Implementing the NIH Data Sharing Policy: Expectations and Challenges

Belinda Seto, Ph.D.
Deputy Director
National Institute of Biomedical
Imaging and Bioengineering
National Institutes of Health (NIH)

NIH Viewpoint

"Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data."

-- NIH Statement on Sharing Research Data
February 26, 2003

National Institute of Biomedical Imaging and Bioengineering

Grants Policy Statement April 1994

National Institute of Biomedical Imaging and Bioengineering

"Restricted availability of unique resources upon which further studies are dependent can impede the advancement of research and the delivery of medical care. Therefore, when these resources are developed with NIH funds and the associated research findings have been published or after they have been provided to the agencies under contract, it is important that they be made readily available for research purposes to qualified individuals within the scientific community. This policy applies to grants, cooperative agreements, and contracts."

NIH Data Sharing Policy

Effective October 1, 2003

- NIH *expects* timely release and sharing of final research data for use by other researchers.
- NIH *expects* grant applicants to include a plan for data sharing or to state why data sharing is not possible, especially if $500K or more of direct cost is requested in any single year
- NIH *expects* contract offerors to address data sharing regardless of cost

Challenges

- **Cultural Challenges**
  - Obtaining data in a traditionally data sharing adverse environment
  - Overcoming the competitive and costly "silo" approach to biomedical research
  - Removing barriers to information flow across the complex, heterogeneous environment
- **Technical Challenges**
  - Dealing with a lack of interoperable technologies, unifying architectures, standards, and terminologies
  - Implementing strategies to process and analyze terabytes of data efficiently
  - Maintaining systems in a biologically changing environment
  - Securing, protecting, and tracking patient data across disparate systems

National Institute of Biomedical Imaging and Bioengineering

Data Sharing Models

- NIH serves as central data repository
- Grantee institutions provide data repositories

NIH Central Data Repositories

- Genome-wide association study
- GenBank
- Protein Cluster
- PubChem
- Many others at: http://www.nlm.nih.gov/databases/

Genome-wide Association Studies (GWAS): Purpose, Goals

- To identify common genetic factors that influence health and disease

- To study genetic variations, across the entire human genome, that are associated with observable traits

- To combine genomic information with clinical and phenotypic data to understand disease mechanism and prediction of disease

- To develop the knowledge base for personalized medicine

National Institute of Biomedical Imaging and Bioengineering

GWAS Data Sharing Policy

All GWAS-funded investigators are expected to submit to the NIH data repository descriptive information, curated and coded phenotype, exposure, genotype, and pedigree data as soon as quality control procedures are completed at the grantee institutions.

Database of Genotype and Phenotype (dbGP)

National Institute of Biomedical Imaging and Bioengineering

- Serves as a single point of access to GWAS data
- To archive and distribute results from studies of the interaction of genotype and phenotype
- Provides pre-competitive data, no IP protection
- Encourages use of primary data from dbGP to develop commercial products or tests

Protection of Research Participants: De-Identification

- NIH does not possess direct identifiers of research participants; does not have access to link between data keycode and identifiable information; such information resides with the grantee institutions
- Research institutions submitting dataset must certify that an IRB and/or Privacy Board has considered and approved the submission
- Investigators must stripped the data of all identifiers before data submission
- Optional: Certificate of Confidentiality

National Institute of Biomedical Imaging and Bioengineering

Protection of Research Participants: Informed Consent

- NIH expects specific discussion and documentation that participants' genotype and phenotype data will be shared for research purposes through dbGP

- If participants withdraw consent for sharing individual-level genotype and phenotype data, the submitting institution will be responsible for requesting the dbGP to remove the data involved from future data distributions.

National Institute of Biomedical Imaging and Bioengineering

Data Access

- Requesters are expected to meet data security measures: physical security, information technology security and user training
- Requires signed data use certification:
  - Proposed research use of data
  - Follows local laws
  - Not sell data elements
  - Not share with individuals not listed in proposal
  - Provide annual progress reports

dbGP Access: Two Levels

National Institute of Biomedical Imaging and Bioengineering

- Open-access data includes:
  - summaries of studies
  - study documents, reports
  - measured variables, e.g., phenotypes
  - genotype-phenotype analyses

dbGP: Controlled-Access

- Requires varying levels of authorization
- Provides data on a per-study basis
- Controlled-access data includes:
  - De-identified phenotypes and genotypes for individual study subjects
  - Pedigrees
  - Pre-computed univariate association between genotype and phenotype

National Institute of Biomedical Imaging and Bioengineering

Controlled-Access Data Requests

- Requester must submit a Data Use Certification

- Access is granted by an NIH Data Access Committee

- Approval of proposed research use will be consistent with patient consent and data provider's institutional terms and conditions

Intellectual Properties?

- Discourages premature claims on pre-competitive information that may impede research
- Encourages patenting of technology for downstream product development, e.g.,
  - Markers for assays
  - Drug targets
  - Therapeutics
  - diagnostics
- Up to one year of exclusivity is allowed for the primary investigators to submit GWAS data analyses for publication
- Clock begins when the GWAS datasets is first made available to the NIH data repository

Example of Grantee Institution Providing Access

The National Longitudinal Study of Adolescent Health (Add Health):

An Example of Sensitive Data and Multi-Tiered Access

National Institute of Biomedical Imaging and Bioengineering

The National Longitudinal Study of Adolescent Health (Add Health)

National Institute of Biomedical Imaging and Bioengineering

- **20,745 adolescents enrolled in grades 7-12, followed between 1994 and 2002.**
- **Data from:**
  - **adolescents and parents;**
  - **90,118 students attending sample schools;**
  - **school administrators;**
  - **independent data on neighborhood/community**
- **Data collected in three waves, 1994 - 2002.**
- **Measures of:**
  - **health**
  - **health-related behaviors (e.g., sex, drugs)**
  - **determinants of health at the individual, family, school, peer group, and community level.**

Add Health: Sensitive Data Sharing Example

Challenges to Sharing Data

- Data sensitivity
- Need to protect confidentiality
- Danger of deductive disclosure

National Institute of Biomedical Imaging and Bioengineering

Add Health: Sensitive Data Sharing Example

National Institute of Biomedical Imaging and Bioengineering

A further challenge…

The timely release of these public use samples is essential. Reviewers understand this to mean that investigators outside of the Carolina Population Center will have ready access to the data as soon as investigators inside the center have such access. Procedures for the guarantee of confidentiality … should apply to all users, both the general public and those at University of North Carolina.

Add Health: Sensitive Data Sharing Example

Solution: a multi-tiered system

- Public use data

- Contractual data sets

- Cold room for on-site data use

Add Health: Sensitive Data Sharing Example

Public use data

- Made available through Sociometrics, a small business data archive
- Contains only a subset of cases (6,504)
- Rare over-samples not included
- Contains most data on included cases
- Potentially identifying information redacted

Add Health: Sensitive Data Sharing Example

Restricted-use contractual data

- Full data set available only under contract
- Available to researchers with:
  - IRB- and UNC-approved data security plan
  - Signed agreement to maintain confidentiality
  - Fee covering costs of providing data & user support; monitoring compliance
- Requires annual progress report and renewal after 3 years

Add Health: Sensitive Data Sharing Example

Cold room for on-site use

- Initial plan required access to some data only on-site at UNC
- Cold room constructed at UNC
- Limited use to date

Add Health: Sensitive Data Sharing Example

Data security caveats

- Security requirements require periodic updating as technology advances
- IRBs often lack understanding of security needs
- Smaller institutions handicapped in creating secure environments for restricted data

National Institute of Biomedical Imaging and Bioengineering

Impact of Sharing Ad Health Data

- 700 publications
- 1000 conferences
- 100 dissertations

Challenges: Sharing Image Data

- Different Data acquisition
- Data processing
- Open architecture
- Open access
- Interoperability

National Institute of Biomedical Imaging and Bioengineering
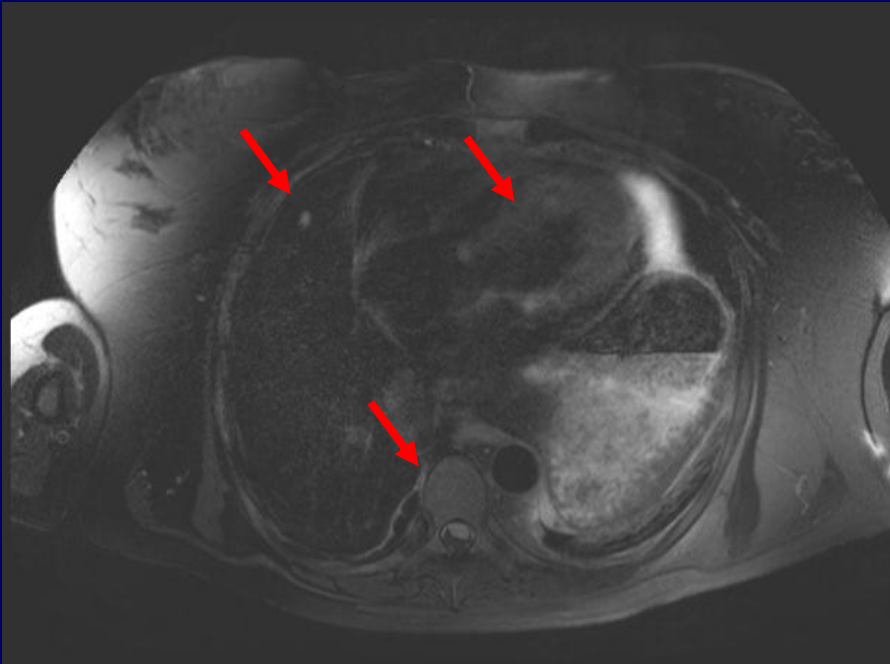
T2 Weighted Images

Turbo Spin Echo
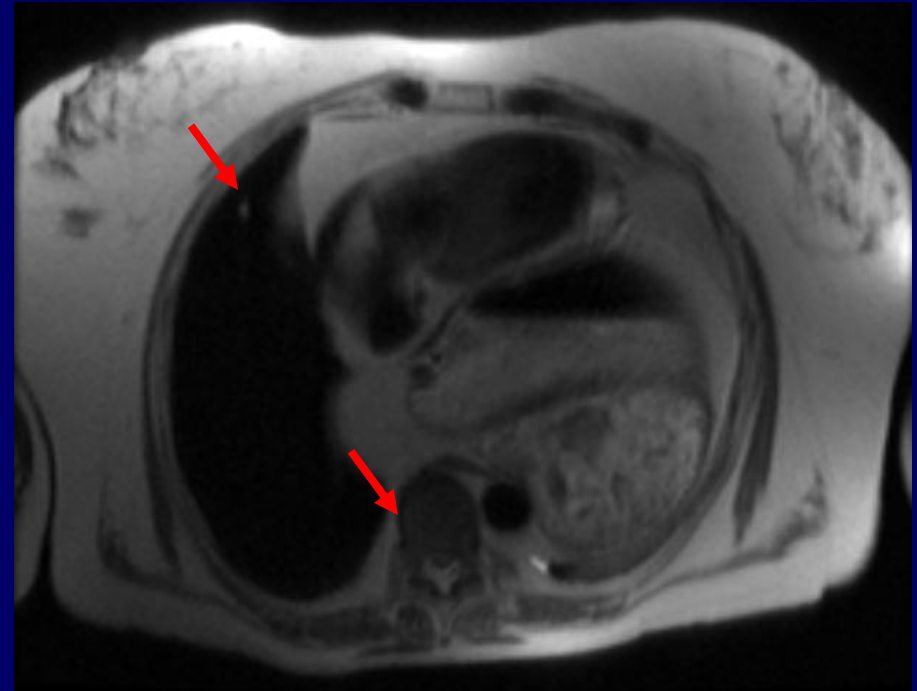3T with fat suppression
Philips
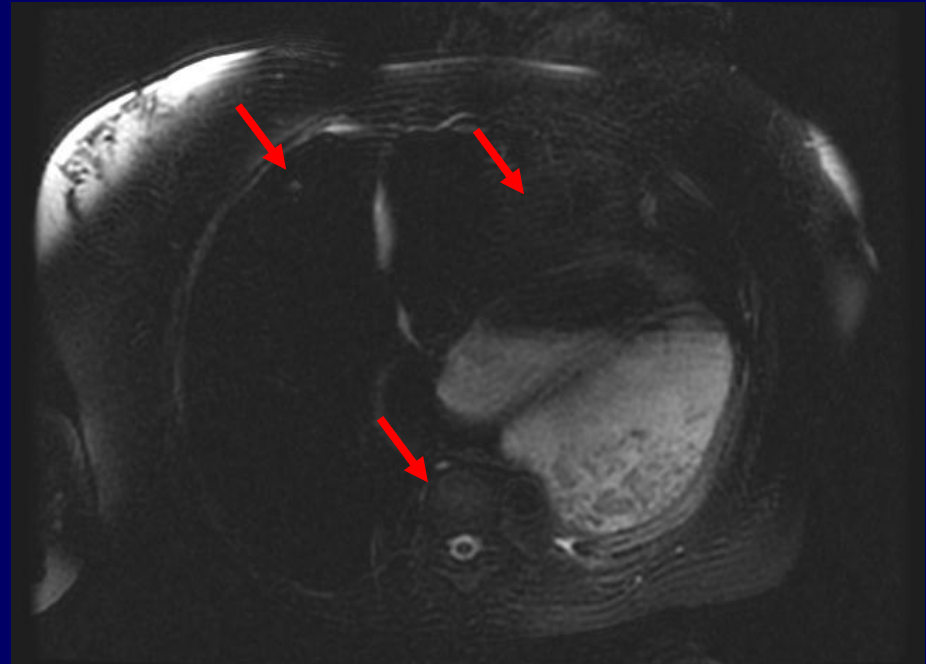
Turbo Spin Echo
1.5T with fat suppression
GE

T2 Weighted Images

Single Shot Fast Spin Echo
3T
Philips

Single Shot Fast Spin Echo
1.5T
GE