ATTACHMENT 14

NIH DATA BASE OF GENOTYPES AND PHENOTYPES (DBGAP) OVERVIEW, DATA ACCESS AND SECURITY PRACTICES

dbGaP Overview

The **d**ata**b**ase of **G**enotypes **a**nd **P**henotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The advent of high-throughput, cost-effective methods for genotyping and sequencing has provided powerful tools that allow for the generation of the massive amount of genotypic data required to make these analyses possible.

dbGaP provides two levels of access - open and controlled - in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Summaries of studies and the contents of measured variables as well as original study document text are generally available to the public, while access to individual-level data including phenotypic data tables and genotypes require varying levels of authorization.

View Certificate of Confidentiality

The data in dbGaP will be pre-competitive, and will not be protected by intellectual property patents. Investigators who agree to the terms of dbGaP data use may not restrict other investigators' use of primary dbGaP data by filing intellectual property patents on it. However, the use of primary data from dbGaP to develop commercial products and tests to meet public health needs is encouraged.

Submission Policy

Submitters who are not Federally-funded and affiliated with an NIH IC will need to work with an NIH DAC so that proposed submission can be reviewed for consistency with appropriate policies to protect the privacy of research participants and confidentiality of their data. Submissions to dbGaP will not be accepted without assurance that the submitting institution approves the submission and has verified that the data submission is consistent with all applicable laws and regulations, as well as institutional policies. Submitters must also identify any limits on research uses of the data that are specifically set by individual research participants, e.g., through their informed consent.

Data Content and Organization

Open-Access Data

Open-access data can be browsed online or downloaded from dbGaP without prior permission or authorization. These data will include, but may not be limited to, the following:

dbGaP Data Type	Where to Find It

Studies	'Study' column when browsing studies
	Result of a search under the tab 'Studies'
	Part of the breadcrumb path of a variable or document
Study Documents	Link from 'Browse Studies'
	Link under 'Associated Documents' on study report
	Result of a search under the tab 'Study Documents'
Phenotypic Variables	Link under 'Browse Studies'
	Link under 'Associated Variables' on study report
	Result of a search under the tab 'Variables'
Genotype-Phenotype Analyses	Link under 'Associated Analyses' on variable report
	Link under 'Associated Analyses' on study report

Please note that this is a general description of what is available to open- access users. Data available to open-access users may vary between studies and may also differ from what is described here without notice. You can find more details regarding data access policies for specific studies on the individual study report pages.

Controlled-Access Data

Controlled-access data can only be obtained if a user has been authorized by the appropriate Data Access Committee (DAC). Information on requesting controlled data access, is available below. Data available to authorized investigators may include the following:

- de-identified phenotypes and genotypes for individual study subjects
- pedigrees
- pre-computed univariate associations between genotype and phenotype (if not made available on the public site)

Since data access policies are determined on a per-study basis, data available to users with controlled access authorization may vary between studies and may also change from what is described here without notice. You can find more details regarding data access policies for a specific study on the study report page along with a link to the appropriate authorizing body.

Requesting Controlled-Access Data

Access to controlled data in dbGaP will be granted by an NIH Data Access Committee or DAC. Users wishing access to controlled data must submit a Data Use Certification, or DUC, to the appropriate NIH DAC for approval. DAC approval for controlled data access will be dependent upon completion of the DUC, and confirmation that the proposed research use is consistent with patient consent forms and any constraints identified by the institutions that submitted the dataset(s) to dbGaP. Links to a study's DAC will be found on the study report page.

Submitters of controlled-access data housed in dbGaP may retain the exclusive right to publish an analysis of their submitted data for a specified period of time. Users of controlled-access data should consult the DUC or the study report page to determine the specific publishing exclusivity period for that study.

Access to individual-level data housed in dbGaP is under the jurisdiction of the sponsoring institute. Therefore, any questions regarding access to controlled data should be directed to the DAC for the study in question, and not to NCBI.

Submitting Data to dbGaP

In addition to providing individual-level phenotype and genotype data to dbGaP, we also require the submission of sufficient metadata to enable NCBI to provide a browseable interface for a study.

The following should be included in submissions to dbGaP:

- study documents (i.e. manual of procedures, protocols, questionnaires, consent forms, etc.)
- data dictionary (a description of measured variables with pointers to those parts of study documents that describe how variables were measured)
- any other supporting documentation
- phenotype, exposure, genotype, and pedigree data without identifiable information, created using a random, unique code whose key will be held by the submitting institution
- a guarantee that the identities of research participants will not be disclosed to dbGaP, or to secondary users of the coded data, without appropriate institutional approvals (due to this guarantee, research participants should not expect the return of individual research results derived from the analyses of submitted data)
- a statement verifying that the data submitted to dbGaP for subsequent sharing and appropriate research is consistent with the initial informed consent process completed by study participants
- a statement identifying any uses of the data that are specifically excluded by the informed consent process
- a statement from the data's originating institution that submission of the data is in accord with all applicable laws and regulations

Please note that these are general submission requirements. Since data submission policies are still being developed by participating studies/institutions, submission requirements may vary between studies and may also change from what is described here without notice.

As dbGaP is a NCBI data distribution service, the control and management of the data housed in dbGaP is under the jurisdiction of the sponsoring institute or study; therefore, any questions regarding submission requirements or other data issues should be directed to the DAC for the study in question.

Glossary

Data Access Committee (DAC): Data Access Committees are established based on programmatic areas of interest as well as technical and ethical expertise. All DACs will operate through common principles and under similar mechanisms to ensure the consistency and transparency of the controlled- data access process.

Data Use Certification (DUC): A Data Use Certification is the application a user submits to a particular study's Data Access Committee (DAC) for consideration for authorized use of controlled dbGaP data. The Data Use Certification should include a list of the controlled data set(s) required by the user and a brief description of the proposed research use of the requested data. The user must also offer the following assurances in the Data Use Certification that:

- the data will only be used for approved research
- data confidentiality will be protected
- all applicable laws, local institutional policies, and terms and procedures specific to the study's data access policy for handling dbGaP data will be followed
- no attempts will be made to identify individual study participants from whom data were obtained
- controlled-access data from dbGaP will not be sold or shared with third parties
- the contributing investigator(s) who conducted the original study and the funding organizations involved in supporting the original study will be acknowledged in publications resulting from the analysis of those data
- all NIH supported genotype/phenotype data and conclusions derived directly from them will remain in the public domain, without licensing requirements
- an annual research progress report will be submitted to the study's DAC

Finally, the completed DUC must be co-signed by a designated official representing the institution for which the applicant works.

Please note that this is a general description of the DUC. Since data access policies are still being developed by participating studies/institutions, controlled access requirements, and hence, DUC requirements may vary between studies and may also change from what is described here without notice. Additional details regarding controlled access requirements for a specific study will be provided on the study report page.

DEPARTMENT OF HEALTH & HUMAN SERVICES Public Health Service
National Institutes of Health
National Human Genome
Research Institute
Ethical, Legal, and Social
Implications (ELSI) Research
5635 Fishers Lane, Suite 4076
Bethesda, MD 20892-9305
Telephone: (301) 402-4997
Fax: (301) 402-1950

CONFIDENTIALITY CERTIFICATE

Number: HG-2009-01
Issued to
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
conducting research known as
"**The database for Genotype and Phenotype (dbGaP)**."

In accordance with the provisions of section 301(d) of the Public Health Service Act 42 U.S.C. 241(d), this Certificate is issued in response to the request of the Principal Investigator, James Ostell, Ph.D., to protect the privacy of research subjects by withholding genotype and phenotype data from all persons not connected with this research. Dr. Ostell is primarily responsible for the conduct of this research, which is supported by the National Library of Medicine, National Institutes of Health. Under the authority vested in the Secretary of Health and Human Services by section 301(d), all persons who:
1. are enrolled in, employed by, or associated with the National Library of Medicine, National Institutes of Health and its contractors or cooperating agencies and

2. have in the course of their employment or association access to genotype or phenotype data from individuals who are the subjects of the research submitted to the project known as "**The database for Genotype and Phenotype (dbGaP)**.",

are hereby authorized to protect the privacy of the individuals who are the subjects of that research by withholding genotype and phenotype data from all persons not connected with the conduct of that research.

This database serves as a central point of research access for NIH-supported and other genotype-phenotype datasets, as well as some DNA sequence-based studies. dbGaP contains both public “open access” information about genetic studies as well as “controlled access” for research datasets with individual-level data. Individual-level data, as well as summary level statistics files of all genotype data, are included in the controlled access sections of dbGaP. Publicly available dbGaP information includes details of study design, participant selection, and aggregate outcomes. Controlled access data provides research access to the genotype and phenotype data for individual participants. The National Center for Biotechnology Information, National Library of Medicine has the linking code back to the identifiers located at the contributing site.

A Certificate of Confidentiality is needed because potentially sensitive genetic information will be collected during the course of the study. The certificate will help researchers avoid involuntary disclosure that could expose subjects or their families to adverse economic, legal, psychological and social consequences.

To protect participants’ confidentiality, there will be on-going assurance of participant protections pertaining to data held within dbGaP based on oversight by the NIH of data access to “secondary” data users. Research access to dbGaP data will be provided through a “Controlled Access” process implemented by NIH Data Access Committees (DACs). Data Access Committees will be constituted by the NIH Institutes with federal employees possessing the appropriate scientific and bioethics expertise, and through the oversight and actions of these committees access to dbGaP datasets will be provided based on the consistency of specific research uses (proposed by data requestors) with the data use limitations set by the institutions submitting the datasets to the NIH. Approved data users will agree, along with their home institutions, to follow specified principles and terms of use for the specific dataset provided. NIH will monitor data use practices over time to assure that policies and procedures for protecting participants and their interests remain robust.

This research began on December 14, 2006 expected to be ongoing. This Certificate will end on October 31, 2018 and is expected to be renewed at that time.

As provided in section 301 (d) of the Public Health Service Act 42 U.S.C. 241(d):

"Persons so authorized to protect the privacy of such individuals may not be compelled in any Federal, State, or local civil, criminal, administrative, legislative, or other proceedings to identify such individuals."

This Certificate does not protect you from being compelled to make disclosures that: (1) have been consented to in writing by the research subject or the subject’s legally authorized representative; (2) are required by the Federal Food, Drug, and Cosmetic Act (21 U.S.C. 301 et seq.) or regulations issued under that Act; or (3) have been requested from a research project funded by NIH or DHHS by authorized representatives of those agencies for the purpose of audit or program review.

This Certificate does not represent an endorsement of the research project by the Department of Health and Human Services. This Certificate is now in effect and will expire on October 31, 2018. The protection afforded by this Confidentiality Certificate is permanent with respect to any individual who participates as a research subject (i.e., about whom the investigator maintains identifying information) in a study that is submitted to the dbGaP project during any time the Certificate is in effect.

Date: _______________ ______________________________________

Elizabeth J. Thomson, DNSc, RN, CGC, FAAN
Program Director
Ethical, Legal, and Social Implications Research

dbGaP Best Practices Requirements

Page 1

SECURITY BEST PRACTICES – Level 2b

Links updated: 11/21/2008

Introduction

The data sets provided in conjunction with this agreement are controlled access data. The procedures described below are based on the assumption that access to deidentified person level detailed genomic data associated with phenome data should be controlled and not publicly available.

The goal of this process is to ensure that data provided by the NIH is kept sufficiently secure and not released to any person not permitted to access the data, either through malicious or inadvertent means. To accommodate these requirements, systems housing these data must not be directly accessible from the internet, and the data must not be posted on any web or ftp server. Data placed on shared systems must be secured and limited to those involved in the research for which the data has been requested. If data is stored on laptops or removable devices, those devices must be encrypted. Protecting the Security of Controlled Data Security Awareness Requirements The controlled access data you received is considered sensitive information. By following the best practices below, you will be doing much towards protecting the information entrusted to your care. This is a minimum set of requirements; additional restrictions may be needed by your institution and should be guided by the knowledge of the user community at your institution.

Think Electronic Security

1. The Single Most Important Advice: Download data to a secure computer or server and not to unsecured network drives or servers.

2. Make sure these files are never exposed to the Internet. Data must never be posted on a PI’s (or institution’s) website because the files can be “discovered” by internet search engines, e.g., Google, MSN.

3. Have a strong password for file access and never share it.

4. If you leave your office, close out of data files or lock your computer.

5. Install a password-enabled screen saver that activates after 15 minutes of inactivity.

Page 2

6. Data stored on laptops must be encrypted. Most operating systems have the ability to natively run an encrypted file system or encrypt portions of the file system. (Windows = EFS or Pointsec and Mac OSX = File Vault) Think Physical Security

1. If the data are in hard copy or reside on portable media, e.g., on a CD, flash drive or laptop), treat it as though it were cash.

2. Don’t leave it unattended or in an unlocked room.

3. Consider locking it up.

4. Exercise caution when traveling with portable media, i.e., take extra precautions to avoid the possibility of loss or theft (especially flash drives which are small and can easily be misplaced).

Protecting the Security of Controlled Data on Servers

1. Servers must not be accessible directly from the internet, (i.e. must be behind a firewall or not connected to a larger network) and unnecessary services disabled.

2. Keep systems up to date with security patches.

3. dbGaP data on the systems must be secured from other users (restrict directory permissions to only the owner and group) and if exported via file sharing, ensure limited access to remote systems.

4. If accessing system remotely, encrypted data access must be used (such as SSH or VPN). It is preferred to use a tool such as RDP, X-windows or VNC that does not permit copying of data and provides “View only” support.

5. Ensure that all users of this data have IT security training suitable for this data access and understand the restrictions and responsibilities involved in access to this data.

6. If data is used on multiple systems (such as a compute cluster), ensure that data access policies are retained throughout the processing of the data on all the other systems. If data is cached on local systems, directory protection must be kept, and data must be removed when processing is complete.

Requesting Investigators must meet the spirit and intent of these protection requirements to ensure a secure environment 24 hours a day for the period of the agreement.

Use Data by Approved Users on Secure Systems

The requesting investigator must retain the original version of the data encrypted data. The requesting investigator must track any copies or extracts made of the data and shall make no copy or extract of the subject data available to anyone except an authorized staff member for the purpose of the research for which the subject data were made available. Collaborating investigators from other institutions must complete an independent data use certification to gain access to the data. When use of the dataset is complete—destroy all individually identifiable data

1. Shred hard copies.

2. Delete electronic files securely.

3. At minimum, delete the files and then empty your recycle bin.

4. Optimally, use a secure method, e.g., an electronic “shredder” program that performs a permanent delete and overwrite.

Additional Resources for testing and best practices:

The Center for Internet Security

CIS is the only distributor of consensus best practice standards for security configuration. The Benchmarks are widely accepted by U.S. government agencies for FISMA compliance, and by auditors for compliance with the ISO standard as well as GLB, SOx, HIPAA, FERPA and other regulatory requirements for information security. End user organizations that build their configuration policies based on the consensus benchmarks can not acquire them elsewhere.

http://www.cisecurity.org/. Appendix A – Has checklists based on CIS best practices, customized for dbGaP data use.

Content for this document has been adapted from CIT/NIH and CIS

Appendix A:

Best Practice Security Requirements for dbGaP Data Recipients

Preface

This appendix has been adapted from the HHS IT Security program for minimal security standards and the Center for Internet Security, and adapted as “Best Practices” for dbGaP

Introduction

The dbGaP Best Practices Guidelines Checklists were created to provide guidance and expectation on how to treat the controlled access data received from dbGaP.

Purpose

The purpose of this appendix is to provide minimum configuration standards for recipients of data from dbGaP. Adhering to these procedures will provide a baseline level of security, ensuring that minimum standards or greater are implemented to secure the confidentiality, integrity, and availability of data resources. If local IT policies are more restrictive, then local policies should apply.

Background

Minimum security configuration standards help to ensure sound control of each system. Adhering to minimum standards helps to mitigate risks associated with implementing applications and software by providing a solid foundation to track changes, the differences between versions, and new components as they are installed. System and application default settings are not optimal from a security perspective. Using default settings increases the risk of exploitation. These risks are mitigated through the use of minimum security configuration standards. These standards are from CIS checklists and are cross mapped to NIST Recommended Security Controls for Federal Information Systems 800-53 Rev. 2.