CODATA Workshop

NIH Data Sharing

Dr. Belinda Seto, Deputy Director
National Institute of Biomedical Imaging and Biomedical Engineering

October, 2006

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

NIH Data Sharing Policy

Effective with October 1, 2003 receipt date for NIH applications

- 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

Caveats for Studies Including Human Research Participants

- **Investigators need to carefully consider**
  - **Studies with very small samples**
  - **Studies collecting very sensitive data**
- **However, even these data can be shared *if* safeguards exist to ensure confidentiality and protect the identity of subjects**

How to Share Data

- Provide in publications
- Share under the investigator's own auspices
- Place datasets in public archives
- Place in restricted access data centers or data enclaves

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

Special topic: Data Sharing and HIPAA Privacy Rule

Option 1: De-identified Health Information

- Completely de-identified information (18 elements removed) and no knowledge that remaining information can identify the subject. OR
- Statistically "de-identified" information where a statistician certifies that there is a "very small" risk that the information could be used to identify the subject.

Special topic: Data Sharing and HIPAA Privacy Rule

The Privacy Rule defines 18 identifiers

- Names
- Geographic info (including city, state, and zip)
- Elements of dates
- Telephone #s
- Fax #s
- E-mail address
- Social Security #
- Medical record, prescription #s
- Health plan beneficiary #s
- Account #s
- Certificate/license #s
- VIN and Serial #s, license plate #s
- Device identifiers, serial #s
- Web URLs
- IP address #s
- Biometric identifiers (finger prints)
- Full face, comparable photo images
- Unique identifying #s

Special topic: Data Sharing and HIPAA Privacy Rule

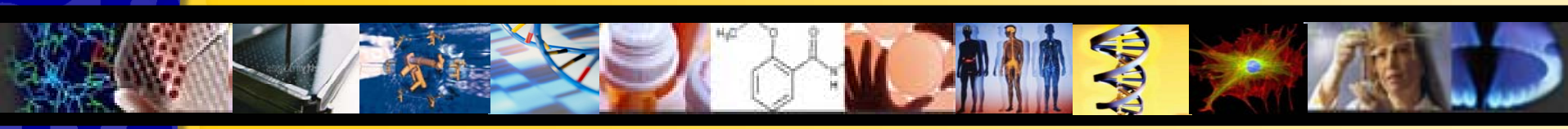
Option 2: Limited Data Set with Data Use Agreement

- **The Privacy Rule permits limited types of identifiers to be released with health information (referred to as a Limited Data Set) -- City, State, Zip; Elements of Date; Unique identifiers not listed as one of 16 direct identifiers**

- **Limited Data Sets can only be used and released in accordance with a Data Use Agreement between the covered entity and the recipient.**

Data Sharing in the Post-Genomic Era

Genome Wide Association Study: Data Sharing



What is Genome-Wide Association Study?

- A genome-wide association study is currently defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition.
- Both clear phenotype information and extensive genotyping (375-500,000 SNPs) are expensive and labor-intensive.

Rationale for Data Sharing

- Genome-wide association studies (GWAS) use scarce human and economic resources, which are diverted from other activities.
- The populations studied by NIH-supported investigators are limited in number, and represent an immensely valuable resource.
- The amount of data obtained exceeds that which can be analyzed by any single group of investigators by many orders of magnitude.

Rationale (continued)

- The cost of extensive genotyping has fallen rapidly, and continues to fall, making studies feasible which would not have been possible even 4 years ago.
- NIH is getting applications for many of these studies representing many millions of dollars.
- If the data remain sequestered, work may be unnecessarily duplicated, and the potential value of multiple analyses and approaches will be lost.

Guiding Principle:

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

Goals of the Proposed Policy

- Advance science for the benefit of the public through the creation of a centralized NIH GWAS data repository.
- Facilitate research and medical science to better address the health needs of people based on their individual genetic information.

Proposed NIH Policy for GWAS

- Data Sharing Procedures
- Data Access Principles
- Intellectual Property
- Protection of Research Participants

NIH GWAS Research Overview

NIH Supported or Conducted GWAS Study

Coded Genotype Data

Coded Phenotype Data

GWAS Data Repository

Study Protocol Descriptive Information → Public Access

Coded Genotype – Phenotype Dataset Pre-computes → **Login Control** → Controlled Access

Dataset specific access rights

Data Access Committee

Peer Review

Submitting investigators & institutions responsible for:
- Compliance with applicable laws/policies
- Data submitted using random code without identifiers
- Notation of any limitations on data use

GWAS application

Data Users and institutions responsible for:
- Compliance with app. laws/policies
- Use for specified research purpose
- Agreement not to identify individuals
- Protecting data confidentiality

Requested Research Use

Data Sharing Procedures

- Central GWAS Data Repository at NCBI, National Library of Medicine
- Data Submission
  - All NIH supported investigators of GWAS are expected to submit protocol, questionnaires, study manuals, measured variables, and other documentation

Data Sharing Procedures
(Continued)

- Data Submission
  - NIH strongly encourages submission of curated and coded phenotype, exposure, genotype and pedigree data
  - All data will be submitted without identifiable information using a random, unique code consistent with the HIPAA Privacy Rule
- Submitted data to be accompanied by:
  - Certification by the responsible IRB
  - Institutional statement that data is in accord with all applicable laws and regulations

Potential Identifiers

- Geographic subdivisions smaller than the state will be needed for genetic-environmental interaction studies
- Dates smaller than a year may be needed for some studies
- A code will be retained to link to data so that it can be updated or withdrawn

Data Access

- Basic descriptive information available to the public
- Access to genotype and phenotype datasets along with pre-computed analyses for research purposes will be provided through NIH Data Access Committee (DAC)
- PIs seeking data will submit a Data Use Certification that is co-signed by the designated Institutional Official for approval by the appropriate DAC
- Confirmation that the proposed research use is consistent with any restraints identified at the Institution submitting the dataset

Publication

- Period of exclusivity for Primary Investigators, proposed 9 months
- Acknowledgement of contributing investigators and funding organization

Intellectual Property

- NIH hopes that genotype-phenotype associations will remain available to all investigators, unencumbered by IP claims
- Discourage premature claims on pre-competitive information
- NIH encourages broad use of NIH supported genotype-phenotype data with NIH's Best Practices for Licensing with Genomic Inventions

Intrinsic limitations

- Return of results
  - To study subjects
  - To primary investigators
- Limited control on use of data by secondary investigators
- Participants can withdraw consent and data, but once data are released, they cannot be retrieved.

Risks

- Sensitive information: ancestry, disease susceptibility can be revealed.
- Genetic information reveals things about people to whom you are related: family, ethnic group.
- This information is not obvious, and requires great sophistication and effort to obtain.
- For real benefits, it must be possible to update information, and a code must be maintained.

Risks (continued)

- Privacy
  - Removed from most identifiers
  - Ultimately identifiable – IF you have extensive genetic information on the person or a close relative
  - Information is subject to FOIA.
- Genetic discrimination
  - The science is way ahead of the law on this.

NIH Public Consultation on Sharing Genetic Data

NATIONAL INSTITUTES OF HEALTH
Office of Extramural Research

NIH Home
OER Home

Search Entire Site
Active Funding Opportunities
Search

Genome-Wide Association Studies (GWAS)

- OER Home
- Funding Opportunities
- Applications & Forms
- Awarded Grants
- Grants Policy
- eRA
- About OER

The NIH is interested in advancing Genome-Wide Association Studies (GWAS) to identify common genetic factors th se because the information derived from such studies will be essential for developing new approaches to reduce disease burden and promote health. GW study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pre or absence of a disease or condition. The goal of the proposed policy is to advance science for the benefit of the public through the creation of a centr . The purpose of this Website is to support the public consultation process to inform policy development activities.

The "Overview" section of this site presents the essential background and respons n GWAS. The remaining sections on this page focus on the notices released to date which will result in a request for information through which th) will solicit advice and comments.

Overview

- Background

Notices and Announcements

- Submit a Comment - Com ough October 31, 2006.
- NIH Press Release (08/30/20 ks Input on Proposed Repository for Genetic Information
- Federal Register Notice (08/30/2006) - Request for Information (RFI): Proposed Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS)
- NIH Guide Notice NOT-OD-06-094 (08/30/2006) - Request for Information (RFI): Proposed Policy for Sharing of Data obtained in NIH supported or conducted Genome-Wide Association Studies (GWAS)
- NIH Guide Notice NOT-OD-06-071 (05/15/2006) - Notice to Applicants for NIH Genome-Wide Association Studies

Comments or Questions?

- Please send email to GWAS@nih.gov.

Google: "GWAS policy"