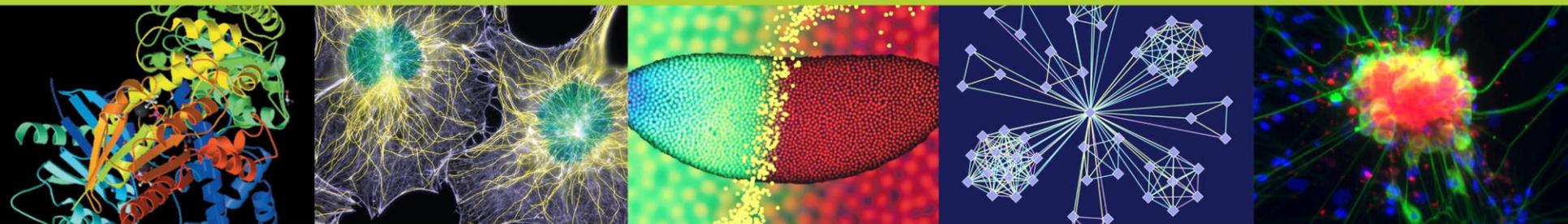


Open Science: Driving Forces and Practical Realities...for Biomedical Research

Susan K. Gregurick, Ph.D.

National Institute of General Medical Sciences

November 12, 2013



National Institutes of Health Mission and Goals

NIH's mission is to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability.

The goals of the agency are:

- To foster fundamental creative discoveries, innovative research strategies, and their applications as a basis for ultimately protecting and improving health;
- To develop, maintain, and renew scientific human and physical resources that will ensure the Nation's capability to prevent disease;
- To expand the knowledge base in medical and associated sciences in order to enhance the Nation's economic well-being and ensure a continued high return on the public investment in research; and
- To exemplify and promote the highest level of scientific integrity, public accountability, and social responsibility in the conduct of science.

Moving from Goals to Outcomes

In realizing these goals, the NIH provides leadership and direction to programs designed to improve the health of the Nation by conducting and supporting research:

- In the causes, diagnosis, prevention, and cure of human diseases;
- In the processes of human growth and development;
- In the biological effects of environmental contaminants;
- In the understanding of mental, addictive and physical disorders; and
- In **directing programs for the collection, dissemination, and exchange of information in medicine and health**, including the development and support of medical libraries and the training of medical librarians and other health information specialists.



Open Science

The umbrella term of the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional

Open Access to:

- Scientific Data and Metadata
 - Quality assurances and privacy concerns
- Scientific Analysis Tools and Methods
 - workflows
- Resulting Information and Publications

Snapshot of Open Science Activities at NIH

- **NIH Data Sharing Policy:**

All investigator-initiated applications with direct costs greater than \$500,000 in any single year will be expected to address data sharing in their application.

http://grants.nih.gov/grants/policy/data_sharing/

- **Overview of NIH Public Access Policy:**

The [NIH Public Access Policy](#) ensures that the public has access to the published results of NIH funded research. It **requires** scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive [PubMed Central](#) *upon acceptance for publication*. To help advance science and improve human health, the Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

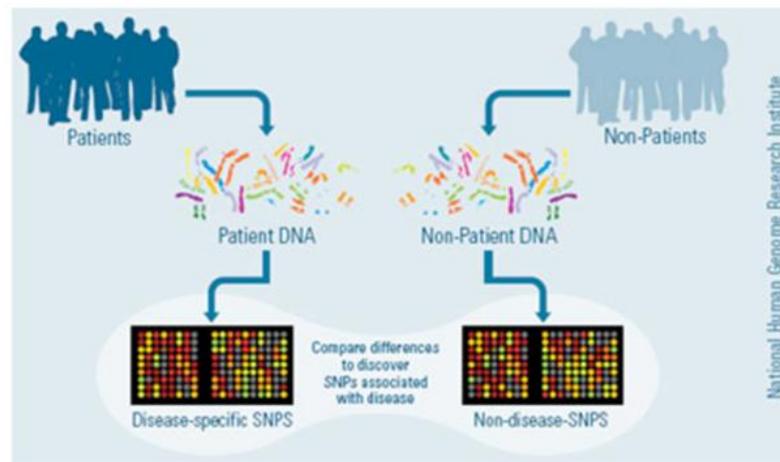
<http://publicaccess.nih.gov/>

Genome Wide Association Studies (GWAS)

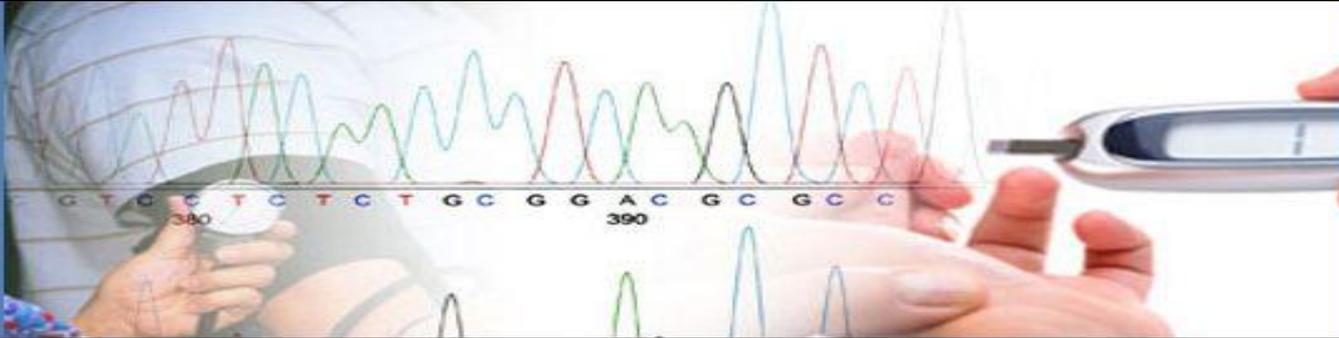
All investigators who receive NIH support to conduct genome-wide analysis of genetic variation in a study population are expected to submit to the NIH GWAS data repository descriptive information about their studies for inclusion in an open access portion of the NIH GWAS data repository. All data and information will be submitted to a high security network within the NCBI through a secure transmission process. Submissions should include the following:

- the protocol,
- questionnaires,
- study manuals,
- variables measured, and
- other supporting documentation

GWAS



Draft Genomic Data Sharing Policy



Overview of the Policy

Advances in DNA sequencing technologies, as well as a steep drop in sequencing costs, have enabled NIH to fund research that generates a greater volume and wide range of genomic data. In light of these developments, on September 20, 2013, NIH released a draft *Genomic Data Sharing Policy* (GDS Policy) for public comment

NIH considers access to such data particularly important because of the opportunities to accelerate research through the power of combining large and information-rich datasets.

Highlights of the GDS Policy

- The draft GDS Policy applies to all NIH-funded research that involves human and non-human genomic data produced by array-based or high-throughput sequencing technologies, such as GWAS, whole-genome, transcriptomic, epigenomic, and gene expression data, irrespective of the funding level and funding mechanism (i.e., grant, contract, or intramural support).
- The draft GDS Policy describes the responsibilities of investigators and institutions for the submission of non-human and human genomic data to data repositories and the secondary research use of such data as well as expectations regarding intellectual property.
- When data sharing involves human data, the protection of research participant privacy and confidentiality is paramount, and the draft GDS Policy reflects NIH's continued commitment to responsible data stewardship. The draft GDS Policy includes a number of provisions to assure the protection of human data.

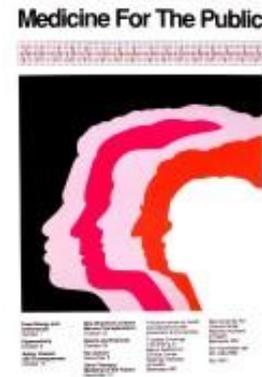
Public Comment Encouraged by November 20, 2013, <http://gds.nih.gov/survey.aspx>

National Library of Medicine

Founded in 1836, NLM is the world's largest biomedical library and NLM supports and conducts research and training in biomedical informatics and health information technology.

Resources include:

- PubMed Central
- MeSH
- UMLS
- ClinicalTrials.gov
- MedlinePlus
- TOXNET
- Images from the history of medicine
- LocatorPlus
- Other NLM Databases maintained by NCBI



National Center for Biotechnology Information

As a national resource for molecular biology information, NCBI's mission is to develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease.



NCBI:

- Maintains the largest collection of interlinked data and databases, tools and educational resources for Genomic Sciences
- Conducts research in fundamental biomedical problems at the molecular level using mathematical and computational methods
- Promotes and fosters scientific communication and cross-training in diverse areas such as computer science, molecular biology and genetics

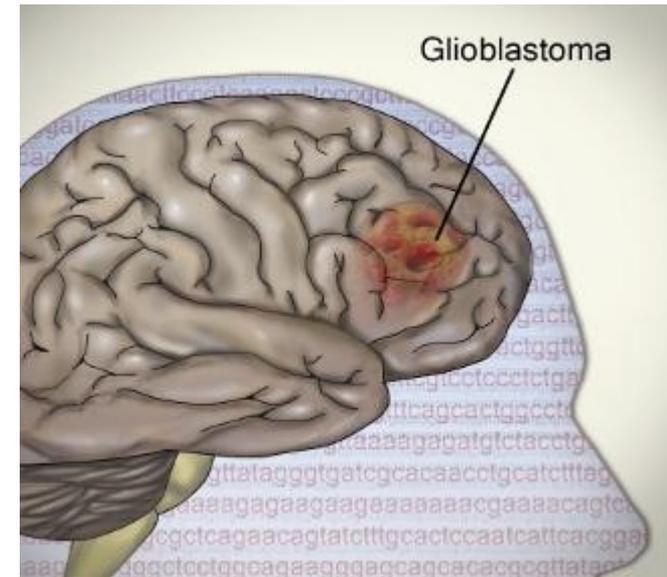


➤ **Each day over 3 million users access 40 interlinked genomic and bibliographic databases and download up to 53 trillion bytes of data**



The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.



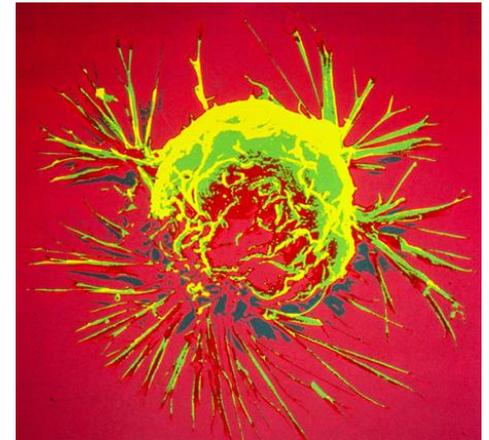
TCGA Data Portal Overview

<https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp>

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes.

The Cancer Imaging Archive (TCIA)

Provides a freely accessible, **open archive** of cancer-specific medical images and metadata accessible for public download. A huge amount of clinical and research images are collected each year. TCIA organizes and catalogs the images so that they may be used for a variety of purposes including:



- **Cancer researchers** can use this data to test new hypotheses and develop new analysis techniques to advance our scientific understanding of cancer.
- **Engineers and developers** can build new analysis tools and techniques using this data as test material for developing and validating algorithms.
- **Professors** can use it as a teaching tool for introducing students to medical imaging technology and cancer phenotypes.
- **The general public** can see how cancer appears in diagnostic images and learn about the instruments doctors use to diagnose cancer and measure the success of treatment.

<http://imaging.cancer.gov/informatics/thecancerimagingarchive>

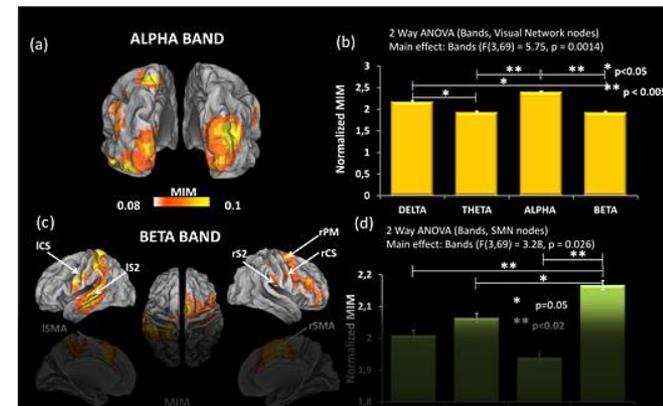
Neuroimaging Tools and Resources Clearinghouse

An open-source clearinghouse for software tools, data and other resources for functional and structural neuroimaging analysis. (www.nitrc.org)

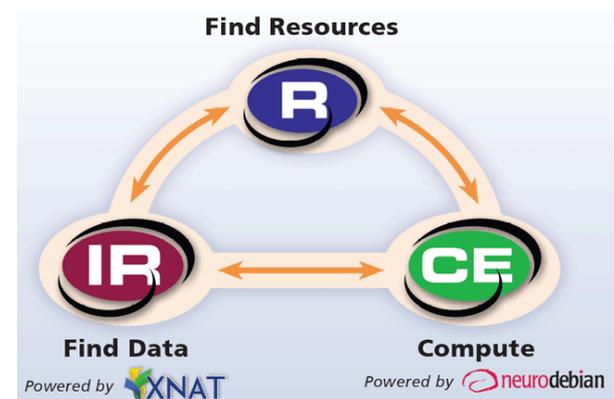
- Data available on NITRC includes brain images from MRI, PET, MEG and other types of brain scans
 - 600+ tools/resources, 4,859 imaging

Cloud enabled computing

- Researcher can compute against data via cloud-based workflow tools (such as best-of-breed neuroimaging workflows or pipelines)
- Pay as you go, and for only what you need for computing
- Released on Amazon's AWS Marketplace and as a public AMI (Jan 2013)



Measuring Interactions Across Resting State Networks. MEG study uses multivariate interaction measures to capture interactions within and across resting state networks over time



The BIOMEDICAL INFORMATICS RESEARCH NETWORK

BIRN provides a user-driven, software-based framework for research teams to share significant quantities of data – rapidly, securely and privately – across geographic distance and/or incompatible computing systems across more than 20 institutions



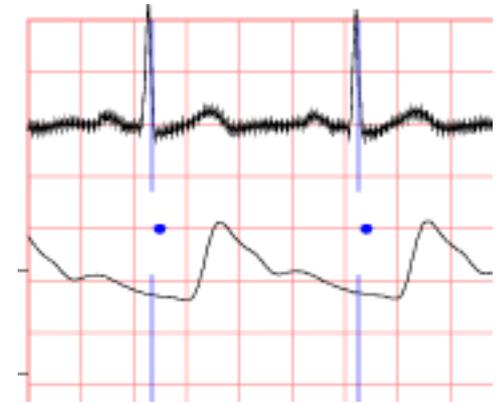
Current Capabilities include:

- Data Grid
- Credential Management
- User Registration Services/Certificate Authority
- Query Mediator
- Workflows for Computational Genomics and Visual Informatics
- BioScholar

<http://www.birncommunity.org/>

Research Resources for Physiologic Signals

PhysioNet offers free web access to large collections of recorded physiologic signals ([PhysioBank](#)) and related open-source software ([PhysioToolkit](#)). Each month, about 45,000 visitors worldwide use PhysioNet, retrieving about 4 terabytes of data



<http://physionet.org/>

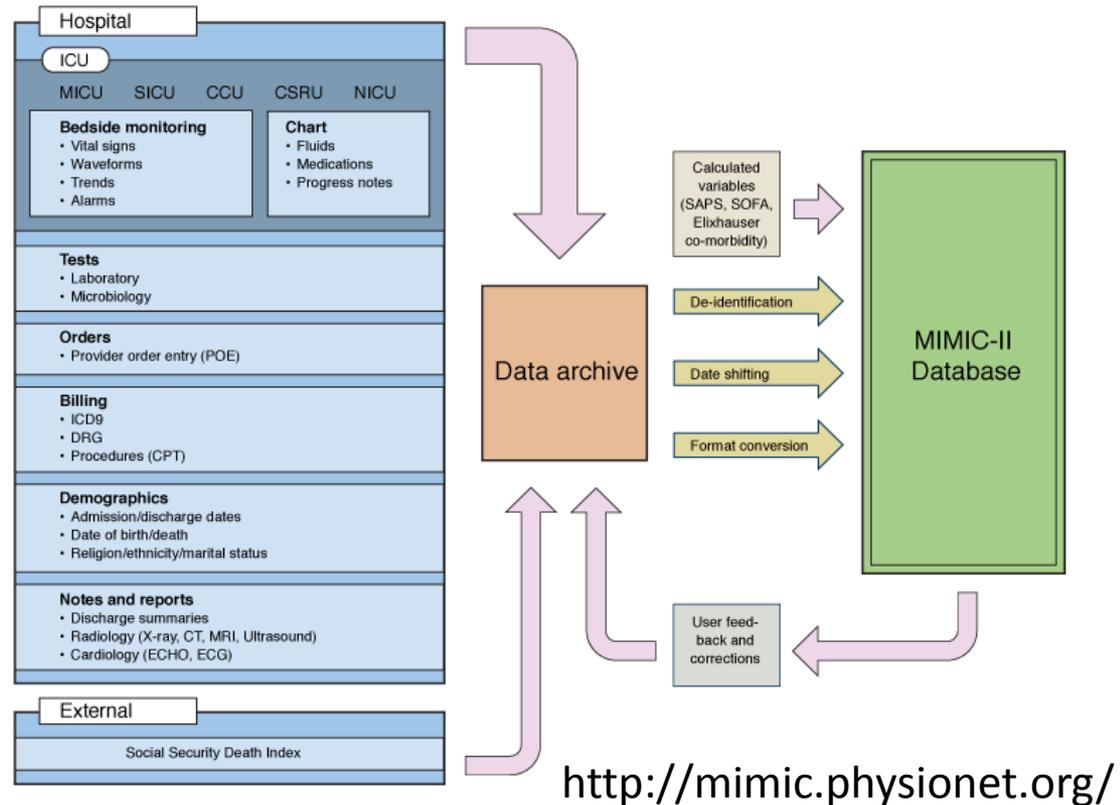
- **PhysioBank** currently includes databases of multi-parameter cardiopulmonary, neural, and other biomedical signals from healthy subjects and patients with a variety of conditions with major public health implications, including sudden cardiac death, congestive heart failure, epilepsy, gait disorders, sleep apnea, and aging.
- **PhysioToolkit** is a large library of software for physiologic signal processing and analysis, detection of physiologically significant events. **All PhysioToolkit software is available in source form under the GNU General Public License (GPL).**

Database for Intensive Care Units

The Multiparameter Intelligent Monitoring in Intensive Care ([MIMIC-II](#)) database: a massive research database from more than 30,000 ICU patients.

Interdisciplinary research partnership from:

- MIT
- Philips Medical Systems and Philips Research North America
- Beth Israel Deaconess Medical Center



Research activities for developing and evaluating advanced ICU patient monitoring and decision support systems that will improve the efficiency, accuracy, and timeliness of clinical decision-making in critical care.

Medical Devices Plug-n-Play (MD PnP) Interoperability Program

<http://www.mdnpn.org/>

- Open source effort to create interoperability of networked medical devices for clinical care.
- Development and support of [open standards](#) (e.g. [ASTM F2761](#), Integrated Clinical Environment, “ICE”).
- Devices source codes on GitHub



- Implementation of prototype use cases in an open “sandbox” environment
- Gov’t Partners: NIH, DoD, NSF, FDA, VHA and NIST
- Academic partners: MGH, UPenn, Kansas State, JHU, UIUC
- Industrial partners: DocBox, Anakena, Kaiser Permanente



NIH Data Sharing Repositories

This table lists NIH-supported data repositories that accept submissions of appropriate data from NIH-funded investigators (and others). Also included are resources that aggregate information about biomedical data and information sharing systems. The table can be sorted according by name and by NIH Institute or Center and may be searched using keywords so that you can find repositories more relevant to your data. Links are provided to information about submitting data to and accessing data from the listed repositories. Additional information about the repositories and points-of-contact for further information or inquiries can be found on the websites of the individual repositories.

Show entries

Search:

IC ▲	Repository Name ▼	Repository Description	Data Submission Policy	Access to Data
NCI	The Cancer Imaging Archive (TCIA)	The Cancer Imaging Archive (TCIA) is a large archive of medical images of cancer accessible for public download. All images are stored in DICOM file format. The images are organized as "Collections", typically patients related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc) or research focus.	How to Submit Data to TCIA	How to Access TCIA Data
NCI (NHGRI, NIGMS)	PeptideAtlas	PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences.	How to Submit Data to PeptideAtlas	How to Access PeptideAtlas Data
NHGRI	FlyBase: A Drosophila Genomic and Genetic Database	Drosophila Genomic and Genetic database that includes proteomics data, microarrays and Tiling BAC's.	How to Submit Data to FlyBase	How to Access FlyBase Data
NHGRI	The Zebrafish Model Organism Database (ZFIN)	ZFIN serves as the zebrafish model organism database. It aims to: a) be the community database resource for the laboratory use of zebrafish, b) develop and support integrated zebrafish genetic, genomic and developmental information, c) maintain the definitive reference data sets of zebrafish research information, d) to link this information extensively to corresponding data in other model organism and human databases, e) facilitate the use of zebrafish as a model for human biology, and f) serve the needs of the research community.	How to Submit Data to ZFIN	How to Access ZFIN Data

Crowdsourcing Science: 2012 NCI-DREAM Challenge in Drug Sensitivity Prediction

DREAM

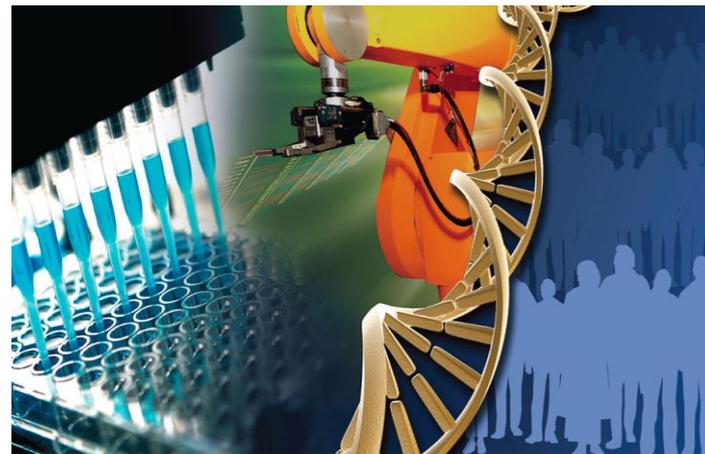


- DREAM is a Dialogue for Reverse Engineering Assessments and Methods. The main objective is to catalyze the interaction between experiment and theory in the area of cellular network inference and quantitative model building in systems biology.
 - The challenge is to use genomic information to build models capable of ranking the sensitivity of cancer cell lines to a set of small molecule compounds or their combinations
 - **Predict the sensitivity of breast cancer cell lines to previously untested compounds**
 - **Predicting compound combinations that have a synergistic effect in reducing viability of a DLBCL cell line**
- Open source codes and standardized computational infrastructure
 - All models behavior and performance reproducible

Computational DREAM Challenge in Toxicogenetics

(<http://www.niehs.nih.gov/funding/challenges/>)

National Institute of Environmental Health Sciences (NIEHS), the National Center for Advancing Translational Sciences (NCATS), University of North Carolina (UNC) and Sage Bionetworks



Goal: *Obtain a greater understanding about how a person's individual genetics can influence cytotoxic response to exposure to widely used chemicals*

Data set: 1086 human lymphoblastoid cell lines, representing 9 distinct geographic subpopulations, were treated with 179 pharmaceutical and environmental chemicals

1. Use the data to develop a model that accurately predicts individual responses to compound exposure based on genomic information
2. Use the data to develop a model that accurately predicts how a particular population will respond to certain types of chemicals.

New Initiatives at NIH

Common Data Element



Many ICs identifying common data element Goals:

- Standardized data elements, case report forms, instruments, lab tests, etc.
- For use in funded clinical research, patient registries, surveillance
- Facilitate cross-study comparison, aggregation of data
- Improve data quality via validated measures
- Most entail a vetting processes for selection



<http://cde.nih.gov>



Encourage Usage: CDE Template Language

Use of Common Data Elements in NIH-funded Research

- *NIH **encourages** the use of common data elements (CDEs) in basic, clinical, and applied research, patient registries, and other human subject research to facilitate broader and more effective use of data and advance research across studies.*
- *CDEs are data elements that have been identified and defined for use in multiple data sets across different studies. Use of CDEs can facilitate data sharing and standardization to improve data quality and enable data integration from multiple studies and sources, including electronic health records.*
- *NIH ICs have identified CDEs for many clinical domains (e.g., neurological disease), types of studies (e.g. genome-wide association studies (GWAS)), types of outcomes (e.g., patient-reported outcomes), and patient registries (e.g., the Global Rare Diseases Patient Registry and Data Repository).*
- *NIH has established a “Common Data Element (CDE) Resource Portal” (<http://cde.nih.gov/>) to assist investigators in identifying NIH-supported CDEs when developing protocols, case report forms, and other instruments for data collection. The Portal provides guidance about and access to NIH-supported CDE initiatives and other tools and resources for the appropriate use of CDEs and data standards in NIH-funded research.*
- *Investigators are **encouraged** to consult the Portal and describe in their applications any use they will make of NIH-supported CDEs in their projects.*



Home

NIH encourages the use of common data elements (CDEs) in clinical research, patient registries, and other human subject research in order to improve data quality and opportunities for comparison and combination of data from multiple studies and with electronic health records. This portal provides access to NIH-supported CDE initiatives and other tools and resources that can assist investigators developing protocols for data collection. [What is a CDE?](#)

NIH CDE Initiatives

Collections of CDEs that have been identified for use in particular NIH-supported research projects or registries after a formal evaluation and selection processes.

NIH CDE Tools and Resources

Databases and repositories of data elements and case report forms that may assist investigators in identifying and selecting data elements for use in their projects.

Summary
Table

Subject
Areas

Summary
Table

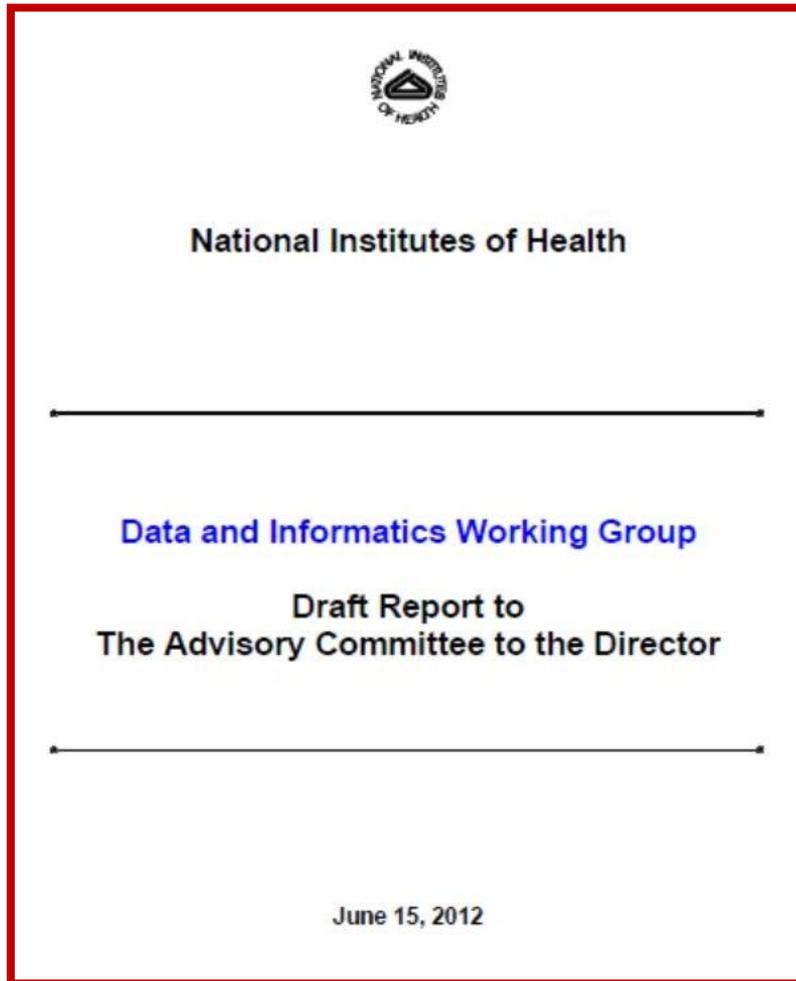
Subject
Areas

The CDE Resource Portal also includes [Other CDE Resources](#) and [Relevant Standards](#). Descriptions of all four groups can be found in the [Glossary](#).

The CDE Working Group of the [Trans-NIH BioMedical Informatics Coordinating Committee](#) (BMIC) developed this Portal to improve the coordination of CDEs. BMIC encourages researchers to use CDEs from the Resources in this Portal where applicable, and to consider existing CDE initiatives before starting additional initiatives.

Are we missing a CDE Resource? [Contact us](#).

NIH Big Data to Knowledge BD2K



Advisory Committee to the NIH Director Data and Informatics Working Group (DIWG*) provided expert advice on the management, integration, and analysis of large biomedical datasets

The DIWG was asked to address:

- Research data spanning basic science through clinical and population research
- Administrative data related to grant applications, reviews, and management
- Management of IT at NIH

*David DeMets & Larry Tabak, Co-Chairs

Rationale for BD2K

- Biomedical research - producing large amounts of heterogeneous data
- Bottleneck
 - Not **data generation, *but***
 - Data management, analysis, visualization and interpretation



BD2K Initiative

Issues enabling community entree to Big Data:

- Sharing and broad use of data
- Software/tools
- Training
- Data science research centers



<http://bd2k.nih.gov>

NIH Big Data to Knowledge (BD2K) Initiatives

1. Facilitating usage and sharing of biomedical big data

- New Policies to Encourage Data & Software Sharing
- Catalog of Research Datasets to Facilitate Data Location & Citation
- Community-based Development of Data & Metadata Standards

2. Development of analysis methods and software

- Software to Meet Needs of the Biomedical Research Community
- Facilitating Data Analysis: Access to Large-scale Computing
- Dynamic Community Engagement of Users and Developers

NIH Big Data to Knowledge (BD2K) Initiatives

3. Enhancing computational training

- Increase Number of Computationally Skilled Trainees
- Strengthen the Quantitative Skills of All Researchers
- Enhance NIH Review and Program Oversight

4. Establishing centers of excellence

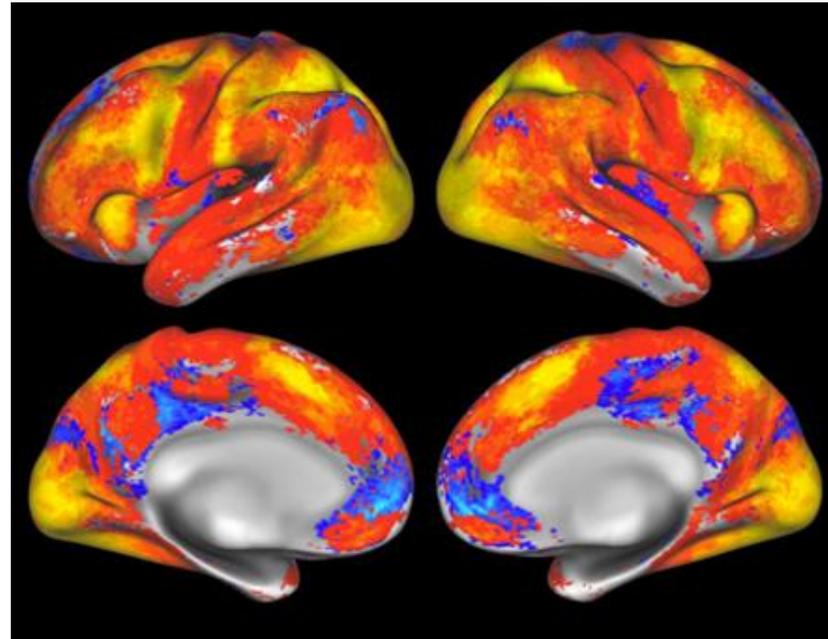
FOA (RFA-HG-13-009): Centers of Excellence for Big Data Computing in the Biomedical Sciences (U54)

- Collaborative environments & technologies
- Data integration
- Analysis & modeling methods
- Computer science & statistical approaches

New Challenges: Preliminary Report on the NIH BRAIN Initiative

Interim Report: Advisory
Committee to the NIH Director,
September 16, 2013

**Brain Research through Advancing
Innovative Neurotechnologies (BRAIN)
Working Group**



The challenge is to map the circuits of the brain, measure the fluctuating patterns of electrical and chemical activity flowing within those circuits, and understand how their interplay creates our unique cognitive and behavioral capabilities

Thematic Core Principles for the NIH BRAIN Initiative

1. **Use appropriate experimental system and models.** The goal is to understand the human brain, but many methods and ideas will be developed first in animal models. Experiments should take advantage of the unique strengths of diverse animal systems.
2. **Cross boundaries in interdisciplinary collaborations.** No single researcher or discovery will crack the brain's code. The most exciting approaches will bridge fields, linking experiment to theory, biology to engineering, tool development to experimental application, human neuroscience to non-human models, and more, in innovative ways.
3. **Integrate spatial and temporal scales.** A unified view of the brain will cross spatial and temporal levels, recognizing that the nervous system consists of interacting molecules, cells, and circuits across the entire body, and important functions can occur in milliseconds, minutes, or take a lifetime.
4. **Establish platforms for sharing data.** Public, integrated repositories for datasets and data analysis tools, with an emphasis on user accessibility and central maintenance, would have immense value.
5. **Validate and disseminate technology.** New methods should be critically tested through iterative interaction between tool-makers and experimentalists. After validation, mechanisms must be developed to make new tools available to all.
6. **Consider ethical implications of neuroscience research.** BRAIN Initiative research may raise important issues about neural enhancement, data privacy, and appropriate use of brain data in law, education and business. Involvement of the President's Bioethics Commission and neuroethics scholars will be invaluable in promoting serious and sustained consideration of these important issues. BRAIN Initiative research should hew to the highest ethical and legal standards for research with human subjects and with non-human animals under applicable federal and local laws.

Open Science...Next Steps

- Future roles of the public and private sectors with regard to open science, as well as the future roles of researchers, librarians, publishers, etc.
 - Interoperability of Software and Technologies
 - Crowdsourcing and Citizen Science as an Open Science Movement
 - Attribution and Provenance of Data and Information

Open Science...Next Steps

- The preservation of the scholarly record – how can/will it be done as the flow of scholarly communication becomes less controlled and more open
 - Open Science in the Multiverse of different research areas
 - Data and Information lifecycles
- Geopolitical/security roadblocks
 - Data and information boundaries when Open Science is international

Discover More about NIH



www.nih.gov

