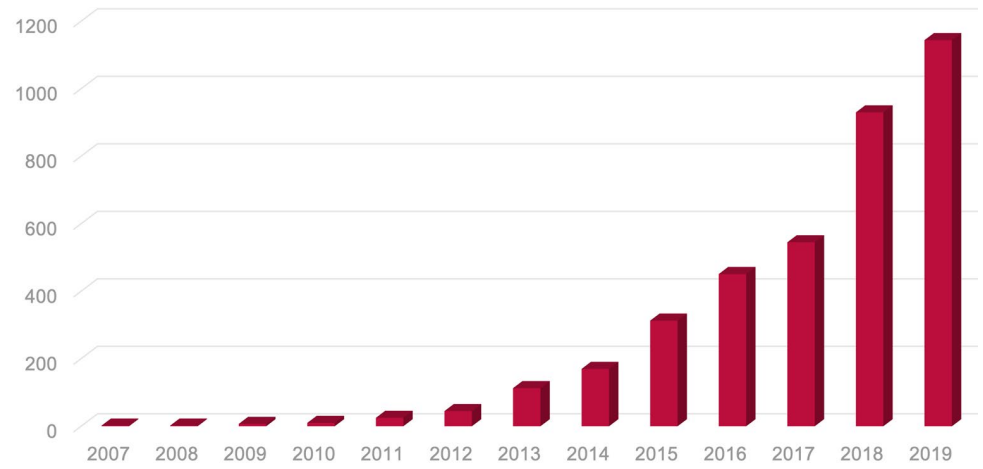
*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

www.cancer.gov/ccg

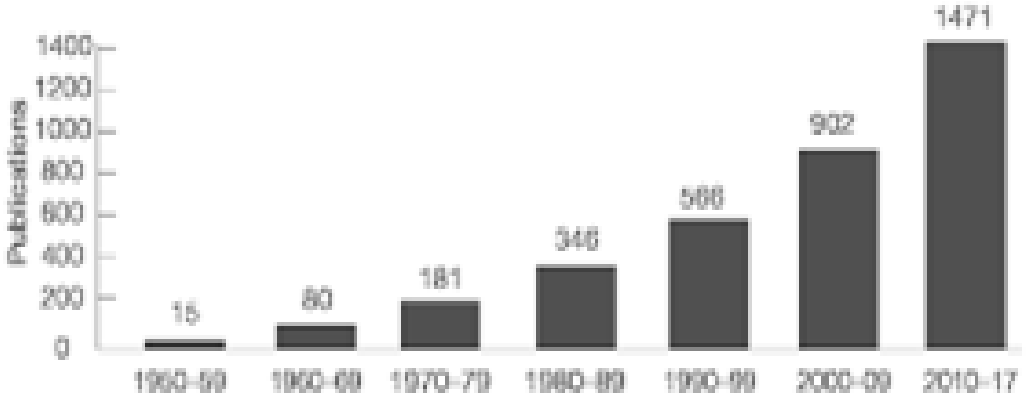
TCGA RESULTS & FINDINGS

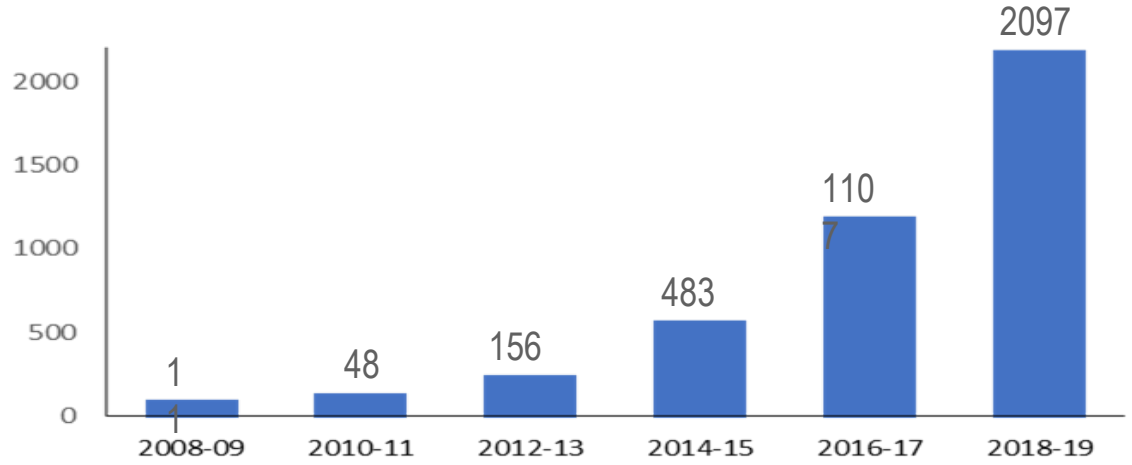
MOLECULAR BASIS OF CANCER	Improved our understanding of the genomic underpinnings of cancer	For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.
TUMOR SUBTYPES	Revolutionized how cancer is classified	TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*
THERAPEUTIC TARGETS	Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development	TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

Number of Publications Using TCGA Data



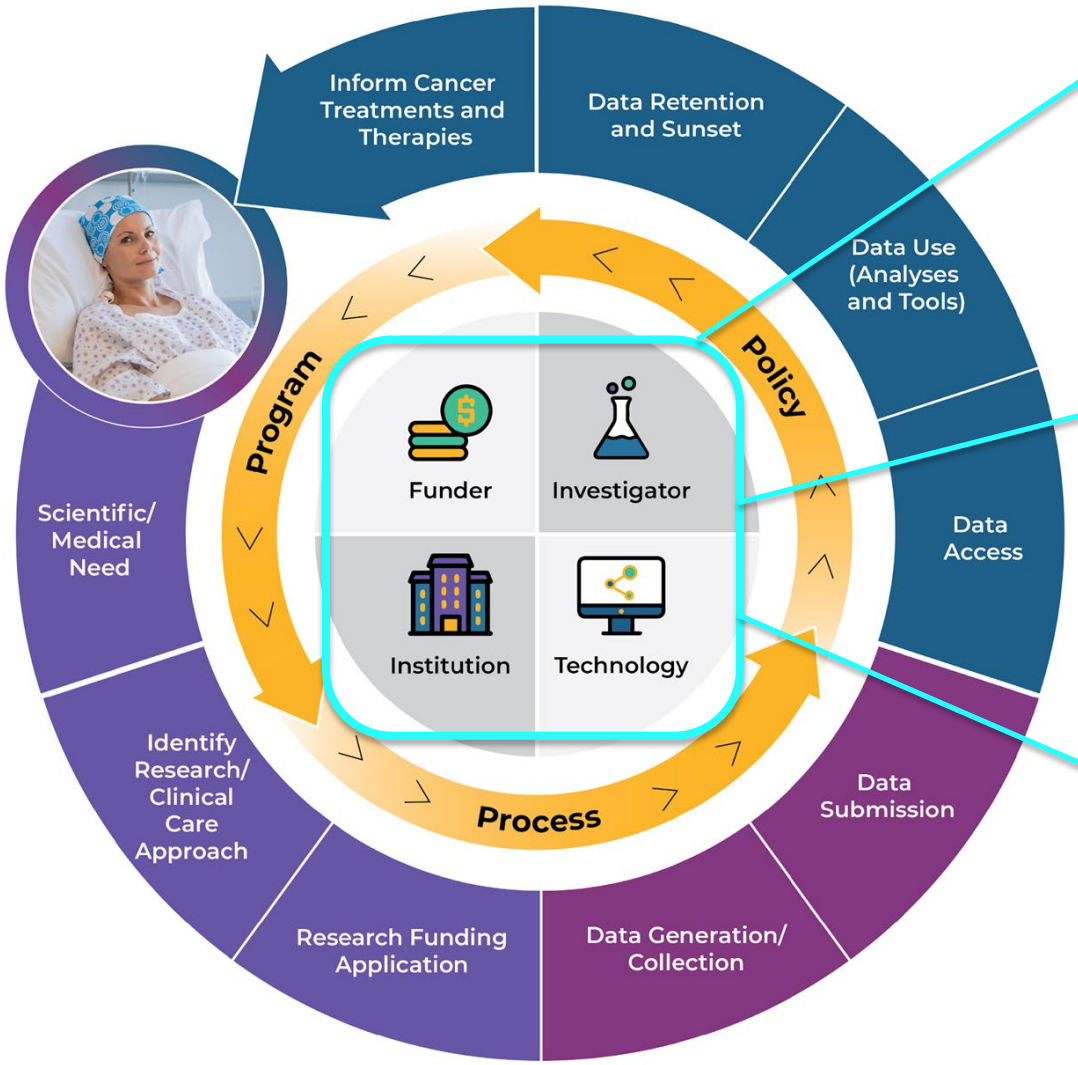
Driving Science through Publications, Data & Collaboration





	Framingham Heart Study	The Cancer Genome Atlas
Study Length	70 years	12 years
Cases Studied	15,144	11,429
Publications	**3,698 (~38,000 PMC)**	**3,747 (~62,000 PMC)**
Controlled-access Data	Consortia; HMB (+IRB/MDS, 2K=NPU)	Collaborative Teams & Public Use of Data; GRU
Authorized Users	715	3,335
Open Data Use & Availability Timing	Little Open Data; mostly available with publication	Some Open Data; All data immediately available to community

Scientific Data Lifecycle: Keys to Impactful Discovery



Critical Questions to Answer

Programs that define therapeutic needs and essential scientific gaps to be filled using structured datasets.

Policies to Promote Broad Use

Implementation of aggressive data management, sharing and access policies that ensure rapid, free and immediate access to all types of data.

Infrastructure to Support FAIR Principles

Technology platforms and tools that employ standards to make data findable, accessible, interoperable and reusable.

"Enable all participants across the cancer research & care continuum to contribute, access, combine & analyze diverse data that will enable new discoveries and lead to lowering the burden of cancer."

- *NCI Cancer Moonshot℠ Mission*

Moonshot Public Access & Data Sharing Policy

Make publications & data immediately and broadly available to the public

Data From

All NCI-Supported Cancer Moonshot Research Projects generating Publications & Data on or after October 1, 2017:

- Extramural grants
- Contracts
- Intramural research

Applies to human & non-human data

Award/Contract Expectations

Submit public access and data sharing plan

Share data to extent feasible, widely and immediately:

- Open-access attribution license (Creative Commons)
- Available through NIH data repository preferably (CRDC, TCIA, NCBI/ dbGaP)

Data Access

Provide final, peer-reviewed manuscript to NLM PubMed Central *(within ~4 weeks of journal release)*

Make publication available immediately with no embargo

The Cancer Moonshot: Success in Mission-Driven Science

Cancer Moonshot℠:
Accelerate discovery, increase collaboration, and expand data sharing

In the Cancer Moonshot's first 4 years (2017–2021):

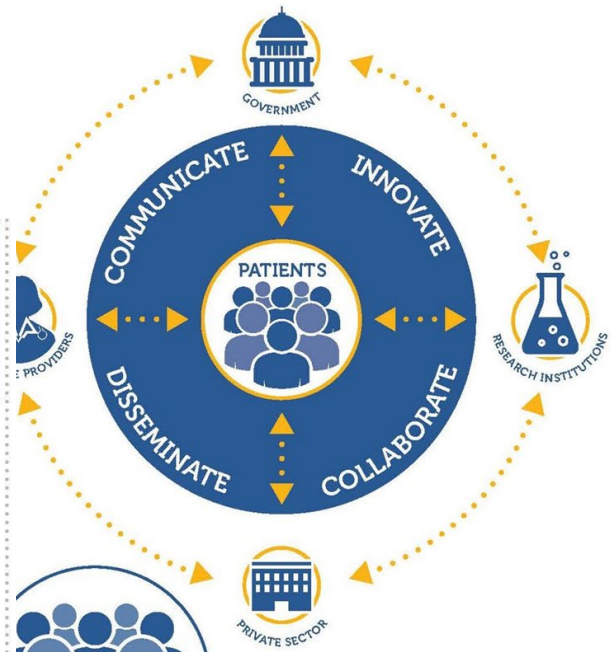
>2,000	49	>30
Publications	Clinical Trials	Patent Filings

CANCER MOONSHOT

INITIATIVES 2017–2022

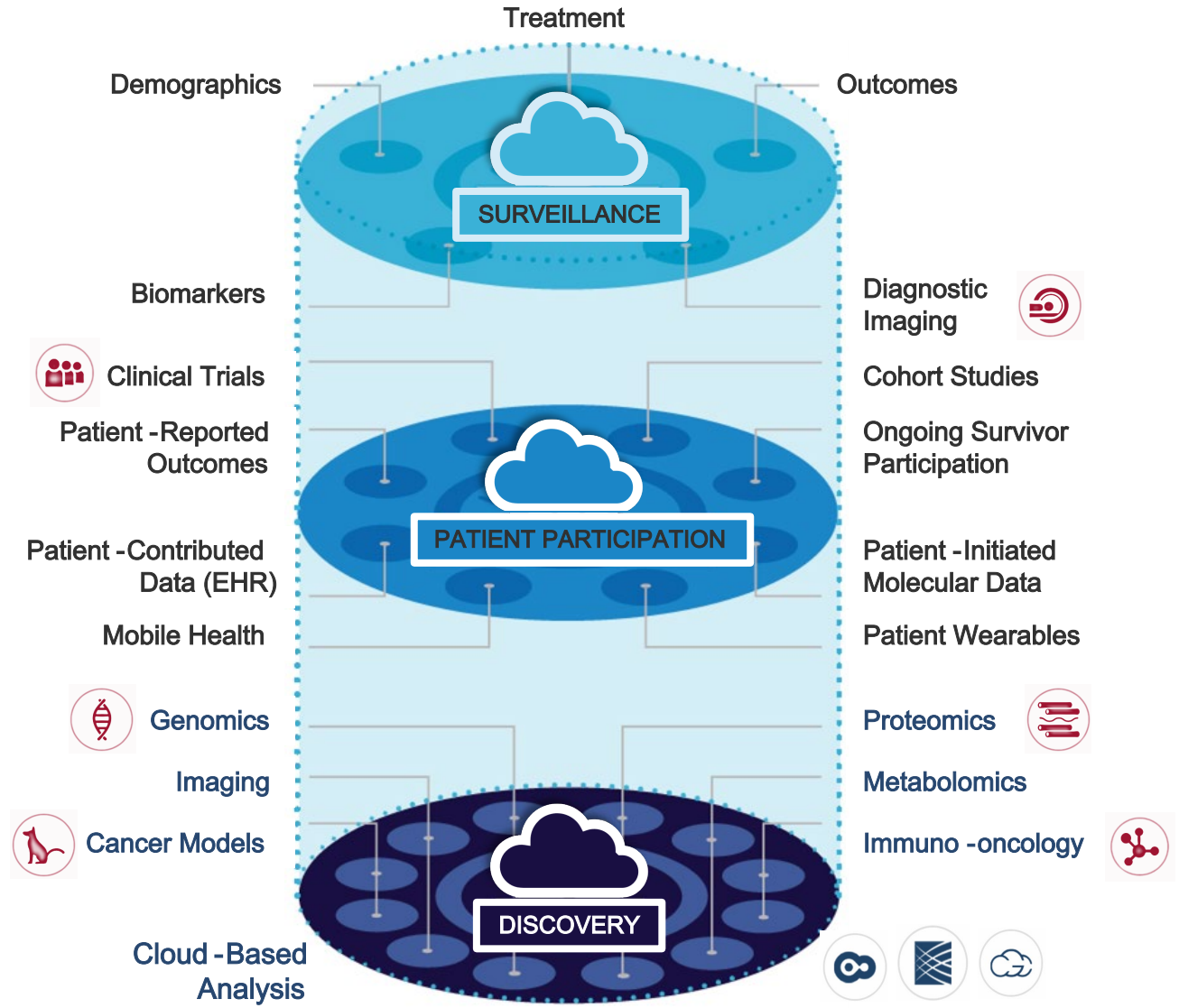
OVER **70** CONSORTIUMS OR PROGRAMS

OVER **240** RESEARCH PROJECTS



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

National Data Ecosystem: Integrating Cancer Research



SEER

AACR American Association for Cancer Research FINDING CURES TOGETHER

PROJECTGENIE Genomics Evidence Neoplasia Information Exchange

APOLLO

HTAN HUMAN TUMOR ATLAS NETWORK

THE CANCER GENOME ATLAS

CCDI BUILDING A COMMUNITY CENTERED AROUND CHILDHOOD CANCER CARE AND RESEARCH DATA — DATA TYPES: CLINICAL, TREATMENT, OUTCOME, MOLECULAR, BIOSPECIMEN, LONGITUDINAL, POPULATION — LEARN FROM EVERY CHILD — Improving the quality, consistency, and accessibility of data to make it easier for researchers to develop new and better treatments for children with cancer.

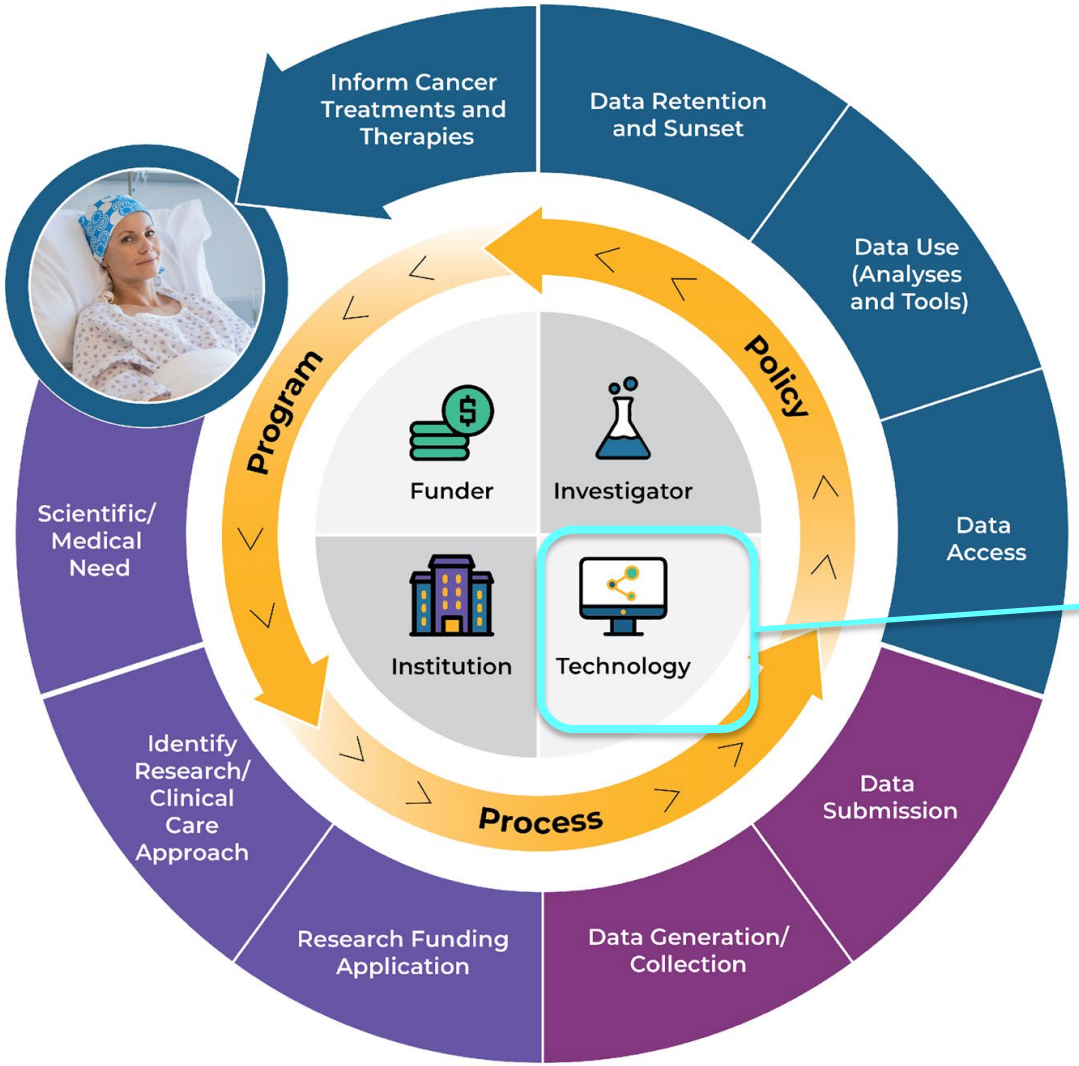
NCI National Clinical Trials Network a National Cancer Institute program

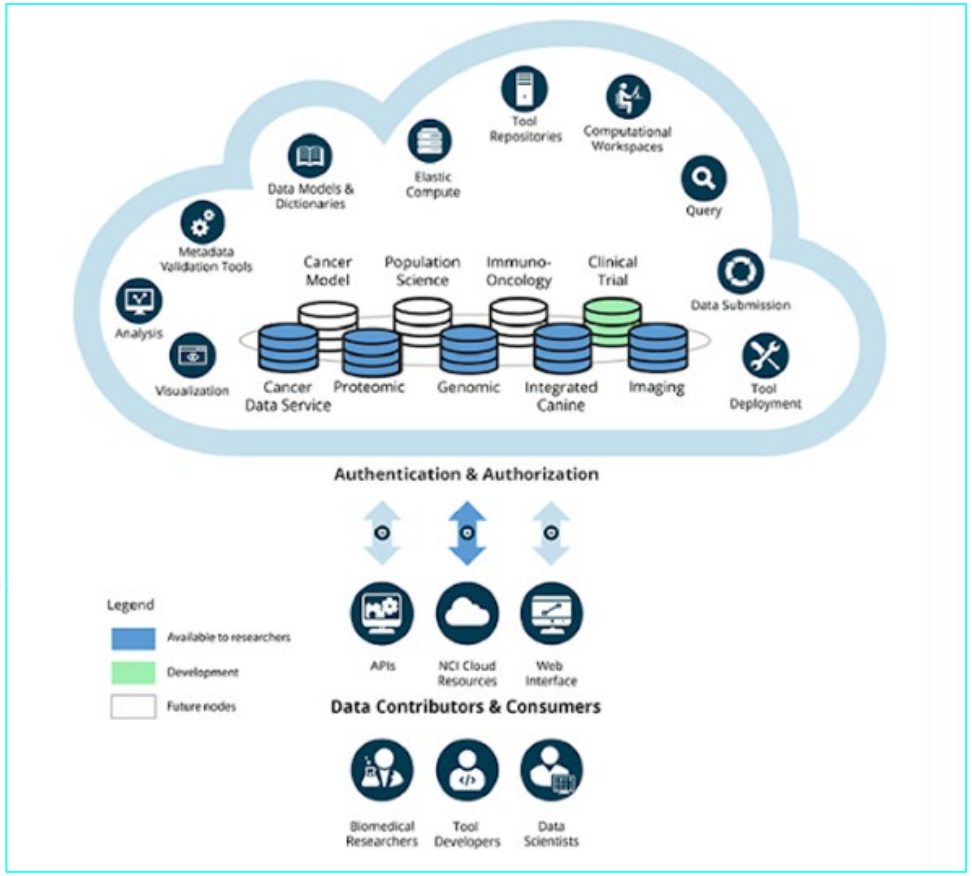
CPTAC National Cancer Institute at the National Institutes of Health

NCI National Cancer Institute — DOE Department of Energy

The Cancer Research Data Commons (CRDC)

Sharing Data Openly through a Cancer Data Ecosystem





NCI Cancer Research Data Commons (CRDC)

Opportunities to Define Impactful Data Sharing

- Think like a ***data user*** rather than a data generator (what reference data is needed for mining or innovation)
- Define what ***data types*** and ***standards*** will have the most value and utility; establish data standards where there are current gaps or needs
- Encourage ***open*** and ***broad usage*** by the largest possible community to promote and accelerate discovery
- Define the repository ahead of time; use existing whenever possible (both fit for purpose and Generalist Repositories)
- Pursue data federation (connecting to data where it lives), not consolidation
- Set expectations on ***timing*** of data availability

Contact Us About Data Sharing

nciofficeofdatasharing@mail.nih.gov

#NCIODS

datasharing.cancer.gov

Questions

https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/funding/public-access-policy

CRDC: Statistics & Impact

CRDC Repositories

Genomic Data Commons

65 K+ users/month

2.9 PB+ data
85,000+ cases

~2 PB data download/month

Proteomic DC

29 TB data
1 M+ peptides

Imaging DC

20 TB data
400 K+ image series

Cancer Data Service

80 TB data
1.3 PB coming soon

Integrated Canine DC

25 TB data
490+ cases

NCI Cloud Resources

12,000+ registered users

2,300+ years of compute

3.8 PB+ data available

1,800+ public tools & workflows

8,000+ user-created workflows

Across the CRDC

200+ Scientific Publications
300+ Studies/Collections Released

CRDC: Statistics & Impact as of Spring 2022