High-Value Data Stimulates Research

Health Research Alliance

Open Science and Data Sharing Session

NIH NATIONAL CANCER INSTITUTE

Tony Kerlavage, PhD
Director, Center for Biomedical Informatics and Information Technology

20 April 2022

Outline

1. *Impact of Data Sharing*
2. *Evolution of Data Sharing Policies*
3. *National Cancer Data Ecosystem*
4. *Challenges & Opportunities*

Impact of Data Sharing

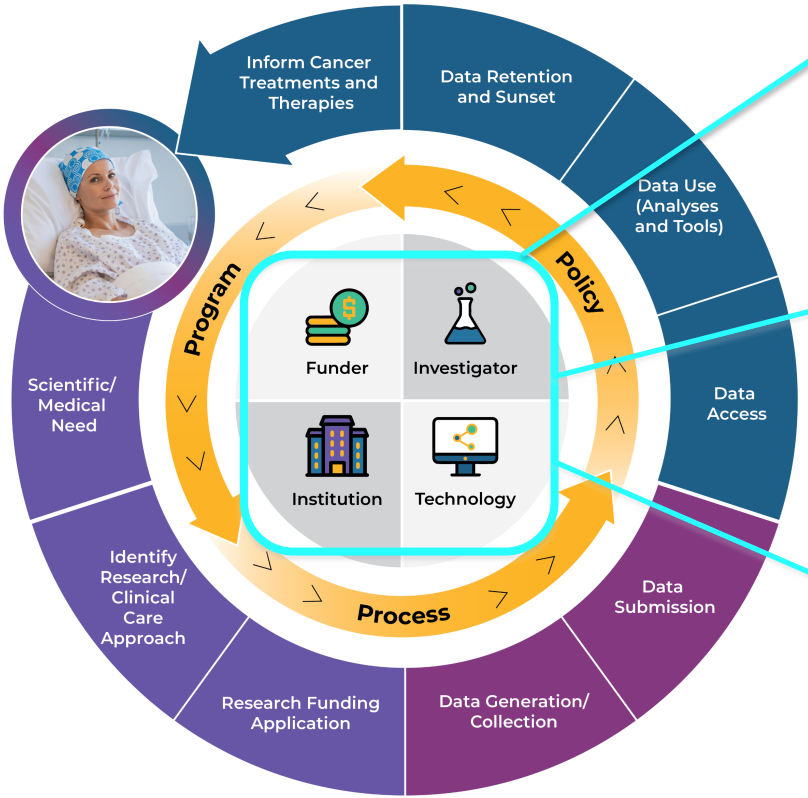
What *IS* Data Sharing?

Data Sharing – The practice of making research data & metadata available for use by the **broader community**

Benefits *OF* Data Sharing

- Stimulate **innovation and discovery:** Secondary use of data leads to **new discoveries** (knowledge, products & procedures)
- Ensure **replication** of results
- Ensure **transparency & openness** - the scientific method

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.

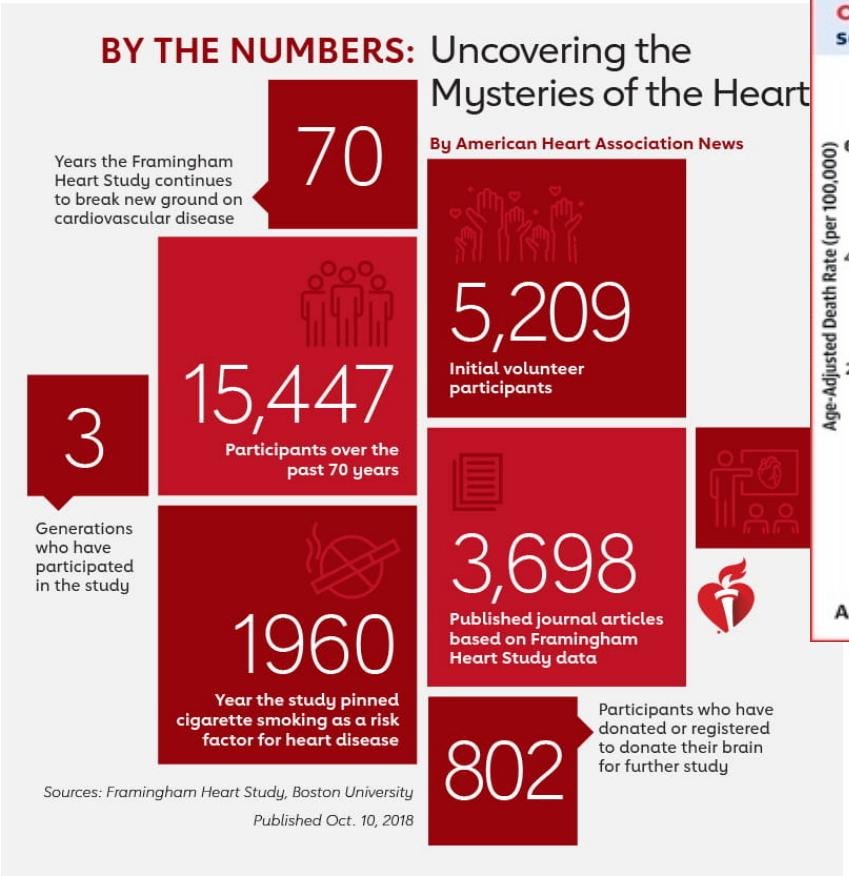
Policies to Promote Broad Use

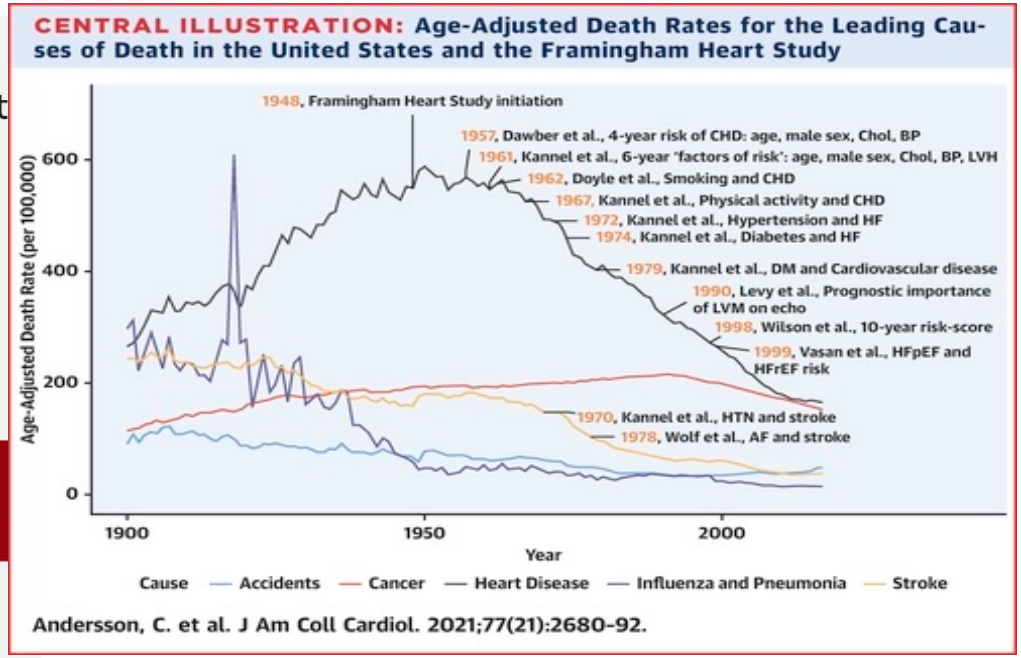
Implementation of aggressive data management, sharing and access policies that ensure rapid, free and immediate access to all types of data.

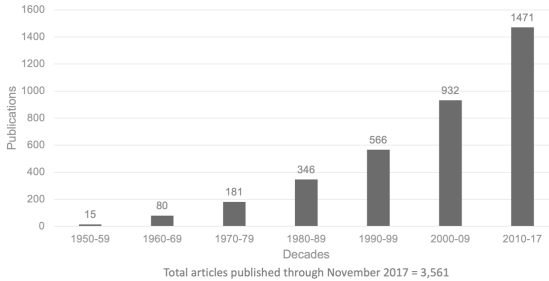
Infrastructure to Support FAIR Principles

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

Framingham Study: Success in Data Collection Over Time





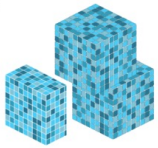


The Cancer Genome Atlas: Success in Open Team Science

TCGA BY THE NUMBERS

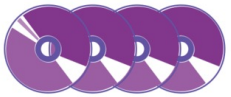
TCGA produced over

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

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.

THE TEAM

20 COLLABORATING INSTITUTIONS across the United States and Canada

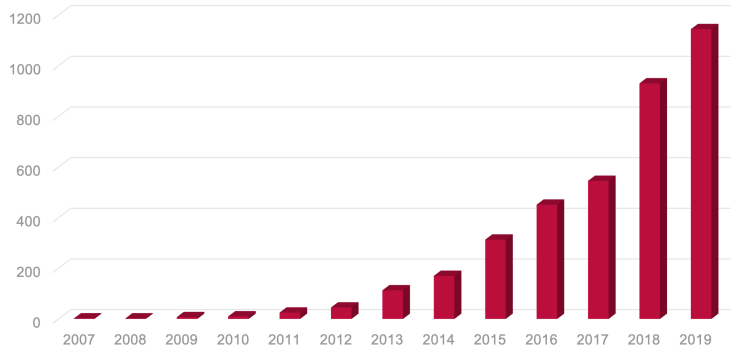
WHAT'S NEXT?

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.

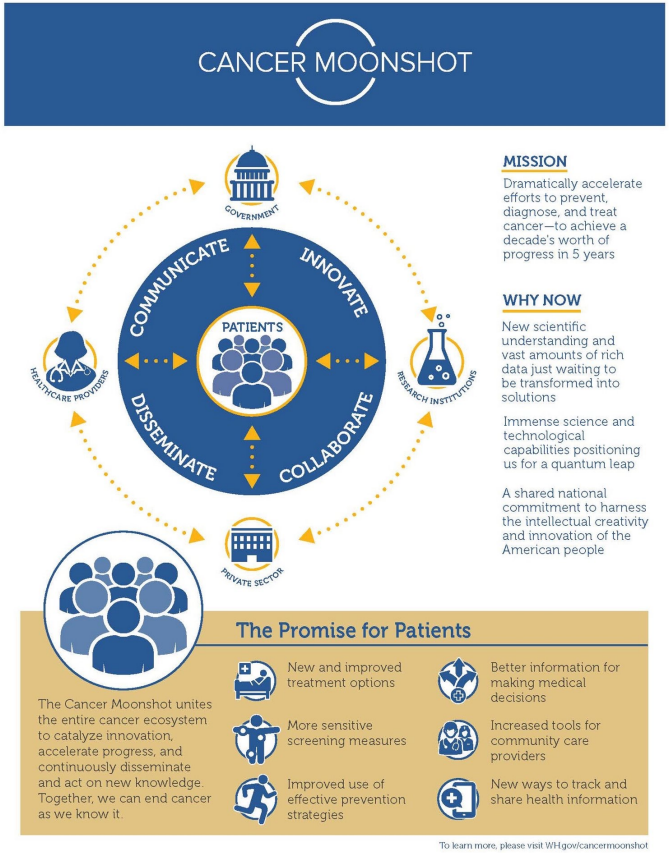
*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

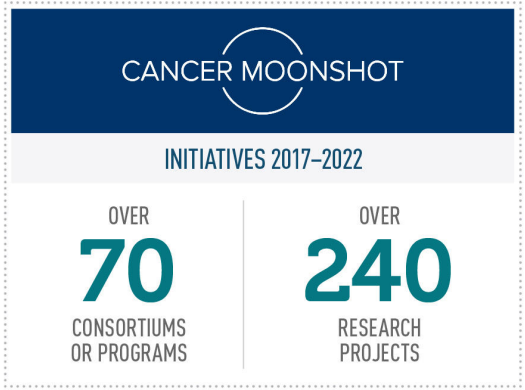
www.cancer.gov/ccg

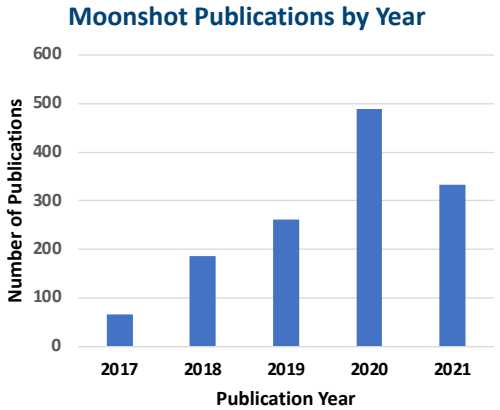
Number of Publications Using TCGA Data



The Cancer Moonshot: Success in Mission-Driven Science



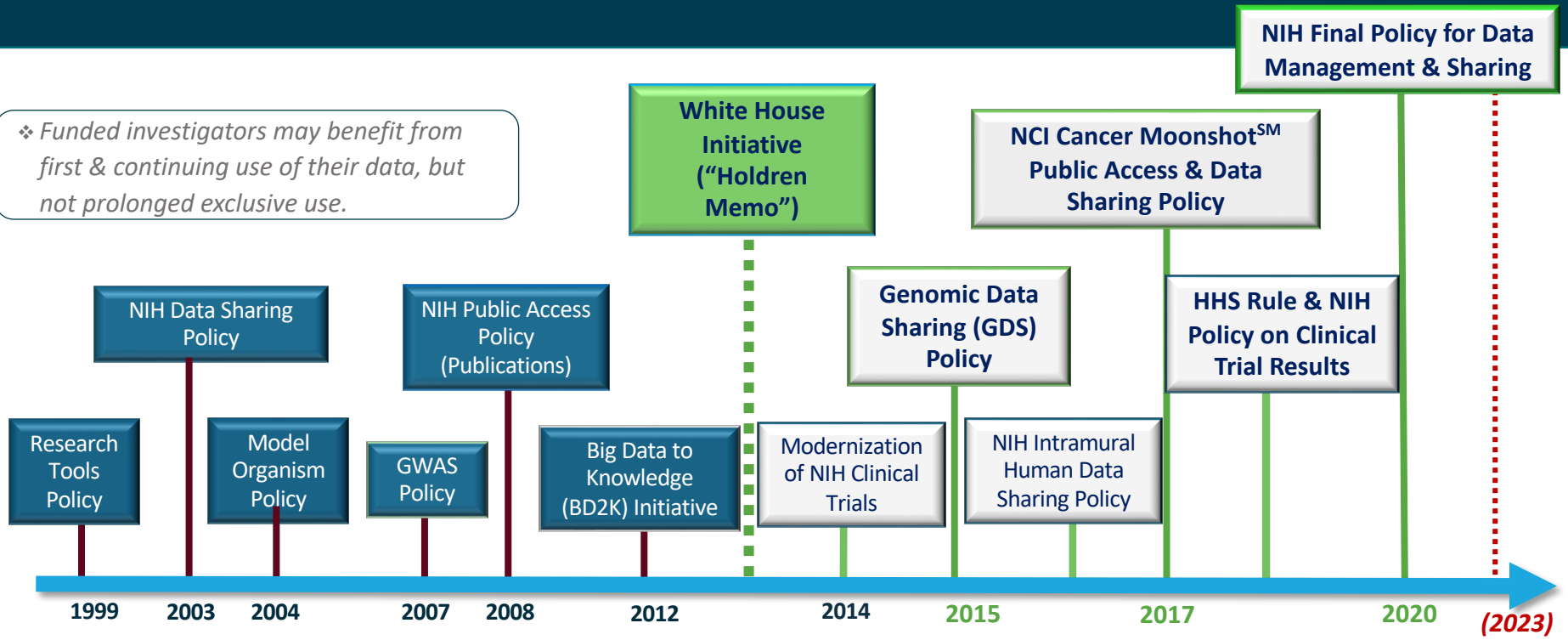




Evolution of Data Sharing Policy

Key NIH & NCI Data Sharing Policies

❖ *Funded investigators may benefit from first & continuing use of their data, but not prolonged exclusive use.*



NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

Investigators must share any information necessary to understand, develop or reproduce published research (raw data, statistical methods, tools, source code)

NIH Genomic Data Sharing (GDS) Policy

Set expectations for making genomics data available in a timely fashion to the research community

Policy Expectations

- Genomics data must be shared broadly with research community for mining & discovery
- Applies to human & non-human data; all types of funding w/o $ threshold: extra/intramural grants, contracts, OTAs
- Timing expectations for release (<9 months for raw data; publication for analyzed results)

Successful Programs

- **Genome-wide association studies** (GWAS) – multiple disease areas
- **The Cancer Genome Atlas** (NCI) – tumor/matched normal samples from 33 cancers
- **TOPMed** (NHLBI) – whole genome sequences on ~186K patients in heart, lung, blood & sleep disorders
- **GTEx** (Genotype-Tissue Expression project) - explore genetic variants across human tissue

Infrastructure

- Platforms to share individual level, potentially identifiable data (controlled-access; DAC approval)
  - Consent defines data use
  - Data de-identified to protect patient privacy & confidentiality
- Make summarized, non-identifiable data broadly available (open access)
- dbGaP, SRA, GDC, BioData Catalyst, AnVIL

Clinical Trials Access Policy

Ensure timely, public availability of results from NCI-supported clinical trials

Policy Expectations

- Share final results of clinical trials (timely, comprehensive): registration 21 days after 1st enrollment, results available 12 months
- NCI-supported clinical trials (intra/ extramural grants/contracts) of:
  - FDA-reg drug, biologics, devices
  - Pediatric post-market surveillance studies (devices under FD&C Act).
- Submit study reports to accessible registries

Successful Programs

- **Molecular Analysis for Therapy Choice Trials** (MATCH; NCI) - precision oncology for relapsed cancer patients
- **Prevention and Early Treatment of Acute Lung Injury** (PETAL; NHLBI) - prevent or provide early treatment for acute respiratory distress syndrome
- **Accelerating Covid-19 Therapeutic Interventions & Vaccines** (ACTIV; NCATS) - vaccine and therapy trials

Infrastructure

- Data available <1 yr Trial's Primary Completion Date
- Platforms to share summary level data of trials results: Clinicaltrials.gov, CTRP, NCTN Archive
- Some individual level data from subjects available (tissues samples collected during clinical trials):
  - NCI TARGET (GDC), Kids First (DRC), PCGC (BDC)
  - Covid-19 (N3C, RADx, IMPORT)

Moonshot (HEAL) Public Access & Data Sharing Policy

Make publications & data immediately & broadly available to the public

Policy Expectations

- Every funded project (intra/ extra-mural grants, contracts) must submit public access and data sharing plan
- Plan becomes term & condition of award (NoA)/ contract deliverable
- Publications & data must be made available to public freely & immediately with no embargo (open access)

Successful Programs

- **Human Tumor Atlas Network (HTAN)**
- **My Pediatric And Rare Tumor Network (MyPART)**
- **Moonshot Biobank / Direct Patient Engagement Network**
- **Cancer Research Data Commons (CRDC)**
- **NIH HEAL Initiative** *(scientific solutions for opioid crisis) - employed same policy*

Infrastructure

- Share data to extent feasible, widely and immediately
  - Preferably available through NCI or NIH data repository (CRDC,NCBI)
  - Open-access attribution license (Creative Commons)
- Creation of a national cancer ecosystem; cloud-based, tools & interfaces for all types of data uses

Final NIH Policy on Data Management & Sharing

Make publications & data immediately & broadly available to the public

Policy Expectations

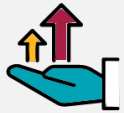
- Every funded project (intra/ extra-mural grants, contracts) must submit data management & sharing plan
  - Plan becomes term & condition of award (NoA)/ contract deliverable
- Good data management (sharing & preservation; FAIR data principles)
- Timely sharing of scientific data: as soon as possible and by publication

❖ *Takes effect **January 25, 2023***

Foundational Initiatives

- **NIH HEAL Initiative**
- **NCI Cancer Moonshot**
- **COVID data efforts**
- **NIH All of Us**
- **NCI Childhood Cancer Data Initiative**

https://sharing.nih.gov/

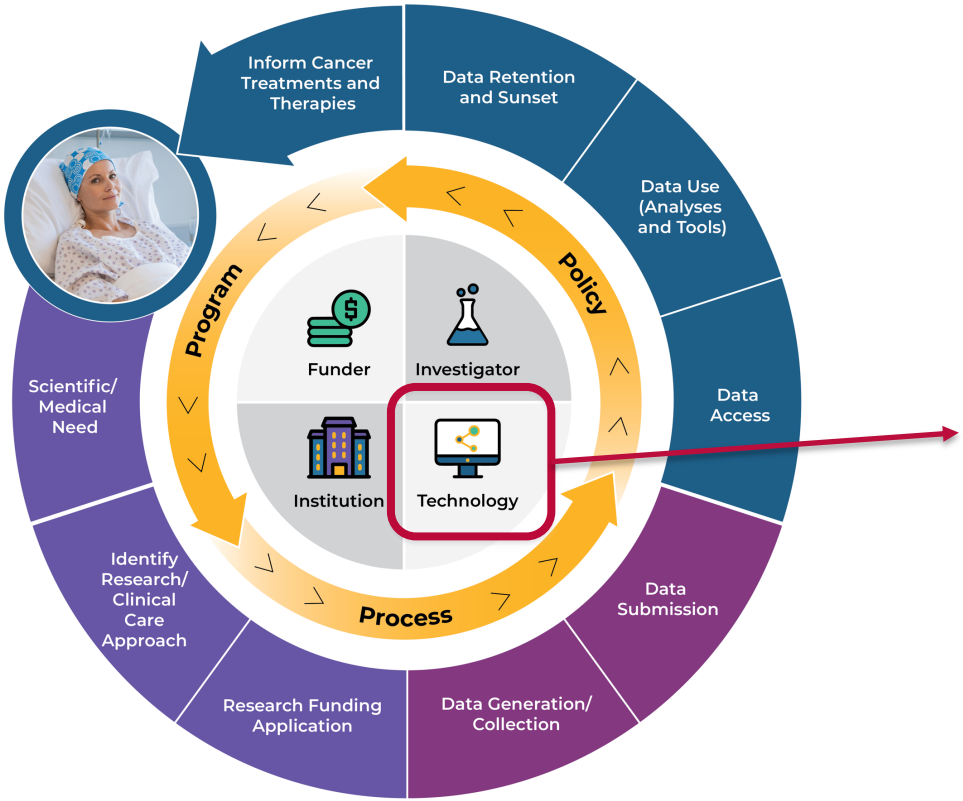


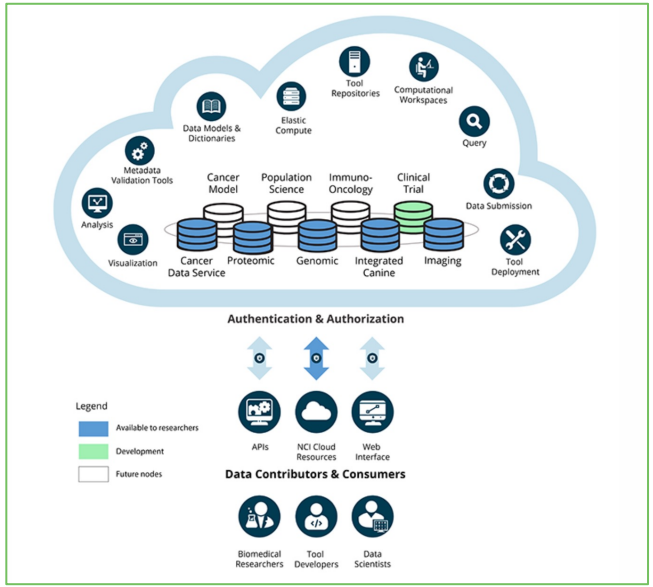
Infrastructure

- Share data to extent feasible, widely and immediately
  - Preferably available through NCI or NIH data repository or ecosystem (CRDC,NCBI)
- Connect research, clinical and public health data for maximal benefit to science & participants
  - N3C, RADx, NCPI, CCDI

National Cancer Data Ecosystem

Scientific Data Lifecycle & the CRDC





Cancer Research Data Commons

National Cancer Data Ecosystem

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

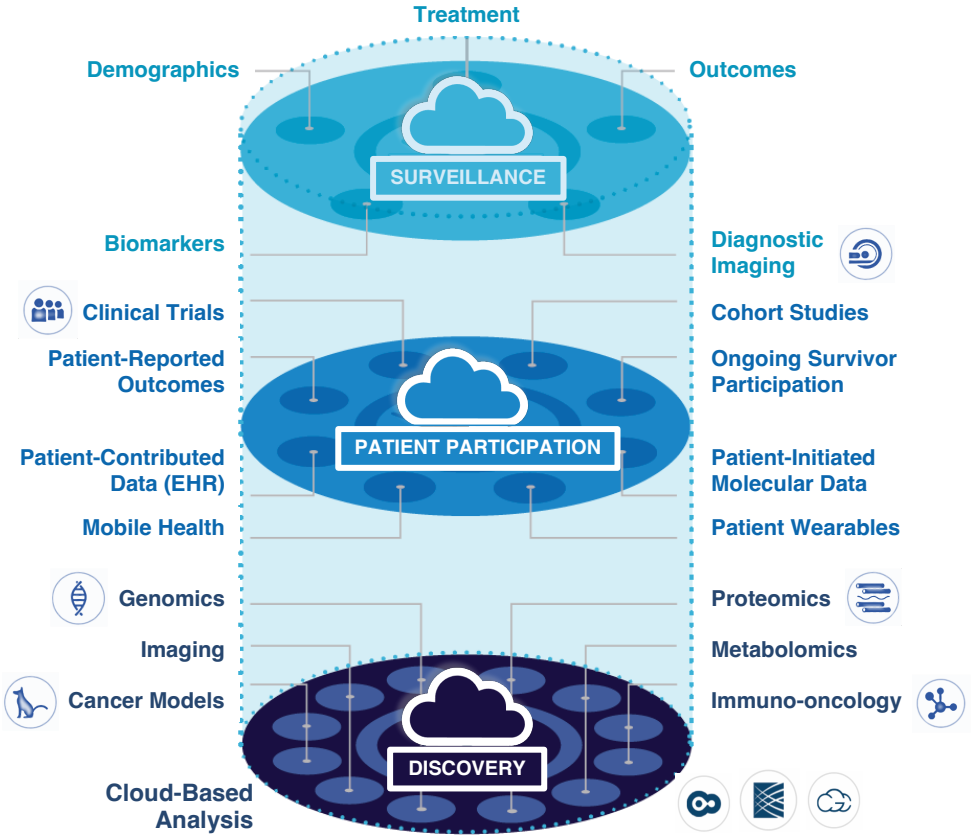
Overarching goals

- **Accelerate progress in cancer, including prevention & screening**
  - From cutting edge basic research to wider uptake of standard of care
- **Encourage greater cooperation and collaboration**
  - Within and between academia, government, and private sector
- **Enhance data sharing**

Recommendations

- **Build a National Cancer Data Ecosystem**
  - Enhanced cloud-computing platforms
  - Services that link disparate information, including clinical, image, and molecular data
  - Essential underlying data science infrastructure, standards, methods, and portals for the Cancer Data Ecosystem

National Data Ecosystem: Integrating Cancer Research



The Cancer Research Data Commons (CRDC)

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

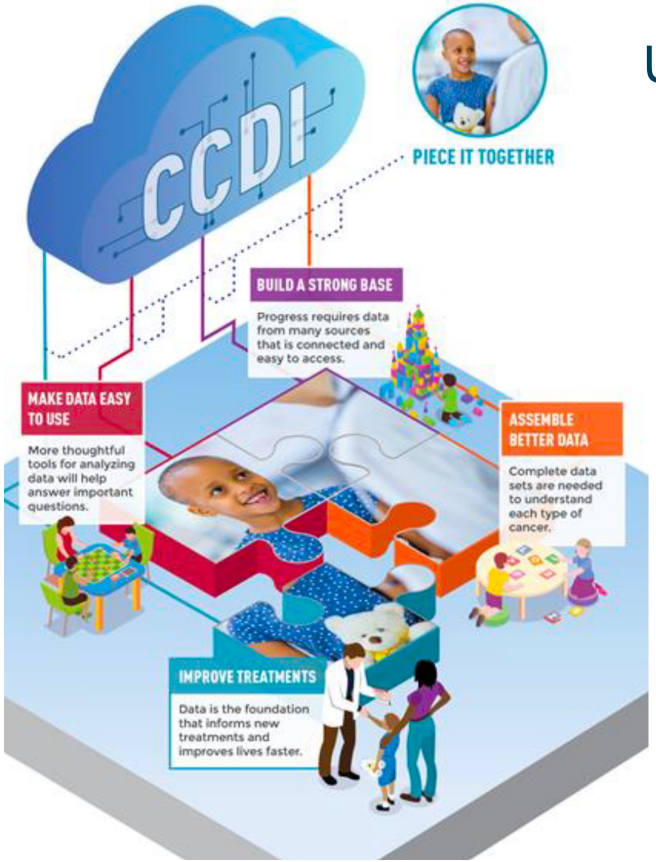
1,800+ public tools & workflows

8,000+ user-created workflows

Across the CRDC

200+ Scientific Publications
300+ Studies/Collections Released

Childhood Cancer Data Initiative



Using data to achieve the goals of CCDI

Piece it together: CCDI is completing the puzzle to learn from and help heal children, teens, and young adults with cancer.

Build a strong base: Progress requires data from many sources that is connected and easy to access.

Make data easy to use: More thoughtful tools for analyzing data will help answer important questions.

Assemble better data: Complete data sets are needed to understand each type of cancer.

Improve treatments: Data is the foundation that informs new treatments and improves lives faster.

CCDI

BUILDING A COMMUNITY CENTERED AROUND CHILDHOOD CANCER CARE AND RESEARCH DATA

LEARN FROM EVERY CHILD

DATA TYPES:

- CLINICAL
- TREATMENT
- OUTCOME
- MOLECULAR
- BIOSPECIMEN
- LONGITUDINAL
- POPULATION

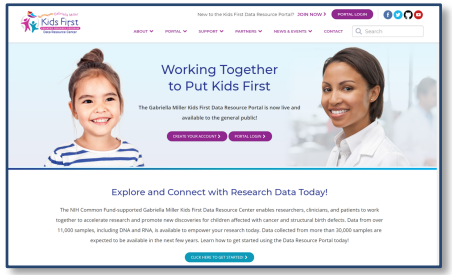
PROGRAM GOALS:

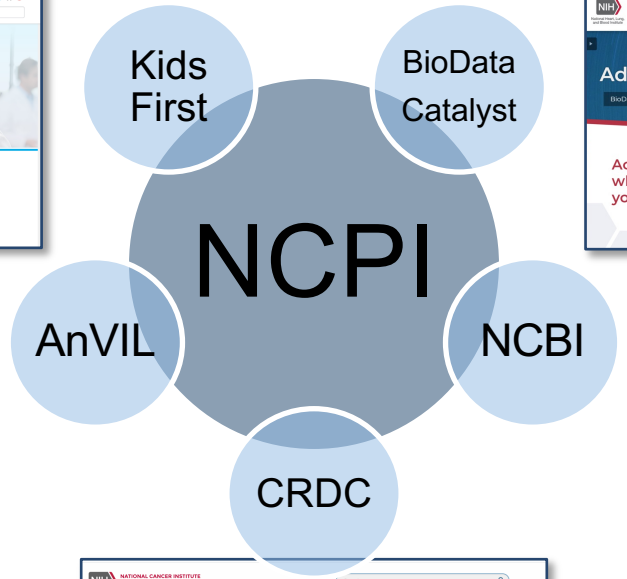
- DEVELOP NEW TREATMENTS
- IMPROVE EXISTING TREATMENTS
- IMPROVE OUTCOMES, QUALITY OF LIFE, AND SURVIVORSHIP

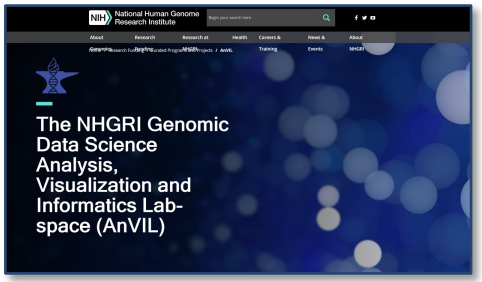
IMPROVING THE QUALITY, CONSISTENCY, ACCESS, AND ACCESSIBILITY OF DATA FROM EVERY CHILD WILL ALLOW US TO ACHIEVE OUR GOALS.

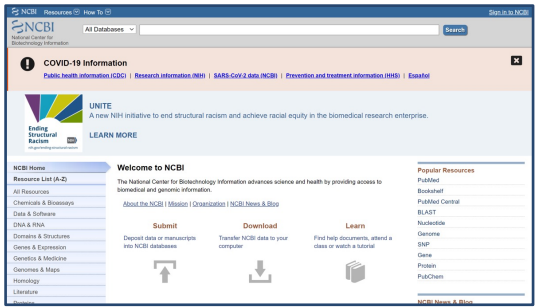
NCPI: NIH Cloud Platforms for Interoperability

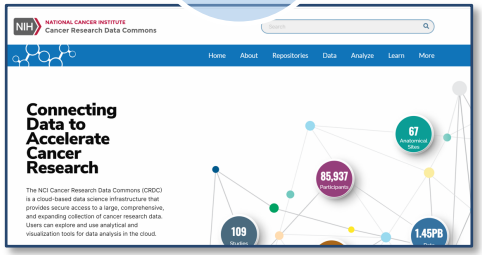
Connecting with a Greater Data Ecosystem











Challenges & Opportunities

Challenges for Data Driven Research

Inconsistent Sharing Policies

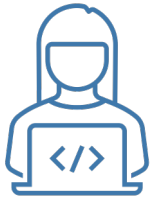
- No universal standard
- Liberal exceptions
- Government or institutional prohibitions
- Discretion of scientist

Data Complexity

- Difficult to find and analyze multiple data types from multiple data sources
  - Basic research
  - Model systems
  - Clinical trials
  - Population-level studies

User Skills and Tools

- Most researchers are not bioinformaticians or data scientists
- Skill levels for data handling and data analysis varies
- Availability of analysis tools varies on platforms



Data Storage and Usage

- Data often stored in separate data repositories for download
- Use of data or combining datasets may require multiple downloads, moving data, or multiple DUAs

Setting Expectations for Researchers Prior to Funding

- Guide researchers to define up front what data will have the most value and build this into the study
- Think like a data user rather than a data generator
- Take the lead and push for a data continuum – work towards high-value data sets
- Encourage open and broad usage by the largest possible community
- Define the repository ahead of time; use existing whenever possible
- Pursue data federation, not consolidation
- Set expectations on timing of data availability

How can the NIH help you today?

- **NCI Data Science**
  - https://datascience.cancer.gov
- **NCI Office of Data Sharing**
  - https://datascience.cancer.gov/data-sharing
- **NIH Data Sharing Website (just launched)**
  - https://sharing.nih.gov
- **NIH Office of Data Science Strategy**
  - https://datascience.nih.gov/about/odss
- **National Center for Biotechnology Information**
  - https://www.ncbi.nlm.nih.gov
- **NIH Generalist Repositories**
  - https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html

Thank You!

www.cancer.gov

www.cancer.gov/espanol