

Balancing Broad & Equitable Data Sharing to Benefit Research & Participant Communities: a NCI ODS Perspective



Jaime M. Guidry Auvil, Ph.D.
Director, ODS, CBIIT

NCI ODS MISSION



- The National Cancer Institute's (NCI) The NCI Office of Data Sharing (ODS), headquartered within the Center for Biomedical Informatics and Information Technology (CBIT), is creating a comprehensive data sharing vision and strategy for NCI and the cancer research community.
- The office advocates for the proper balance of open-access and broad data sharing policies to enable reproducibility, secondary use, and knowledge sharing. ODS respects the rights of the public to participate in and benefit from publicly funded research while considering the critical importance of intellectual property concerns for individuals and organizations to support a healthy commercial marketplace



Office of Data Sharing Activities

- Advises on ethical data access and sharing issues, policies, and practices
- Enhances the accessibility and utility of research data and metadata, in part by refining data and metadata standards (prevent data silos)!
- Develops sustainable, achievable, and meaningful incentives for data sharing
- Supports equitable sharing through a robust, sustainable data management ecosystem
- Encourages participation in major data-sharing initiatives, including contributing to NIH-supported data repositories
- Creates educational resources to guide the cancer community on the importance and processes of sharing data

To contact the NCI Office of Data Sharing,
email nciofficeofdatasharing@nih.gov

Islands of data
Disconnected data silos



Different:
Standards
Quality
Databases
Semantics

Talk Objectives



- 1. Define “Data Sharing” and Establish the importance of responsible and broad data sharing to advance disease knowledge and improve care**
2. Outline current NIH data sharing policies (including GDS Policy) and dbGaP procedures.
3. Discuss barriers to sharing and potential ways to overcome them
 - Lessons learned and potential solutions to barriers in broad & equitable sharing (Informed Consent, submission to public dBs)
 - Clinical data/phenotype, open data → “coded” data sets
4. Introduce ways that the new NCI Office of Data Sharing is advocating for the proper balance of broad and open data sharing/access while respecting the right of patients to participate in and benefit from research as they see fit.

What *IS* Data Sharing??

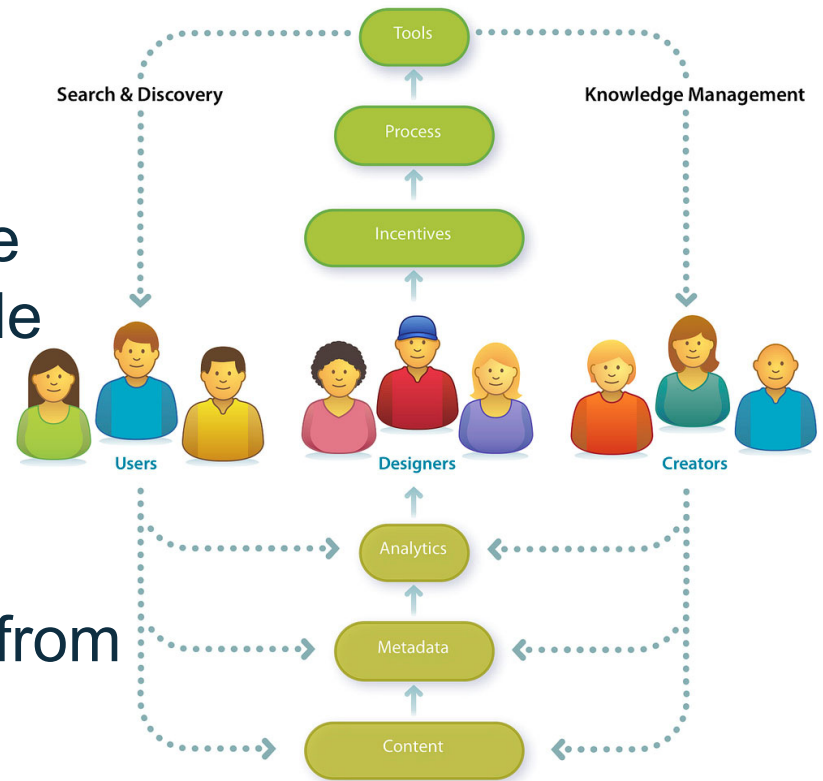
- Data sharing – practice of making data & metadata used for scholarly research available to other investigators.
 - Can be done a variety of ways
 - Replication is key
 - Transparency & openness are part of the scientific method



- Various funding agencies and science/biomedical journals require authors of peer-reviewed papers to make available (“share”) any supplemental information (ie. raw data, statistical methods or source code) necessary to understand, develop or reproduce published research.

Benefits of Data Sharing

- Enables data generated from one study to be used to explore a wide range of additional research questions
- Increases statistical power and scientific value by enabling data from multiple studies to be combined
- Facilitates reproducibility and validation of research results
- Facilitates innovation of methods and tools for research
- Reduces duplication and saves time, valuable resources & experimental costs



Public Wants Broad Data Sharing



The NEW ENGLAND
JOURNAL of MEDICINE

Michelle Mello, JD, PhD, Van Lieou, BS, & Steven Goodman, MD, PhD. "Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing" NEJM, June 7, 2018

- **93%** were very or somewhat likely to allow their own data to be shared with university scientists.
- **82%** were very or somewhat likely to share with scientists in for-profit companies.
- >8% of respondents felt that the potential negative consequences of data sharing outweighed the benefits.
- Willingness to share data did not vary appreciably based on purpose for data use (fewer participants were willing to share their data for use in litigation).
- Majority of participants believe there would be **“great benefit”** for healthcare companies & medical production (72%), physicians treating patients (81%), and scientists improving knowledge and treatment protocols (85%)
- Greatest concerns were:
 - others may be less willing to enroll in clinical trials (37% very/ somewhat concerned)
 - data would be used for marketing purposes (34%)
 - data could be stolen (30%).
 - Less concern about discrimination (22%) or exploitation of data for profit (20%).

Talk Objectives



1. Define “Data Sharing” and Establish the importance of responsible and broad data sharing to advance disease knowledge and improve care
2. **Outline current NIH data sharing policies (including GDS Policy) and dbGaP procedures.**
3. Discuss barriers to sharing and potential ways to overcome them
 - Lessons learned and potential solutions to barriers in broad & equitable sharing (Informed Consent, submission to public dBs)
 - Clinical data/phenotype, open data → “coded” data sets
5. Introduce ways that the new NCI Office of Data Sharing is advocating for the proper balance of broad and open data sharing/access while respecting the right of patients to participate in and benefit from research as they see fit.

NIH View on Data Sharing

- As part of NIH's long-standing policy to share and make available to the public the results and accomplishments of the activities that it funds, NIH reaffirms its support for the concept of data sharing.
- We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health.
- The NIH endorses the sharing of final research data to serve these and other important scientific goals.
- The NIH expects and supports the timely release (by the time of publication) and sharing of final research data from NIH-supported studies for use by other researchers.
- Investigators submitting an NIH application seeking \$500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible.

Seeking Appropriate Balance

- Natural tension between values & needs:
 - Protect privacy and research integrity
 - Respect broad range of participant wishes
 - Promote health advances through research
 - Support investigators and their ability to do good work



Setting NIH Policy Guidelines

NIH Office of Science & Policy focuses on two broad areas of Scientific Data Sharing:

- **Genomics and Health** - analyzes the scientific, ethical, and social implications of genetic and genomic research on health and, as warranted, provides policy recommendations on particular issues or concerns raised through genomic research and emerging technologies. Engages in trans-NIH and inter-agency collaborations to advance implementation of genomic medicine.
- **Scientific Data Management** - evaluates opportunities and challenges regarding the generation, management, sharing, and access of scientific research data. Provides policy recommendations, as needed, on issues/concerns associated with data sharing and management (ie. data science, public access, open science, big data)



NIH OSP Scientific Data Sharing

Genomics and Health

- NIH Genomic Data Sharing
- HeLa Cell Whole Genome Sequence Data Sharing
- Genomic Medicine
- Genetic Testing Registry

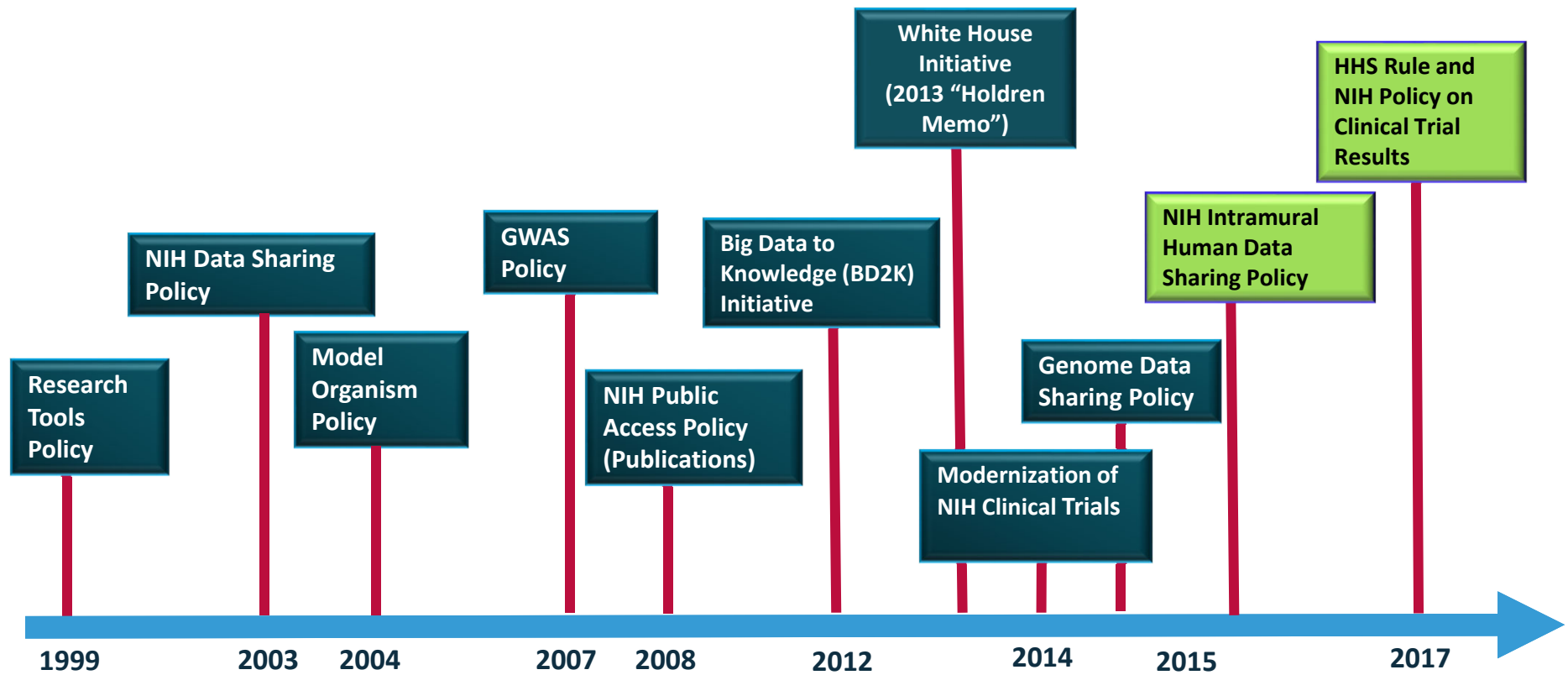


Scientific Data Management

- NIH Data Management and Sharing Activities Related to Public Access and Open Science
- NIH Data Science Policy Council
- Interagency and International Open Science Efforts



History of NIH Data Sharing Policy



NIH Intramural Human Data Sharing Policy (08/2015)

- The NIH's mission is to improve the health of the public through support of biomedical research and the training of biomedical scientists. To further advance and accelerate research to benefit the public health, data developed in the NIH Intramural Research Program (IRP) should be collected in a manner that permits and promotes the broadest sharing possible.
- Data sharing may be complicated or limited, in certain cases, by agreements with outside collaborators, e.g., Clinical Research and Development Agreements (CRADAs) or Clinical Trial Agreements, by Institutional Review Board (IRB) rules or by other constraints. NIH IRP investigators should share data broadly for secondary research purposes, in all cases consistent with applicable laws, regulations and policies.

Key Points of Policy:

- Applies to all human data in the NIH IRP (NIH Clinical Center & NIH Institutes/Centers).
- A Data Sharing Plan must be developed for any research involving human data and will be included in the IC scientific review.
- Clinical investigators are expected to develop protocols and consent processes/forms to enable broad data sharing for secondary research consistent with this Policy.
- Sharing data for secondary research purposes shall comply with human subjects research regulations and procedures, if applicable.
- All IRP investigators are encouraged to deposit data in publicly accessible research repositories for sharing to the extent feasible and appropriate.

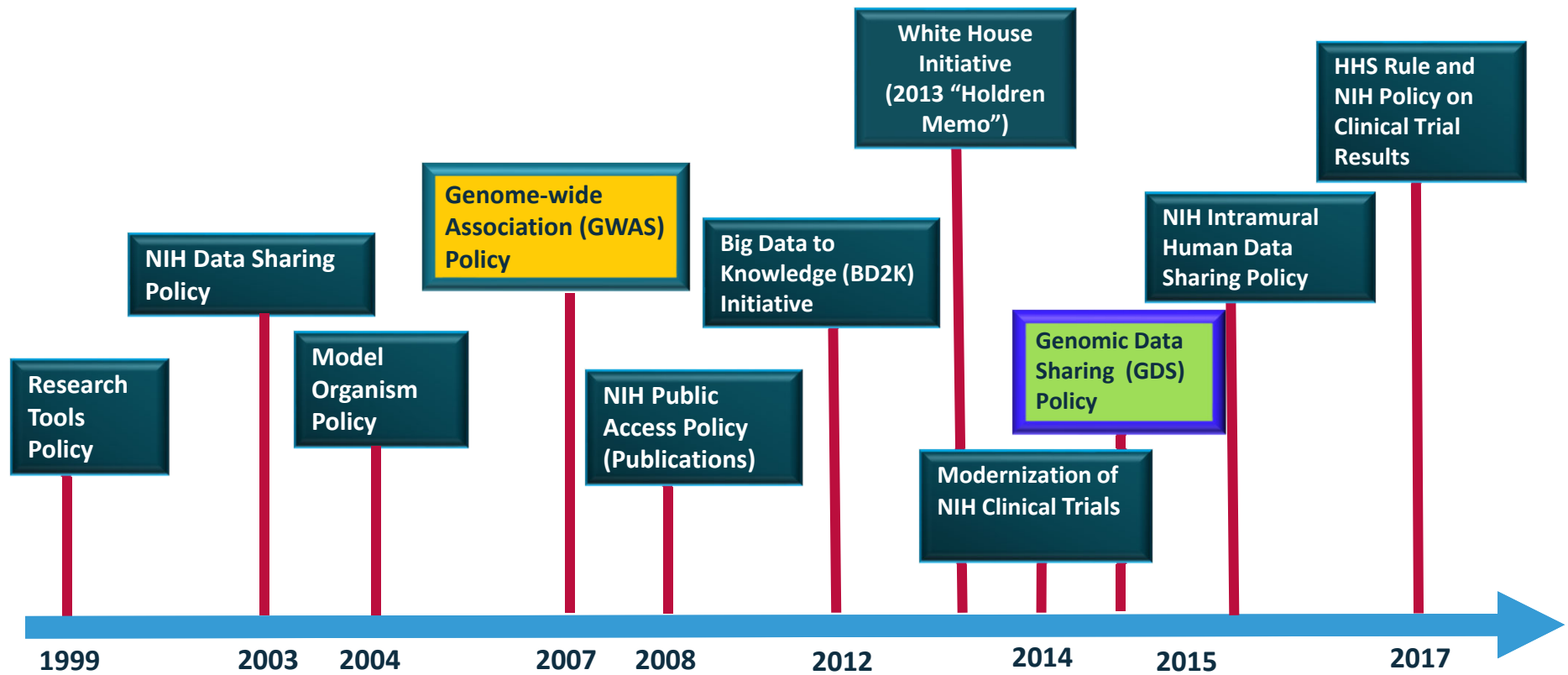
Policy on Dissemination of NIH-Funded Clinical Trial Information (01/2017)

The National Institutes of Health (NIH) is issuing this policy to promote broad and responsible dissemination of information from NIH-funded clinical trials through ClinicalTrials.gov. The policy establishes the expectation that all investigators conducting clinical trials funded in whole or in part by the NIH will ensure that these trials are registered at ClinicalTrials.gov, and that results information of these trials is submitted to ClinicalTrials.gov.

Key Points of Policy

- Applies to all intramural and extramural clinical trials funded wholly or partially by NIH of FDA-regulated drug, biological, and device products and pediatric post-market surveillance studies of devices required by the FDA under the FD&C Act. (Does not apply to phase 1 trials or small feasibility device studies).
- Trials, including data elements, must be registered on ClinicalTrials.gov no later than 21 days after enrollment of the first participant.
- Results information is to be submitted to ClinicalTrials.gov no later than 12 months after primary completion date; possible delay of up to an additional 2 years for trials of unapproved products or of products for which initial FDA marketing approval or clearance is being sought, or approval or clearance of a new use is being sought.
- For federally funded trials, grant funding can be withheld if required reporting cannot be verified. Civil monetary penalties of up to \$10,000/day (amount to be adjusted going forward)/ May lead to suspension or termination of grant or contract funding/ Can be considered in future funding decisions.

History of NIH Data Sharing Policy



Guiding Principle of the NIH Genomic Data Sharing (GDS) Policy

The greatest public benefit will be realized if large-scale genomic data are made available in a timely manner to the largest possible number of investigators. For human data, data are made available under terms and conditions consistent with the informed consent provided by individual participants.

Benefits/Rationale for GDS

- Enhance scientific progress and accelerate translation of genomic research into treatments, products and procedures that benefit public health
- Examine relationships between genomic data and phenotypes while respecting rights of research participants
- Unshared Information represents lost opportunity to improve public health
- Encourage data access and sharing unencumbered by intellectual property claims (discourage premature claims on pre-competitive information)
- Increased availability of data to a wide range of secondary data users not engaged in human subjects research (45 CFR 46)

The infographic is divided into two main sections: 'Genomics Fun Facts' on the left and 'Genome data' on the right. The text states: 'Sequencing **one person's genome** generates around **200GB** of data, the capacity of a typical computer'. To the right of the text is an illustration of a man in a suit standing next to a computer desk with a monitor displaying '200GB'. Above the man are floating letters representing DNA bases: G, C, A, T.

NIH Genomic Data Sharing Policy Details

- **Purpose**

- Sets forth expectations and responsibilities for investigators and their institutes that ensure the broad and responsible sharing of genomic research data in a timely manner (within 6 months of clean, QC'ed data)

- **Scope**

- All NIH-funded research generating large-scale human or non-human genomic data and the use of these data for subsequent research
- Applies to all funding mechanisms (grants, contracts, intramural support) and there is no minimum threshold for cost

- **Data Sharing**

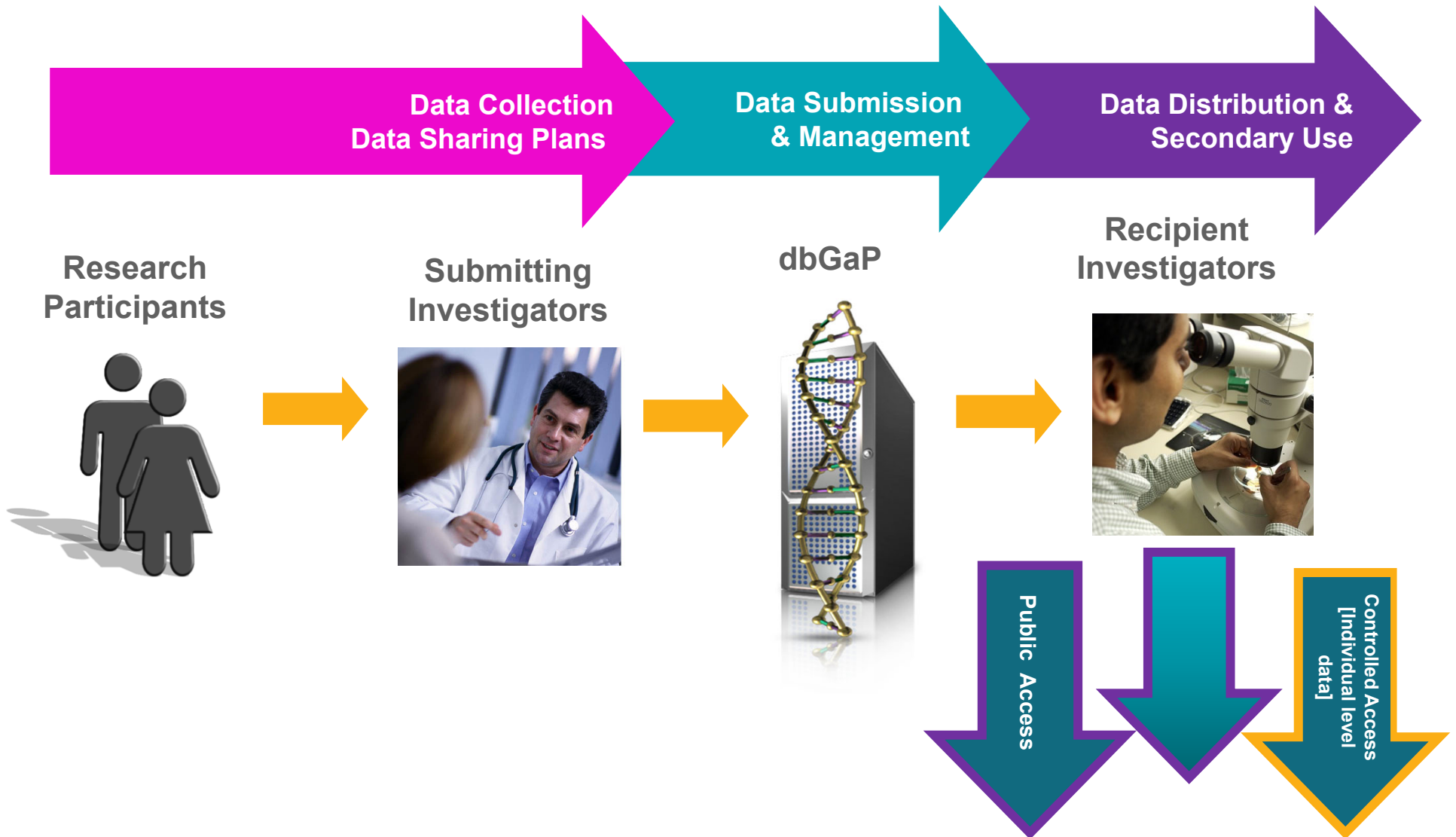
- *Non-human data*: made available through current databases and resources remain standard mechanism; any widely used data repository (e.g., GenBank, SRA, ZFIN)
- *Human data*: studies with data derived from human specimens registered in dbGaP

Unrestricted- vs Controlled-access to Human Genomic Data

- Informed consent is the basis for institutions to determine the appropriateness of submitting human data to unrestricted or controlled-access NIH data repositories*
- ***Unrestricted/ Open-access Tier***: data are publicly available to anyone (ie. The 1000 Genomes Project); includes study protocols, metadata, certain phenotype data, genomic summary results (sensitive populations may apply for all controlled-access).
- ***Controlled-access Tier***: investigators must obtain approval from NIH Data Access Committees to use the requested data (e.g., dbGaP); includes individual-level sequence data and potentially identifiable phenotype and analyzed data (**unless informed consent explicitly states unrestricted-access to individual-level human genomics data is appropriate*)



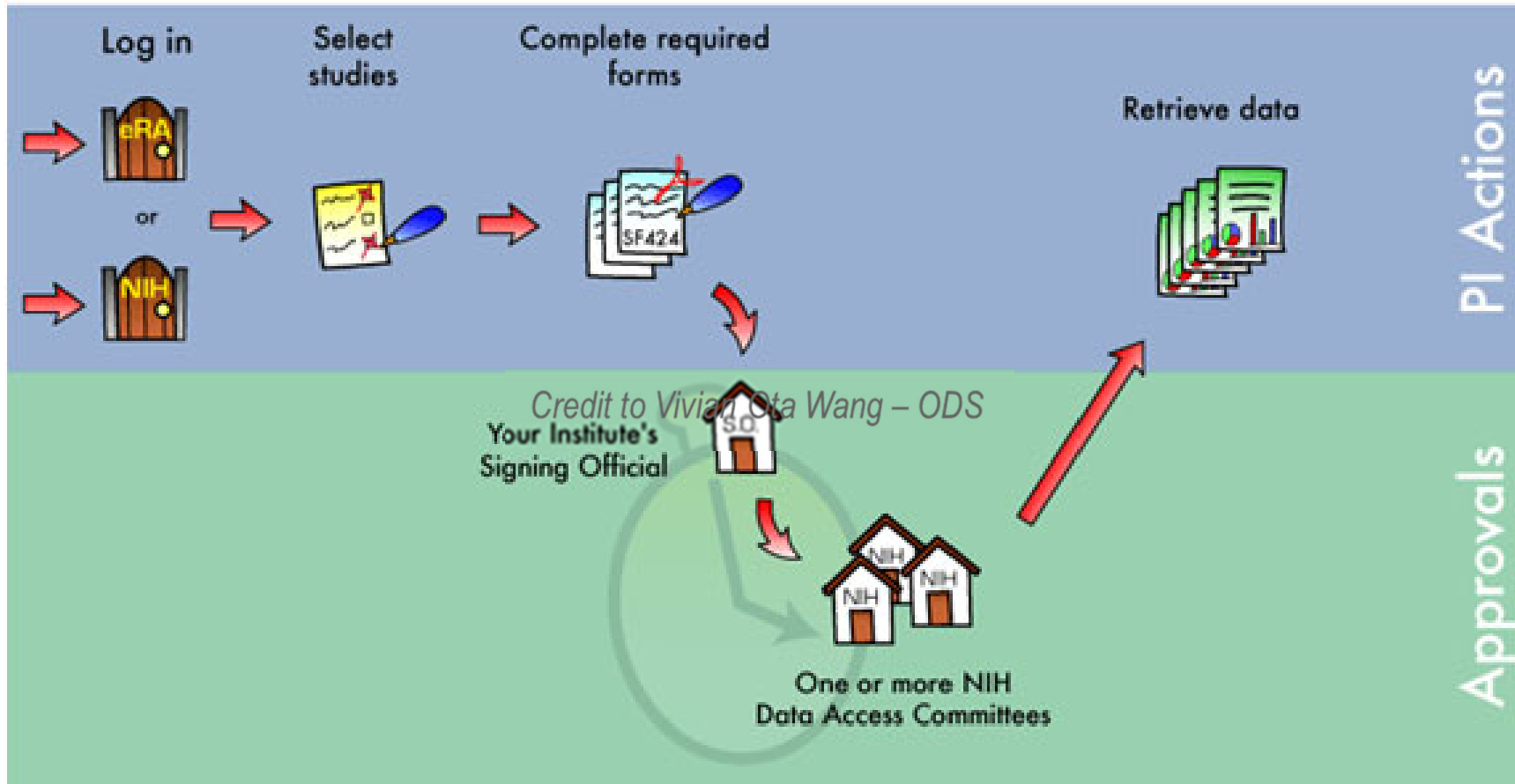
dbGaP Controlled-Data Management Process



NIH Data Sharing Plans (DSP)

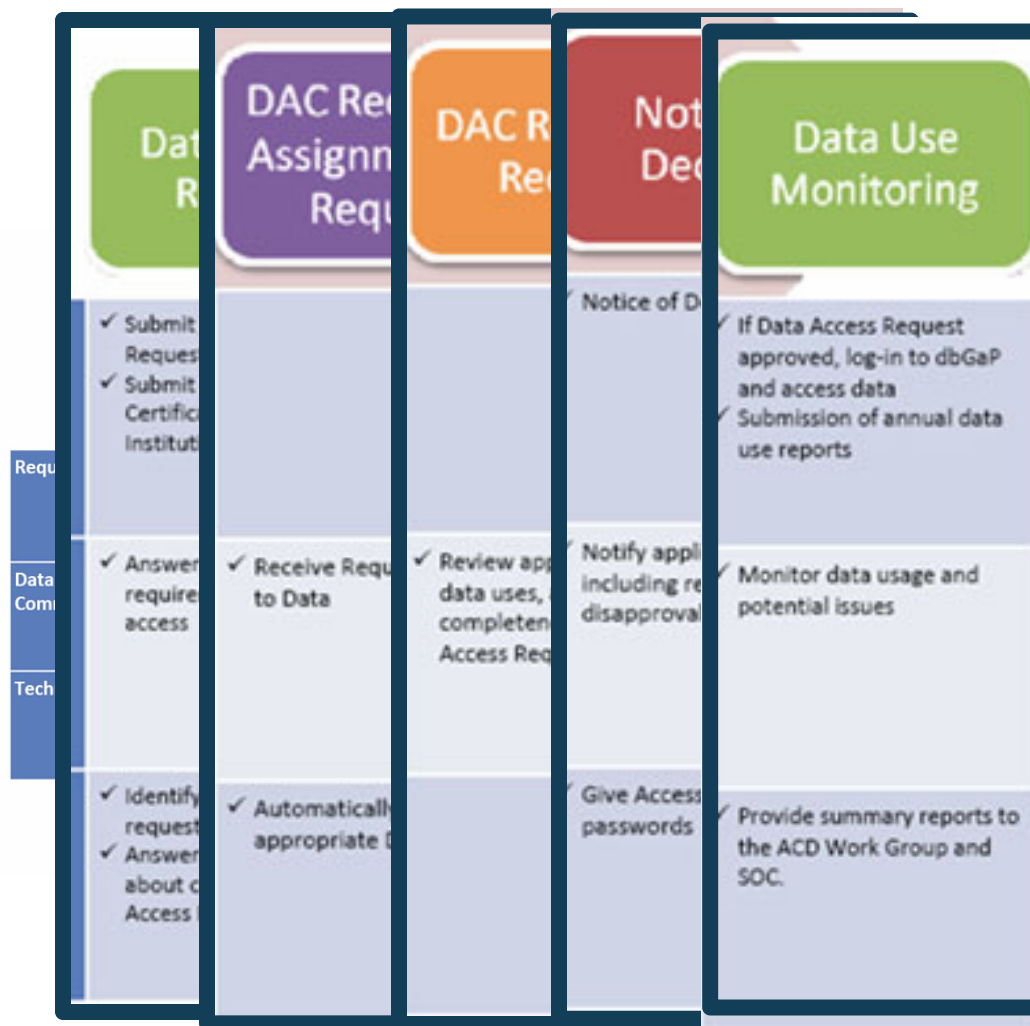
- Applicants should contact IC program staff prior to submission & include DSP with funding applications, as data sharing is specific by funding announcement
 - Program Announcements (PA) may request data sharing plans for applications that are less than \$500,000 direct costs in any single year.
 - Reviewers will not factor the proposed data-sharing plan into the determination of scientific merit or priority score.
 - Program staff will be responsible for overseeing the data sharing policy and for assessing the appropriateness and adequacy of the proposed data-sharing plan.
- DSP should state clearly how data and metadata will be shared for secondary use (including as many details as possible; data platforms, levels of data, associated clinical/ phenotype data)
- Data release for public secondary use is expected to be no later than the acceptance for publication of the main findings from the final data set.
 - ❖ *NIH continues to expect that the initial investigators may benefit from first and continuing use but not from prolonged exclusive use.*
- Grantees are free to choose their own NIH Data Repository, and the decision process can be referenced in the DSP
- If data cannot be shared in accordance with NIH policy, applicants must clearly outline the reasons and provide an alternate plan for program staff to consider.

The dbGaP Data Request Process



The dbGaP Data Request Process

- Requester submits Data Access Request (DAR) to Institutional Signing Official
- Signing Official (SO) approves and submits to Data Access Committee (DAC) staff
- DAC Staff DAR pre-review
- Full Data Access Committee (DAC) review
- Data Access Request (DAR) is approved or disapproved
- Requestor is notified by email of DAC decision
- Requestor downloads data
- Requester completes Annual Report
 - Renew access or closeout.



NIH Code of Conduct for Genomic Data Use

Investigators agree to:

- Use requested datasets **only** for the research described in their Data Access Request (DAR)
- Not distribute data to individuals not specified in their DAR
- Not attempt to contact or identify research participants
- Adhere to *dbGaP Best Practices that ensures data security*
- Report Data Management Incidents
- The proposed Research Use Statement is consistent with the Genomic Data Sharing Policy
- Not re-deposit data in public databases

**External collaborators must independently apply for data access.*

The screenshot shows the 'My Research Projects' page on the dbGaP website. At the top, there are navigation tabs: 'My Projects', 'My Requests', 'Downloads', 'Downloaders', and 'My Profile'. The main heading is 'My Research Projects' followed by 'General Instructions'. Below this, there are several bullet points providing instructions on how to use the application, including generating a Data Access Request (DAR) number, project-specific requests, and the need for approval for new projects. A section titled 'Before You Get Started' lists the information required to complete the application, such as a research statement, institutional signing official, list of investigators, and IRB approval. A note at the bottom of this section states: '* You can navigate to each study DUC from the public study home page in dbGaP. Look for the "individual-level data" section.' Below this is a red-bordered box containing the 'dbGaP APPROVED USER CODE OF CONDUCT'. This section explains that the following code of conduct applies to approved users and lists seven specific rules regarding data use, distribution, security, and reporting. At the bottom of the page, there is a green button labeled 'Begin New Research Project'.

My Projects My Requests Downloads Downloaders My Profile

My Research Projects

General Instructions

- This application will automatically generate a Data Access Request (DAR) number and a project number. Please keep track of this number for future communications with dbGaP and relevant Data Access Committee(s) (DAC)
- A completed request for data access includes this form as well as a review of and agreement to the terms, conditions, and statements in the Data Use Certification (DUC) for each respective dataset requested.
- Dataset requests are project-specific. If you were granted access to a dataset(s) for another project, that approval does not carry over to this new proposed project. You must request access to all datasets that you plan to use in the new project.
- Please note that fields marked as "*" are required fields.

Before You Get Started

In order to complete the application for data access you will need to collect the following information:

- A research statement and a nontechnical summary statement describing your planned use of the data.
- The name of the institutional signing official who will certify the terms of use assurances on behalf of your institution.
- A list of all internal investigators at your institution who will share access to the data for the proposed research.
- A list of external collaborating investigators.
- The name of the information technology (IT) Director.
- Some datasets may require local Institutional Review Board (IRB) approval for use. These are noted in the study list. Please check the individual study pages in dbGaP for these additional requirements.
- Some datasets may require supplemental documentation to accompany this standard application. Review the DUC* instruction pages for detailed information about how to prepare these materials in a single PDF file.

* You can navigate to each study DUC from the public study home page in dbGaP. Look for the "individual-level data" section.

dbGaP APPROVED USER CODE OF CONDUCT

The following is the Code of Conduct that research investigators agree to abide by as Approved Users of data received through the database of Genotypes and Phenotypes (dbGaP). Failure to abide by any term within this Code of Conduct may result in revocation of approved access to any or all datasets obtained through dbGaP.

The elements of the NIH Code of Conduct for Data Use include:

1. Investigator(s) will use requested datasets solely in connection with the research project described in the approved Data Access Request for each dataset;
2. Investigator(s) will make no attempt to identify or contact individual participants from whom these data were collected without appropriate approvals from the relevant IRBs;
3. Investigator(s) will not distribute these data to any entity or individual beyond those specified in the approved Data Access Request;
4. Investigator(s) will adhere to computer security practices that ensure that only authorized individuals can gain access to data files;
5. Investigator(s) will not submit for publication or any other form of public dissemination analyses or other reports on work using or referencing NIH datasets prior to the embargo release date listed for the dataset (or dataset version) on dbGaP;
6. Investigator(s) acknowledge the Intellectual Property Policies as specified in the Data Use Certification; and,
7. Investigator(s) will report any inadvertent data release in accordance with the terms in the Data Use Certification, breach of data security, or other data management incidents contrary to the terms of data access.

Begin New Research Project

What *IS* NIH Data Sharing

IC	Data Sharing Policy Name	Description of Data Sharing Policy	Repositories
NIH	NIH Data Sharing Policy	Expects investigators seeking more than \$500K in direct support in any given year to submit a data sharing plan with their application or to indicate why data sharing is not possible.	No specific repository listed
NIH	NIH Policy on Deposit of Atomic Coordinates into Structural Databases	NIH policy requires that atomic coordinates from X-ray crystallographic and nuclear magnetic resonance experiments that were supported by NIH grants be deposited into the appropriate structural database at the time of submission of a research article drawing conclusions from these data.	Protein Data Bank
NHGRI	ENCODE Consortia Data Release, Data Use, and Publication Policies	Requires resource producers to release primary data along with an initial interpretation, in the form of genome features, to the appropriate public databases as soon as the data is verified. Consortia members will also identify validation standards that will be applied in subsequent analyses of the data or with additional experimentation where appropriate. All data will be deposited to public databases, such as GenBank or the ENCODE/modENCODE Data Coordination Centers (DCCs) and these pre-publication data will be available for all to use.	ENCODE
NIH	Genomic Data Sharing Policy	Expects that large-scale genomic research data from NIH-funded studies involving human specimens, as well as non-human and model organisms, will be shared through a publicly available data repository. All studies with human genomic data should be registered in dbGaP , and the data should be submitted to an NIH-designated data repository . Non-human data may be submitted to any widely used data repository.	dbGaP (for registration) NIH-designated data repository (for data)
NHLBI	NHLBI Policy for Data Sharing from Clinical Trials and Epidemiological Studies	Encourages all applicants to include a plan to address data sharing or to state why data sharing is not possible. For studies that meet the the following criteria, applicants are are required to provide a data sharing plan, which will be reviewed and approved by the relevant NHLBI program official: a) research applications/proposals requesting \$500000 direct costs; b) research studies that have 500 or more participants c) ancillary studies based on NHLBI-funded parent studies d) applications/proposals submitted in response to FOAs that specify inclusion of data sharing plans; or e) other research studies deemed appropriate for data sharing by NHLBI program official investigators.	NHLBI data repository , BioLINCC
NIA	Alzheimer's Disease Genetics Sharing Plan	NIA policy in the area of human Alzheimer's disease genetics applies to all NIA funded research in this area regardless of cost. NIA follows the NIH GWAS Policy and extends NIA's existing policy on sharing data on Alzheimer's disease genetics to include secondary analysis of data resulting from a genome wide association study. It is the policy of the NIA that useful specimens and Associated Phenotypic Data for the genetics of late onset Alzheimer's disease be deposited at the National Cell Repository for Alzheimer's Disease (NCRAD) whenever possible. It is the policy of the NIA that all Genetic Data derived from NIA funded studies for the genetics of late onset Alzheimer's disease be deposited at the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) or another NIA approved site or both whenever possible. It is the policy of the NIA that all GWAS data, including secondary analysis data, derived from NIA funded studies for the genetics of late onset Alzheimer's disease be deposited at the NIH GWAS data repository (dbGaP) or another NIA approved site or both, wherever possible.	NCRAD , NIAGADS , dbGaP
NIA	Alzheimer's Disease Neuroimaging Initiative (ADNI) Data Sharing and Publication Policy	The ADNI Executive Committee and the NIA expect that ADNI deidentified data will be made available to the general scientific community within a very short timeframe. ADNI recommends full, open access of all de-identified ADNI imaging and clinical data to individuals who register with the ADNI and agree to the conditions in the "ADNI Data Use Agreement" and who undergo limited screening.	ADNI
NIAID	NIAID/DMID Data Sharing and Release Guidelines	Establishes general principles and specific guidelines for data release plans across NIAID/DMID Omics Centers including Genomic Sequencing Centers for Infectious Diseases (GSCID) and other NIAID-funded large scale Centers and projects. Indicates that plans should specify that genomic and other data types collected in NIAID-funded research will be submitted as rapidly as possible into publicly accessible and searchable international databases such as GenBank, dbGaP, the sequence read archive, the DMID Bioinformatics Resource Center, or other databases designated and approved by NIAID.	GSCID , DMID Bioinformatics Resource Center , Trace Archive or, as appropriate, to the Short Read Archive , dbGaP , dbSNP , BEI Resources Repository
NIAID	Data Sharing Guiding Principles for the NIAID/DMID Systems Biology Program	The NIAID/DMID Systems Biology Program (SBP) encourages center-wide joint sharing and analysis of data and can be accomplished by: 1) making raw data available to center investigators, including raw data where final analysis may not be complete, or 2) where feasible and to maximize information content generated by each center, analyses of samples should be performed with multiple -omics platforms, versus a single profiling technology. By SBP contract requirement, research data, protocols and computational and statistical models must be made freely and publicly available to the scientific community through the centers' websites or other public databases within four weeks of publication, or within one year of generation.	Systems Biology Program (SBP)
NIAID	Human Immunology Project Consortium Data Sharing Plan	HIPC investigators agree to deposit their data into the Immunology Database and Analysis Portal (ImmPort) system (www.Immport.org) according to a timeline determined together with the NIAID Program Officer for each study. To fulfill the HIPC data sharing objectives, the investigators will enter all study data and meta-data into ImmPort.	ImmPort

<https://grants.nih.gov/policy/sharing.htm>

What *IS* NIH Data Sharing

NICHD	Revised Resource Sharing Plan Instructions for Genetic Screens to Enhance Zebrafish Research and Enhancing Zebrafish Research with Research Tools and Techniques (PAR-08-138 and PAR-08-139)	Regardless of the amount requested, investigators are expected to include a brief 1-paragraph description of how final research data will be shared, or explain why data-sharing is not possible. Applicants are encouraged to discuss data-sharing plans with their NIH program contact. When preparing their Resource Sharing Plan, participants are strongly encouraged to contact the Zebrafish International Resource Center (ZIRC, http://zfin.org/zirc/home/stckctr.php) to discuss their plans for sharing resources created under their proposed application and to receive a cost estimate for deposition of materials at ZIRC. Plans to share materials generated by projects under the FOA through ZIRC, including but not limited to mutant fish, embryos, and sperm, genetic screens, mutagenesis protocols, mutagenesis vector constructs, and genetic and phenotypic data for all mutant strains, should include evidence/documentation of coordination with staff at the Resource. A reasonable time frame for periodic deposition of mutants, sperm, reagents, and data should be specified in the application and will be considered during the review of the plan for sharing.	Zebrafish International Resource Center (ZIRC)
NIDA	NIDA Data Sharing Policy	Requires data for all NIDA-funded human genetics studies to be available for sharing, independent of direct costs, membership in the NIDA Genetics Consortium, or the type of genetics data generated.	NIDA Genetics Consortium, NIDA Center for Genetic Studies Repository
NIDA	Clinical Trials Network Data Share Policy	The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers to expedite the translation of research results into knowledge, products and procedures to improve human health. Data sets for CTN protocols will be available after (1) the primary paper has been accepted for publication, or (2) the data is locked for more than 18 months, whichever comes first.	CTN Data Share Website
NIDDK	The Environmental Determinants of Diabetes in the Young	All investigators who receive TEDDY resources must agree to acknowledge the TEDDY Study and the NIDDK central repository. This approach is fully compliant with the NIH public data sharing policy. That policy states that the NIH expects and supports the timely release and sharing of final research data from NIH-sponsored studies for use by other investigators and that the definition of "the timely release and sharing" to be no later than the acceptance for publication of the main findings from the final data set.	NIDDK Repository
NIH	NDAR Grantees Data Sharing Policy	All data resulting from this autism-related NIH-funded research involving human subjects are expected to be submitted to the National Database for Autism Research (NDAR), along with appropriate supporting documentation to enable efficient use of the data.	NDAR
NINDS	Sharing Data via the Federal Interagency Traumatic Brain Injury Research (FITBIR) Informatics System	Investigators submitting FITBIR data are expected to: a) provide descriptive information about their studies, b) submit coded genotypic and phenotypic data to the FITBIR Informatics System; and c) submit a data submission for providing assurance that all data are submitted to the DOD and the NIH in accord with applicable laws and regulations, and that the identities of research participants will not be disclosed to the FITBIR Informatics System.	FITBIR
ORDR	The Collaboration, Education, and Test Translation (CETT) Program's Guidelines for Data Collection and Sharing	Specifies that deidentified clinical data will be submitted and stored at the NIH for future distribution for research purposes. To facilitate the widest access to data, CETT Collaborative teams agree to the following principles: a) follow de-identification procedures defined within the GWAS policy b) develop procedures and educational/informational documents and c) de-identified clinical data will be submitted and stored at the NIH for future distribution for research purposes.	dbGaP

https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

Talk Objectives



1. Define “Data Sharing” and Establish the importance of responsible and broad data sharing to advance disease knowledge and improve care
2. Outline current NIH data sharing policies (including GDS Policy) and dbGaP procedures.
3. **Discuss positive examples of and barriers to sharing and potential ways to overcome them**
 - **Lessons learned and potential solutions to barriers in broad & equitable sharing (Informed Consent, submission to public dBs)**
 - **Clinical data/phenotype, open data → “coded” data sets**
5. Introduce ways that the new NCI Office of Data Sharing is advocating for the proper balance of broad and open data sharing/access while respecting the right of patients to participate in and benefit from research as they see fit.

Precision Medicine is a Grand Challenge



Courtesy of P. Kuhn (USC)

It Requires:

- Deep biological understanding
- Advances in scientific methods
- Advances in instrumentation
- Advances in technology
- Advances in data management and computation

*Cancer Research and Care generate detailed **data** that are critical to create a learning health system for cancer*

NCI Signature Adult Genomics Initiative

NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

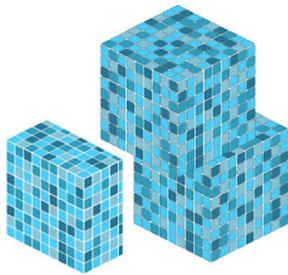
TCGA BY THE NUMBERS

TCGA produced over

2.5

PETABYTES

of data



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



TUMOR
SUBTYPES

Revolutionized how
cancer is classified



THERAPEUTIC
TARGETS

Identified genomic
characteristics of tumors
that can be targeted with
currently available
therapies or used to help
with drug development



A Data Sharing Platform to Promote Precision Oncology

The Genomic Data Commons

NCI Signature Precision Oncology Initiative

NATIONAL CANCER INSTITUTE NCI-MATCH CLINICAL TRIAL

THIS PRECISION MEDICINE TRIAL
EXPLORES TREATING PATIENTS
BASED ON THE MOLECULAR
PROFILES OF THEIR TUMORS

NCI-MATCH* IS FOR ADULTS WITH:

- solid tumors (including rare tumors) and lymphomas
- tumors that no longer respond to standard treatment



ABOUT 3,000
CANCER PATIENTS
WILL BE
SCREENED WITH A
TUMOR BIOPSY

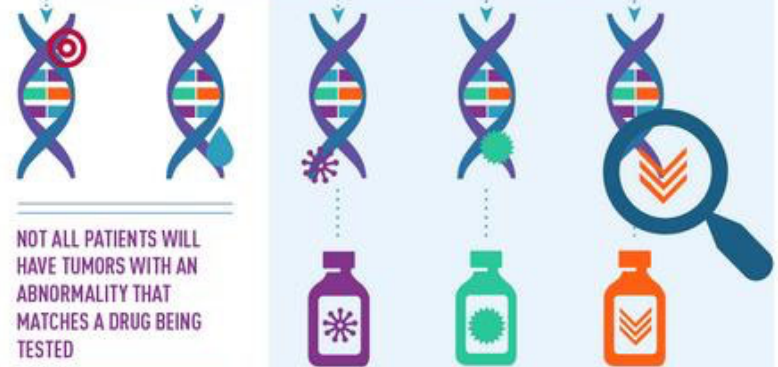


GENE SEQUENCING WILL LOOK FOR CHANGES IN 143 GENES

THE BIOPSIED
TUMOR TISSUE
WILL UNDERGO
GENE
SEQUENCING

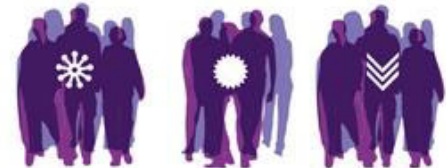


IF A PATIENT'S TUMOR HAS A GENETIC ABNORMALITY THAT MATCHES ONE TARGETED BY A DRUG USED IN THE TRIAL, THE PATIENT WILL BE ELIGIBLE TO JOIN THE TREATMENT PORTION OF NCI-MATCH



NOT ALL PATIENTS WILL
HAVE TUMORS WITH AN
ABNORMALITY THAT
MATCHES A DRUG BEING
TESTED

PATIENTS WITH TUMORS
THAT SHARE THE SAME
GENETIC ABNORMALITY,
REGARDLESS OF TUMOR
TYPE, WILL RECEIVE THE
DRUG THAT TARGETS
THAT ABNORMALITY



*NCI-Molecular Analysis for Therapy Choice

www.cancer.gov/nci-match
To learn more, call 1-800-4-CANCER

NCI National Clinical
Trials Network

Application of Cancer Genomics to Clinical Research

Content in the Genomic Data Commons

Current

- ❖ TCGA 11,353 cases
- ❖ TARGET 3,178 cases

Coming soon

- ❖ Foundation Medicine 18,000 cases
- ❖ Cancer studies in dbGaP ~4,000 cases
- ❖ Multiple Myeloma RF ~1,000 cases
- ❖ GENIE ~59,000 cases

Planned (1-3 years)

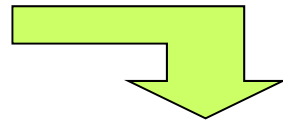
- ❖ NCI-MATCH ~3,000 cases
- ❖ Clinical Trial Sequencing Program ~3,000 cases
- ❖ Cancer Driver Discovery Program ~5,000 cases
- ❖ Human Cancer Models Initiative ~1,000 cases
- ❖ APOLLO – VA and DoD ~8,000 cases

~116,000 cases

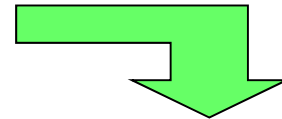
Strategy to Develop Novel Treatments for Childhood Cancers



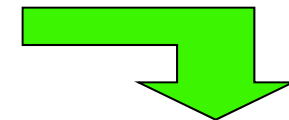
**Genomic
Discovery**



**Preclinical
Evaluation/
Data Mining**



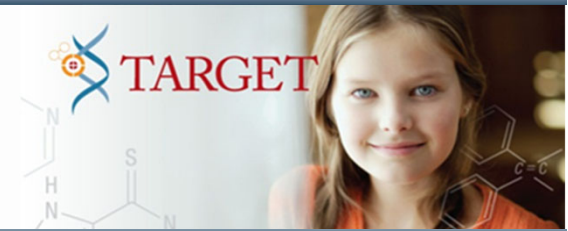
**Phase 1
Clinical Trial
(COG)**



**Definitive
Clinical Trial/
Standard Tx**

**Goal:* To rapidly identify viable molecular targets that will improve understanding and treatment of cancer.

TARGET Data Types & Platforms



Disease	Patient Data	Transcriptome Sequencing/ Gene Expression			DNA Sequencing/ Chr. Copy #			Methylation	Other Platforms
ALL Pilot	Y	mRNA-seq	U133A+2		WGS	WGS-lite	SNP 6.0		Targeted Seq
ALL P2	Y	mRNA-seq	U133A+2	miRNA-seq	WGS	WXS	SNP 6.0	HELP	
AML	Y	mRNA-seq	Gene ST	miRNA-seq	WGS	WXS	SNP 6.0	Infinium 27K/450K	Targeted Seq
AML-IF	Y	mRNA-seq		miRNA-seq	WGS			Infinium 450K	
NBL	Y	mRNA-seq	HuEx ST	miRNA-seq	WGS	WXS	Infinium 550K	Infinium 450K	Targeted Seq
OS	Y	mRNA-seq	HuEx ST	MegaPlex Taqman	WGS	WXS	SNP 6.0	Infinium 450K	Targeted Seq
WT	Y	mRNA-seq	U133A+2	miRNA-seq	WGS	WXS	SNP 6.0	Infinium 450K	Targeted Seq
CCSK	Y	mRNA-seq	U133A+2		WGS		SNP 6.0	Infinium 450K	
RT	Y	mRNA-seq		miRNA-seq	WGS			Methyl-seq	ChIP-seq
PPTP			U133A+2			WXS	SNP 6.0		

Translating Key TARGET Discoveries

Genomic Discovery

Clinical Translation

Ph-like Acute Lymphoblastic Leukemia

- Gene expression profile similar Ph+ ALL
- Poor outcome
- Frequency increases w/ age (prevalent in young adults)
- Kinase activating lesions in ~90% cases
- High frequency of rearrangements & fusions highly responsive to TK inhibitors
- COG ADVL1011/ AALL1521 → Phase I & Phase II trials of Janus kinase (JAK) inhibitor ruxolitinib in patients with known CRLF2 rearrangements or JAK pathway mutations
- COG AALL1131 Phase III → efficacy combination chemo with dasatinib for Ph-like ALL & ABL-class fusions

(Roberts, *NEJM*, 2014; Mullighan, *NEJM*, 2009; & others)

High-Risk Wilms Tumors

Results from key genetic mutations across a number of genes important in two major cellular processes that occur early in kidney development: one pathway regulates miRNA biogenesis and another interferes with normal maturation of the kidney (induction) by regulating gene transcription. (*Gadd, Nature Gen, 2017*)

Relapsed Favorable Histology WT

- Recurrent mutations in key miRNA processing genes (DGCR8, DROSHA) & SIX1/2 homeobox genes → higher rate relapse/death
- Activating MLLT1 mutation, early renal development → WT

Diffuse Anaplastic WT

- Confirm TP53 defining mutation develop anaplasia (>95%)

****Reduction in toxic tx for patients not likely to relapse through stratification of patients by 1q gain & loss of 16q,1p (new standard protocol for WT)**

Current Barriers to Data Sharing

❖ Inability to integrate data due to disparate consenting language and processes

- Large volumes of data being generated at a feverish pace without consistent formats or data and metadata standards
- Lack of searchable and interconnected data repositories with associated tools and services
- Lack of agreed upon ontologies, vocabularies, and data models severely impacts interoperability, integration, and analysis across multiple datasets
- Policy and procedural obstacles preventing patients and researchers from contributing their data to certain databases
 - ❑ Mandates and legal issues from funding sources (GDPR)
 - ❑ Lack of resources to format data and metadata files, and further submit them to databases
 - ❑ How to choose the best database to house the data
- Consent and data-use agreements (Electronic, trackable, machine-readable consents and terms-of-use agreements for data and other services to enable monitoring, computationally enforcing, and updating these agreements)

Patient Consent Language

NIH strongly encourages consent language that allows for future use and broad sharing of data without additional use restriction placed on it.

- Ideally this would be unrestricted or General Research Uses, which allows data to be combined and analyzed with any other dataset.
- Disease-specific references should be avoided, particularly if the language could be interpreted as exclusive use (ie. Permission for “kidney cancer” and related disorders cannot be easily interpreted nor combined with any other type of disease data); if necessary, say “must include disease....”
 - What and who define “related disorders”; is it by histology, developmental pathway, symptoms, treatment?
 - This data could not be combined with other disease research that might be beneficial including other kidney diseases
 - Genetic diseases may be related more by variation in a gene or pathway that could lead to treatment
 - Alternate disease groups provide ready control sets
- Modifiers that place additional restrictions on data use should be avoided (ie. Additional IRB consent for secondary use, restriction of methods development, collaboration required)

Current Barriers to Data Sharing

- Inability to integrate data due to disparate consenting language and processes
- **Large volumes of data being generated at a feverish pace without consistent formats or data and metadata standards**
- **Lack of searchable and interconnected data repositories with associated tools and services**
- **Lack of agreed upon ontologies, vocabularies, and data models severely impacts interoperability, integration, and analysis across multiple datasets**
- **Policy and procedural obstacles preventing patients and researchers from contributing their data to certain databases**
 - ❑ **Mandates and legal issues from funding sources (GDPR)**
 - ❑ **Lack of resources to format data and metadata files, and further submit them to databases**
 - ❑ **How to choose the best database to house the data**
- **Consent and data-use agreements (Electronic, trackable, machine-readable consents and terms-of-use agreements for data and other services to enable monitoring, computationally enforcing, and updating these agreements)**

The Beau Biden Cancer Moonshotsm

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
- Blue Ribbon Panel – October, 2016
 - Recommendations include:
 - 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

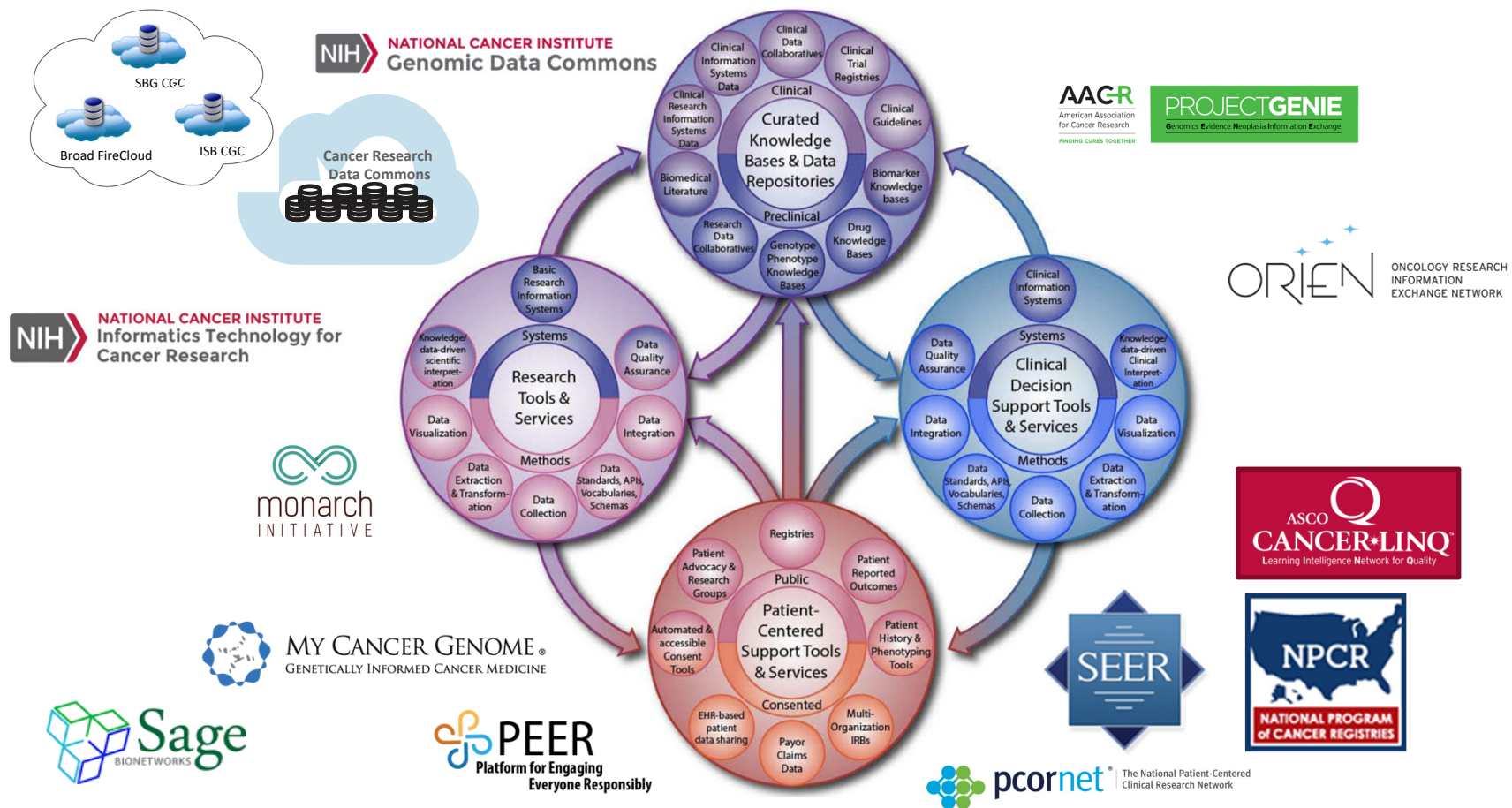
National Cancer Data Ecosystem Recommendations

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.”

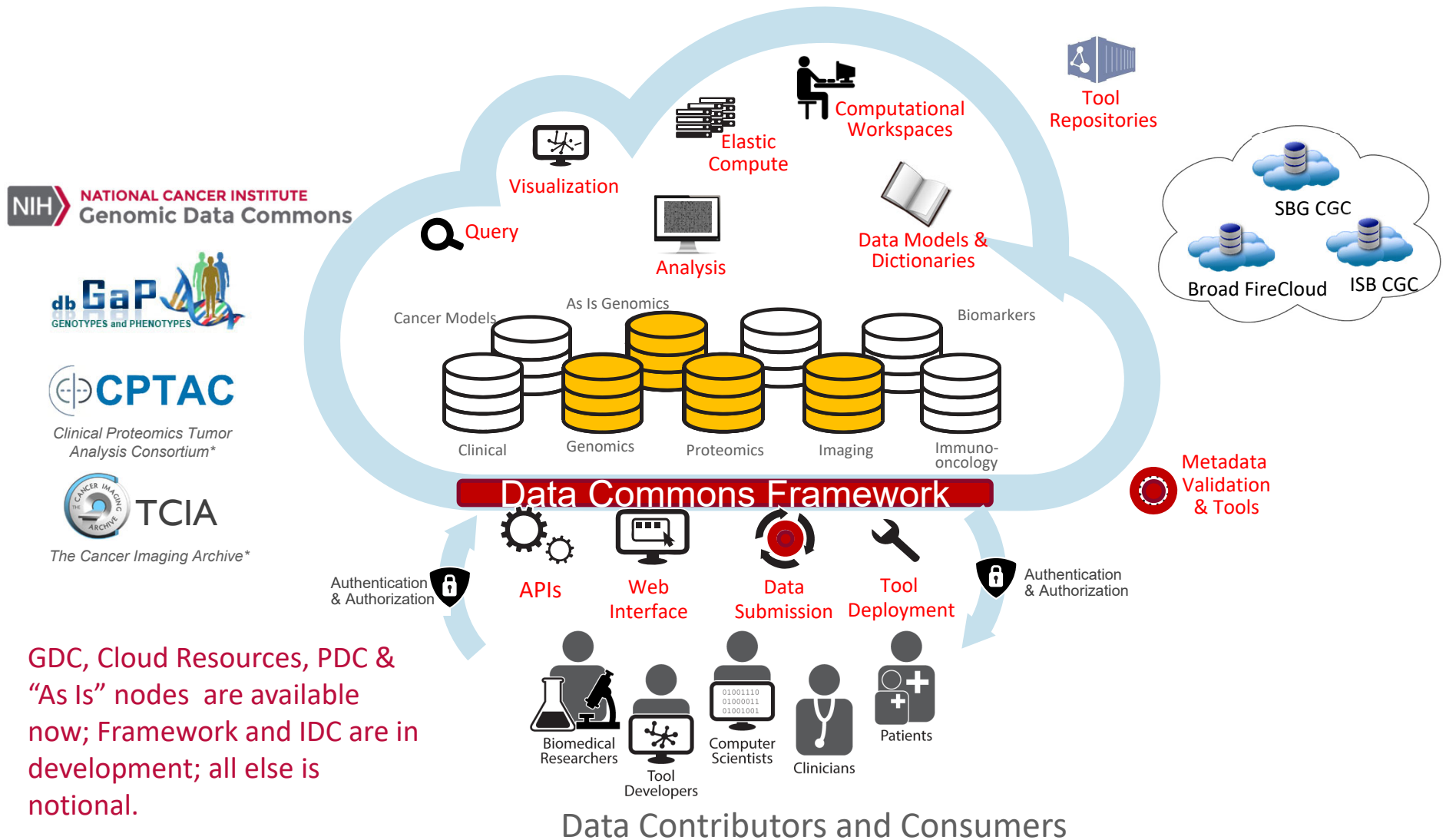
Recommendations

- **Build a National Cancer Data Ecosystem**
 - Enhanced cloud-computing platforms.
 - Essential underlying data science infrastructure and portals for the Cancer Data Ecosystem.
 - Services that link disparate information, including clinical, image, and molecular data.
 - Develop standards and tools so that data are interoperable.
 - Address sustainability and data governance to ensure long-term health of the Ecosystem.
- The National Cancer Data Ecosystem is broader than NCI
 - An NCI Cancer Research Data Commons is envisioned as part of the National Cancer Data Ecosystem

Enhanced Data Sharing Working Group Recommendation: *The Cancer Data Ecosystem*



NCI Cancer Research Data Commons



GDC, Cloud Resources, PDC & "As Is" nodes are available now; Framework and IDC are in development; all else is notional.

NCI Broaden Genomic Data Storage to...

- Take alignments and variant calls **as they are**
- Take study data model **as it is**
 - Rather than forcing a common model
- Take vocabulary **as it is**

- Obligation on investigator to communicate the above
- Do so by capturing what investigators are doing anyway
 - Create no additional burden

- Goal: Usable by those not engaged in the creation or production of the dataset
 - FAIR as a general principle
 - Ensure the context of the samples is captured

GDC: Data Retrieval

GDC Website

Data Portal

The screenshot shows the GDC Data Portal search interface. It includes a search bar, filters for Project, Primary Site, Cancer Program, and Data Type. A table of projects is displayed with columns for ID, Disease Type, Primary Site, Program, Count, and Available Cases per Data Type.

Data Transfer Tool

The screenshot shows the GDC Data Portal interface with a terminal window open. The terminal displays a command to download data for a specific case: `1 $ parcel udt -t token -m gdc_manifest 97a589423b4c1e15fd29a6cb58a6c6652c21`. The terminal output shows the download process, including file names and sizes.

The screenshot shows the GDC Website homepage. It features a navigation menu, a search bar, and a main content area with a "Case Distribution by Disease Type" pie chart and a "Data Availability Summary" table.

Visualization Tools

The screenshot shows a GDC visualization tool. It displays a genomic track for the TP53 gene with various mutations highlighted. Below the track is a table of mutations with columns for Sample ID, Cancer Study, Mutation, and other details.

GDC API

```
{
  "data": {
    "hits": [
      {
        "project_id": "TCGA-SKCM", "primary_site": "Skin"
      },
      {
        "project_id": "TCGA-PCCPG", "primary_site": "Nervous System"
      },
      {
        "project_id": "TCGA-LAML", "primary_site": "Blood"
      },
      {
        "project_id": "TCGA-CNTL", "primary_site": "Not Applicable"
      },
      {
        "project_id": "TCGA-UVM", "primary_site": "Eye"
      }
    ]
  }
}
```

Legacy Archive

The screenshot shows the GDC Legacy Archive interface. It includes a search bar, filters for Case, Case Submitter ID Prefix, Primary Site, Cancer Program, and Project. A table of cases is displayed with columns for Access, File Name, and Case Project.

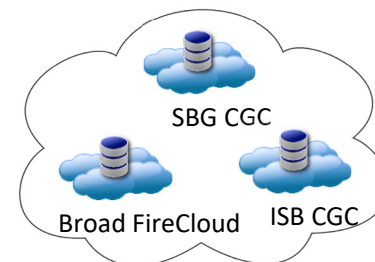
API URL Endpoint URL parameters Query

https://gdc-api.nci.nih.gov/projects?fields=project_id,primary_site

NCI Cloud Resources

Cloud Resources provide:

- Access to large genomic data sets without need to download
- Ability for researchers to bring their own tools and pipelines to the data
- Ability for researchers to bring their own data and analyze in combination with existing genomic data
- Workspaces, for researchers to save and share their data and results of analyses



- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

Data



- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Compute




- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

Security

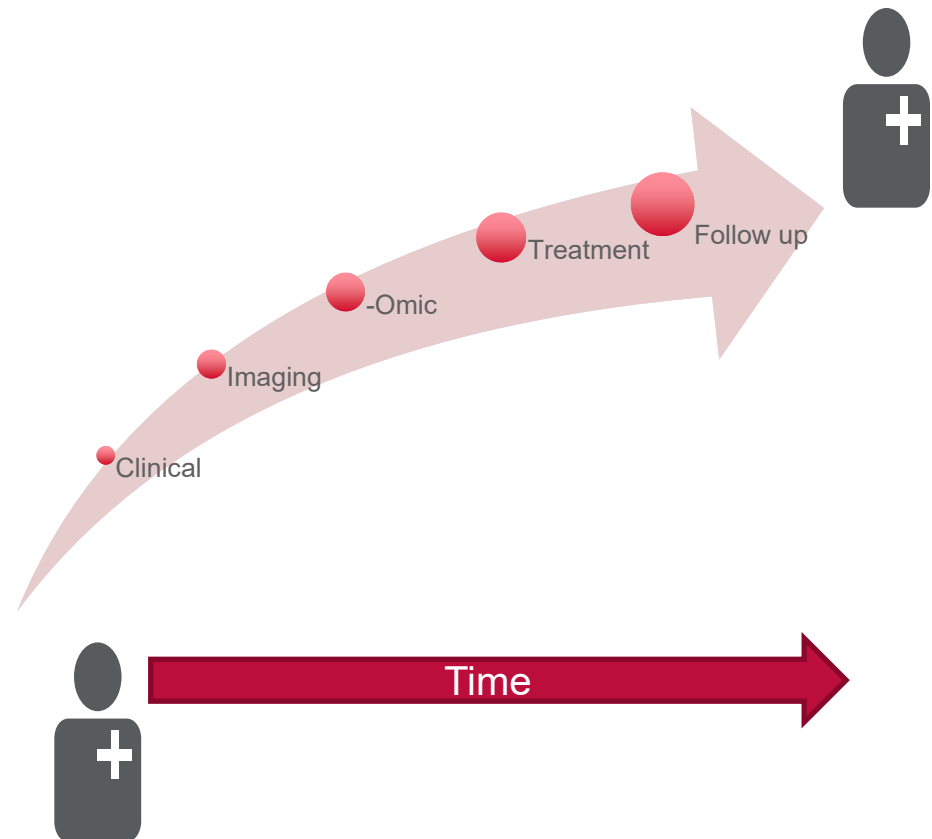


Democratize access to NCI-generated genomic and related data, and to create a cost-effective way to provide scalable computational capacity to the cancer research community.

 #NCICloud

Development of an Encrypted Unique Patient Identifier

- Pressing need to connect patient-level data across multiple data sources, data types and research studies—over time.
- Challenges include:
 - Protecting patient confidentiality
 - Consistency of identifying data (personally identifiable information, PII) available across diverse sources
 - Accuracy of linkage with varying PII
 - Scalability
- Encrypted hashed token
 - Allows linkage of diverse data.
 - Permits data sharing across multiple sources without release of PII.



NCI is Creating Partnerships

- Administrative supplements for Cancer Centers in GENIE and GA4GH coordination.
- Coordination with and support of Moonshot Programs
 - Assistance for U24 programs, e.g., Human Tumor Atlas & Immuno-oncology Data Coordinating Centers
- Work across related initiatives/programs
 - NCI, other NIH Institutes, NIH Data Commons Pilot Phase Consortium, All of Us, Chan Zuckerberg Initiative, GA4GH
- Workshops and RFIs to gather community input, feedback, and participation
- Establish CRDC governance process, including Scientific and Technical Advisory Board and Steering Committee.
- Establishing NCI Office of Data Sharing as a resource to NCI staff, external investigators and the broader research and participant communities.

Talk Objectives



1. Define “Data Sharing” and Establish the importance of responsible and broad data sharing to advance disease knowledge and improve care
2. Outline current NIH data sharing policies (including GDS Policy) and dbGaP procedures.
3. Discuss positive examples of and barriers to sharing and potential ways to overcome them
 - Lessons learned and potential solutions to barriers in broad & equitable sharing (Informed Consent, submission to public dBs)
 - Clinical data/phenotype, open data → “coded” data sets
4. **Introduce ways that the new NCI Office of Data Sharing is advocating for the proper balance of broad and open data sharing/access while respecting the right of patients to participate in and benefit from research as they see fit.**

Reasons why Investigators Resist Using dbGaP

- Difficulty of navigating the system (registration, submission, and access)
- Lack of understanding what data or metadata standards to use and how to analyze or integrate them
- Delayed time to obtaining approvals
- Fear of how data will be used or shared
- Mistrust of the government
- Federal rule

Immediate Issues to be resolved:

- Investigators are frustrated with the “dbGaP” processes, in general
 - Challenging submission process
 - Delayed time to obtaining approvals for access
 - Lack of understanding surrounding the processes

Reasons why Investigators Resist Using dbGaP

- Difficulty of navigating the system (registration, submission, and access)
- Lack of understanding what data or metadata standards to use and how to analyze or integrate them
- Fear of how data will be used or shared
- Mistrust of the government
- Federal rule

ODS Strategies Currently in Progress

- Centralize NCI DAC operations – create a single NCI DAC to provide a more efficient overall process approvals; increase consistency, decrease duplication and review time
- Streamline the Rejection/Revision Process to align across NIH ICs
- Expedite approvals for projects that request access to datasets with “General Research Use” & “Health, Medical & Biomedical” DULs
- Develop user-friendly, automated central interface to walk investigators through dbGaP registration, submission and access processes
 - Accessible by NIH and external accounts such that investigators, program staff and Institutional officials can complete tasks and communicate as necessary
 - Interoperates with existing NCI/NIH databases, tracking systems and data repositories
- Enhance communication within NCI/NIH and to external investigators to better educate each group regarding the process and the components of their various roles

ODS Strategies in Progress

Issues to be resolved:

- Wide spectrum of views on data sharing among patients, investigators, commercial entities
 - Extreme privacy to unrestricted access
 - How to ethically address wishes of all types of patient, investigator and consumer views?

Mission Strategies:

- Engage stakeholders (thought leaders, patient advocates, policy developers, PIs) to help refine NCI and NIH data-sharing strategies
- Implement an ELSI (Ethical, Legal & Social Issues)-type program for NCI
 - Include advocacy, outreach and community engagement across cancer research stakeholders (particularly noting health disparities among underserved individuals/communities)
 - Programmatic focus on data sharing issues
- Work to create and innovate around a “healthy” commercial marketplace that includes less restrictive business models (e.g. not based on “controlled” access cancer research/care data)

ODS Acknowledgements

- CBIIT Leadership & ODS Personnel

Anthony Kerlavage

Vivian Ota Wang

Freddie Pruitt

Sylvia Gayle

CIB Staff

- NIH OSP Personnel

- NIH Data Sharing & Policy Committee Members

- NCI Center for Cancer Genomics Staff



How Can this work for members of HRA?

- Do you allow your recipients to use funds to cover data-sharing costs? If yes, how much? is there a limit? Do you grant extra money on top of the grant?
- Do you check on compliance? What happens if they do not comply?
- What challenges did you face when developing and implementing the policy? How did you address those?
- What guidance can you give nonprofit nongovernmental funders (with fewer resources!) about how to develop a policy that works for “their” organization. Then how to market that to your scientists – or is mandating it enough?
- How does the community pay (funders pay, the scientists themselves??) for the myriad of costs incurred during the process of sharing data.

Additional Resources

- **Websites:** <https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/>
 - NIH OSP – <http://osp.od.nih.gov/>
 - NCI ODS – <https://cbit.cancer.gov/data-sharing>
 - NCI GDS information – <https://www.cancer.gov/grants-training/policies-process/nci-policies/genomic-data>

- **For General Inquiries:**
NCIofficeofdatasharing@nih.gov

- **Subscribe to NIH OD LISTSERVs:**
GDS: GENOMIC-DATA-SHARING-GDS-L
OSP: LISTSERV@list.nih.gov
(with the message: **Subscribe OSP_News**)



Bringing Science Policy Into Focus

Learn more about the Office of Science Policy from our blog “Under the Poliscope”

<http://osp.od.nih.gov/under-the-poliscope>

Additional Resources

- **dbGaP FAQs**

- <https://www.ncbi.nlm.nih.gov/books/NBK5298/>

- **dbGaP YouTube tutorial**

- https://www.youtube.com/watch?annotation_id=annotation_747461745&feature=iv&src_vid=-3tUBeKbP5c&v=m0xp_cCO7kA

- **dbGaP Help Contact** – dbgap-help@ncbi.nlm.nih.gov

- **eRA Commons** – <https://public.era.nih.gov/commons>

- **Genomic Data Sharing (GDS) Policy** – <https://osp.od.nih.gov/scientific-sharing/policies/>



**NATIONAL
CANCER
INSTITUTE**

www.cancer.gov/ccg