

Excerpt from http://grants.nih.gov/grants/policy/nihgps_2003/NIHGPS_Part7.htm

Availability of Research Results: Publications, Intellectual Property Rights, and Sharing Research Resources

It is NIH policy that the results and accomplishments of the activities that it funds should be made available to the public. PIs and grantee organizations are expected to make the results and accomplishments of their activities available to the research community and to the public at large. (See also [“Public Policy Requirements and Objectives—Availability of Information—Access to Research Data”](#) for policies related to providing access to certain research data at public request.) If the outcomes of the research result in inventions, the provisions of the Bayh-Dole Act of 1980, as implemented in 37 CFR Part 401, apply.

As long as grantees abide by the provisions of the Bayh-Dole Act, as amended by the Technology Transfer Commercialization Act of 2000 (P.L. 106-404), and 37 CFR Part 401, they have the right to retain title to any invention conceived or first actually reduced to practice using NIH grant funds. The principal objectives of these laws and the implementing regulation are to promote commercialization of federally funded inventions, while ensuring that inventions are used in a manner that promotes free competition and enterprise without unduly encumbering future research and discovery.

The regulation requires the grantee to use patent and licensing processes to transfer grant-supported technology to industry for development. Alternatively, unpatented research products or resources—“research tools”—may be made available through licensing to vendors or other investigators. Sharing of copyrightable outcomes of research may be in the form of journal articles or other publications.

The importance of each of these outcomes of funded research is reflected in the specific policies pertaining to rights in data, sharing of research data and unique research resources, and inventions and patents described in the following subsections.

Rights in Data (Publication and Copyrighting)

In general, grantees own the rights in data resulting from a grant-supported project. Special terms and conditions of the award may indicate alternative rights, e.g., under a cooperative agreement or based on specific programmatic considerations as stated in the applicable RFA. Except as otherwise provided in the terms and conditions of the award, any publications, data,^[12] or other copyrightable works developed under an NIH grant may be copyrighted without NIH approval. Rights in data also extend to students, fellows, or trainees under awards whose primary purpose is educational, with the authors free to copyright works without NIH approval. In all cases, NIH must be given a royalty-free, nonexclusive, and irrevocable license for the Federal government to reproduce,

publish, or otherwise use the material and to authorize others to do so for Federal purposes. Data developed by a consortium participant also is subject to this policy.

As a means of sharing knowledge, NIH encourages grantees to arrange for publication of NIH-supported original research in primary scientific journals. Grantees also should assert copyright in scientific and technical articles based on data produced under the grant where necessary to effect journal publication or inclusion in proceedings associated with professional activities.

Journal or other copyright practices are acceptable unless the copyright policy prevents the grantee from making copies for its own use (as provided in 45 CFR 74.36 and 92.34). The disposition of royalties and other income earned from a copyrighted work is addressed in “[Administrative Requirements—Management Systems and Procedures—Program Income](#).”

For each publication that results from NIH grant-supported research, grantees must include an acknowledgment of NIH grant support and a disclaimer stating the following:

“This publication was made possible by Grant Number _____ from _____” or
“The project described was supported by Grant Number _____ from _____” and
“Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the [name of awarding office or NIH].”

If the grantee plans to issue a press release concerning the outcome of NIH grant-supported research, it should notify the NIH awarding office in advance to allow for coordination.

One copy of each publication resulting from work performed under an NIH grant-supported project must accompany the annual or final progress report submitted to the NIH awarding office (see “[Administrative Requirements—Monitoring—Reporting—Non-Competing Grant Progress Reports](#)” and “[Administrative Requirements—Closeout—Final Reports—Final Progress Report](#)”).

Sharing of Research Data

NIH believes that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. NIH endorses the sharing of final research data to serve these and other important scientific goals and expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. “Timely release and sharing” is defined as no later than the acceptance for publication of the main findings from the final data set. Effective with the October 1, 2003 receipt date, investigators submitting an NIH application seeking \$500,000 or more in direct costs in any single budget period are expected to include a plan for data sharing or state why data sharing is not possible.

NIH recognizes that data sharing may be complicated or limited, in some cases, by organizational policies, local IRB rules, and local, State and Federal laws and regulations, including the “Privacy Rule” (See “[Public Policy Requirements and Objectives—Requirements Affecting the Rights and Welfare of Individuals as Research Subjects, Patients, or Recipients of Services—Confidentiality—Standards for Privacy of Individually Identifiable Health Information](#)”). The rights and privacy of individuals who participate in NIH-sponsored research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects.

Sharing of Unique Research Resources

Investigators conducting biomedical research frequently develop unique research resources. Categories of these resources include synthetic compounds, organisms, cell lines, viruses, cell products, and cloned DNA, as well as DNA sequences, mapping information, crystallographic coordinates, and spectroscopic data. Specific examples include specialized or genetically defined cells, including normal and diseased human cells; monoclonal antibodies; hybridoma cell lines; microbial cells and products; viruses and viral products; recombinant nucleic acid molecules; DNA probes; nucleic acid and protein sequences; certain types of animals, such as transgenic mice; and intellectual property, such as computer programs.

NIH considers the sharing of such unique research resources (also called research tools) an important means to enhance the value of NIH-sponsored research. Restricting the availability of unique resources can impede the advancement of further research. Therefore, when these resources developed with NIH funds and the associated research findings have been published or after they have been provided to NIH, it is important that they be made readily available for research purposes to qualified individuals within the scientific community.

To provide further clarification of the NIH policy on disseminating unique research resources, NIH published *Principles and Guidelines for Recipients of NIH Research Grants and Contracts on Obtaining and Disseminating Biomedical Research Resources* (64 FR 72090, December 23, 1999), which is available on the NIH website (http://www.ott.nih.gov/policy/rt_guide_final.html). This document will assist grantees in determining reasonable terms and conditions for disseminating and acquiring research tools.

The terms of those agreements also must reflect the objectives of the Bayh-Dole Act and the Technology Transfer Commercialization Act of 2000 to ensure that inventions made are used in a manner to promote free competition and enterprise without unduly encumbering future research and discovery.

In addition to sharing research resources with the research community, upon request of the NIH awarding office, the grantee also must provide a copy of documents or a sample

of any material developed under an NIH grant award. The grantee may charge a nominal fee to cover shipping costs for providing this material. Income earned from these charges must be treated as program income (see “[Administrative Requirements—Management Systems and Procedures—Program Income](#)”).

To facilitate the availability of unique or novel biological materials and resources developed with NIH funds, investigators may distribute the materials through their own laboratory or organization or submit them, if appropriate, to entities such as the American Type Culture Collection or other repositories. Investigators are expected to submit unique biological information, such as DNA sequences or crystallographic coordinates, to the appropriate data banks so that they can be made available to the broad scientific community. When distributing unique resources, investigators are to include pertinent information on the nature, quality, or characterization of the materials.

Investigators must exercise great care to ensure that resources involving human cells or tissues do not identify original donors or subjects, directly or through identifiers such as codes linked to the donors or subjects.

Organizations that believe they will be unable to comply with these requirements should promptly contact the GMO to discuss the circumstances, obtain information that might enable compliance, and reach an understanding in advance of an award.

Inventions and Patents

The Bayh-Dole Act of 1980 (Public Law 96-517; 35 U.S.C. 200-212) and the related EO 12591 (April 10, 1987) provide incentives for the practical application of research supported through Federal funding agreements. To be able to retain rights and title to inventions made with Federal funds, so-called “subject” inventions, the grantee must comply with a series of regulations that ensure the timely transfer of the technology to the private sector, while protecting limited rights of the Federal government.

The regulations apply to any subject invention—defined as any invention either conceived or first actually reduced to practice in the performance of work under the Federal award—and to all types of recipients of Federal funding. This includes non-profit entities and small businesses or large businesses receiving funding through grants, cooperative agreements, or contracts as direct recipients of funds, or as consortium participants or subcontractors under those awards.

NIH grantees may retain intellectual property rights to subject inventions provided they do the following:

- Report all subject inventions to NIH.
- Make efforts to commercialize the subject invention through patent or licensing.

- Formally acknowledge the Federal government’s support in all patents that arise from the subject invention.
- Formally grant the Federal government a limited use license to the subject invention.

Exhibit 5 summarizes recipient responsibilities for invention reporting as specified in the regulations in 37 CFR Part 401. Grantees should refer to 37 CFR Part 401 (available on the Interagency Edison site: <https://s-edison.info.nih.gov/iEdison/>) for a complete discussion of the regulations.

| Exhibit 5. Extramural Invention Reporting Compliance Responsibilities | | | |
|---|--|--|------------------------------|
| Action required | When action must be taken | Discussion | 37 CFR 401 reference |
| Employee Agreement to Disclose All Inventions | | | |
| The PI (employee) must sign an agreement to abide by the terms of the Bayh-Dole Act and the NIHGPS as they relate to intellectual property rights. | At time of employment. | Grantee organizations and consortium participants must have policies in place regarding ownership of intellectual property. | 401.14(f)(2) |
| Invention Report and “Disclosure” | | | |
| The grantee organization must submit to NIH a report of any subject invention. This includes a written description (the so-called “invention disclosure”) of the invention. | Within 2 months of the inventor’s initial report of the invention to the grantee organization. | There is no single format for disclosing the invention to the Federal government. The report must identify inventor(s), NIH grant number, and date of any public disclosure. | 401.14(a)(2) 401.14(c)(1) |
| Rights to Consortium Participant Inventions | | | |
| Consortium participants under NIH grants retain rights to any subject inventions they make. | Within 2 months of the inventor’s initial report of the invention to the consortium participant. (The consortium participant has the same invention reporting obligations as the grantee.) | The grantee cannot require ownership of a consortium participant’s subject inventions as a term of the consortium agreement. | 401.14(g)(1) 401.14(g)(2) |

| Exhibit 5. Extramural Invention Reporting Compliance Responsibilities | | | |
|--|--|--|---|
| Action required | When action must be taken | Discussion | 37 CFR 401 reference |
| Election of Title to Invention | | | |
| The grantee must notify NIH of its decision to retain or waive title to invention and patent rights. | Within 2 years of the initial reporting of the invention to NIH. | | 401.14(b) 401.14(c)(2) 401.14(f)(1) |
| Confirmatory License | | | |
| For each invention, the grantee must provide a use license to NIH for each invention. | When the initial non-provisional patent application is filed. | | 401.14(f)(1) |
| Patent Application | | | |
| The grantee must inform NIH of the filing of any non-provisional patent application. The patent application must include a Federal government support clause. | Within 1 year after election of title, unless there is an extension. | Initial patent application is defined as a non-provisional U.S. application. The patent application number and filing date must be provided. | 401.14(c)(3) 401.2(n) |
| Assignment of Rights to Third Party | | | |
| If the grantee is a non-profit organization, it must ask NIH approval to assign invention or U.S. patent rights to any third party, including the inventor(s). | As needed. The NIH Office of Technology Transfer serves in an advisory capacity to OER for the processing of such assignment requests. | Grantees that are for-profit entities (including small businesses) do not need to ask approval. | 401.14(k) |
| Issued Patent | | | |
| The grantee must notify NIH that a patent has been issued. | When the patent is issued. | The patent issue date, number, and evidence of Federal government support clause must be provided. | 401.5(f)(2) |

| Exhibit 5. Extramural Invention Reporting Compliance Responsibilities | | | |
|---|--|--|----------------------|
| Action required | When action must be taken | Discussion | 37 CFR 401 reference |
| Extension of Time to Elect Title or File Patent | | | |
| The grantee may request an extension of up to 2 years for election of title, or 1 year for filing a patent application. | As needed. | Request for extension of time must be made. Such requests are preapproved. | 401.14(c)(4) |
| Change in Patent Application Status | | | |
| The grantee must notify NIH of changes in patent status. | At least 30 days before any pending patent office deadline. | This notification allows NIH to consider continuing the patent action. | 401.14(f)(3) |
| Invention Utilization Report | | | |
| The grantee must submit information about the status of commercialization of any invention for which title has been elected. | Annually. | This report gives an indication of whether the objectives of the law are being met. Specific reporting requirements can be found in iEdison (https://s-edison.info.nih.gov/iEdison/). | 401.14(h) |
| Annual Invention Statement | | | |
| The grantee must indicate any inventions made during the previous budget period on all grant awards. | Part of all competing applications and non-competing grant progress reports. | The information is requested as a checklist item on the PHS 398 application and on the non-competing grant progress report. | PHS 398 and PHS 2590 |
| Final Invention Statement and Certification | | | |
| The grantee must submit to the NIH awarding office GMO a summary of all inventions made during the entire term of each grant award. | Within 90 days after the project period (competitive segment) ends. | Required information is specified on the HHS 568 form. If no inventions occurred during the project period, a negative report must be submitted. | 401.14(f)(5) |

Failure of the grantee to comply with any of these or other regulations cited in 37 CFR Part 401 may result in the loss of patent rights or a withholding of additional grant funds.

The Bayh-Dole Act includes provisions for the grantee to assign invention rights to third parties. Grantees that are non-profit organizations must request NIH approval for the assignment. If the assignment is approved and the rights are assigned to a third party, invention and patent reporting requirements apply to the third party. The grantee should review existing agreements with third parties and revise them, as appropriate, to ensure they are consistent with the terms and conditions of their NIH grant awards and that the objectives of the Bayh-Dole Act are adequately represented in the assignment.

Any invention made using funds awarded for educational purposes, e.g. fellowships, training grants or certain types of career development awards, is not considered a subject invention and therefore is not subject to invention reporting requirements (as provided in 45 CFR 74. and 37 CFR 401.1(b)). The grantee should seek the advice of NIH to verify whether any invention made under a career development award should be considered a subject invention.

Details regarding invention reporting and iEdison are discussed under “[Administrative Requirements—Monitoring—Reporting—Invention Reporting.](#)”

All issues or questions regarding extramural technology transfer policy and reporting of inventions and their utilization should be referred to the following address:

Extramural Inventions and Technology Resources Branch
Division of Grants Policy
Office of Policy for Extramural Research Administration
Office of Extramural Research
NIH
6705 Rockledge Drive, MSC 7980
Bethesda, MD 20892-7980
301-435-1986 (voice)
301-480-0272 (fax)

Examples of Resource and Data Sharing Plans

RESOURCE SHARING PLAN

We will adhere to the NIH Grant Policy on Sharing of Unique Research Resources including the Sharing of Biomedical Research Resources Principles and Guidelines For Recipients of NIH Grants and Contracts issued in December, 1999

<http://ott.od.nih.gov/policy/rt_guide_final.html>. Our lab has demonstrated its commitment to sharing over the last few years, often prior to publication. We fill requests in a timely fashion. In addition, we will provide relevant protocols and published genetic and phenotypic data upon request. Material transfers will be made with no more restrictive terms than in the Simple Letter Agreement (SLA) or the Uniform Biological Materials Transfer Agreement (UBMTA) and without reach through requirements. Should any intellectual property arise which requires a patent, we will ensure that the technology (materials and data) remains widely available to the research community in accordance with the NIH Principles and Guidelines document.

RESOURCE SHARING

Novel compounds emanating from the studies will be made available to the scientific community upon publication.

SHARING RESEARCH RESOURCES

At time of manuscript acceptance, we will readily share _____, as well as cell lines that we created which would allow other investigators to expand our studies.

SHARING RESEARCH RESOURCES

Research resources generated by this project including mouse models, expressed proteins and antibodies will be shared with other investigators upon request.

SHARING RESEARCH RESOURCES

All reagents generated in this project will be shared with the scientific community.

SHARING RESEARCH RESOURCES

All data and research tools including the new animal line will be freely available to the general scientific community.

DATA SHARING

Data, methodologies and research accomplishments will be reported in publicly accessible scientific journals.

SHARING RESEARCH DATA

Research data obtained during the course of these proposed investigations will be shared and disseminated via seminars and publications.

DATA SHARING

Information will be disseminated via lectures and scientific publications.

DATA SHARING

Data collected during the studies will be initially shared among the investigators working on the project. Results will be shared with the scientific community through submission of abstracts of talks or posters presented at scientific meetings and papers for publication in scientific journals. Data sharing will be available through a Lawrence Berkeley National Laboratory web site with references to published articles. No data-sharing agreement will be required. The documentation to be provided includes results published in peer-reviewed scientific journals, progress reports to NIH, internal LBNL reports, abstracts/papers and talks submitted and presented.

SHARING MODEL ORGANISMS

As for our plan to share materials and our management of intellectual property, we will adhere to the NIH Grant Policy on Sharing of Unique Research Resources including the Sharing of Biomedical Research Resources Principles and Guidelines for Recipients of NIH Grants and Contracts issued in December, 1999 http://ott.od.nih.gov/NewPages/RTguide_final.html. All 'model organisms' generated by this project will be distributed freely or deposited into a repository/stock center making them available to the broader research community, either before or immediately after publication.

If we assume responsibility for distributing the newly generated model organisms, we will requests in a timely fashion. In addition, we will provide relevant protocols upon request. Material transfers will be made with no more restrictive terms than in the Simple Letter Agreement (SLA) or the Uniform Biological Materials Transfer Agreement (UBMTA) and without reach through requirements. Should any intellectual property arise which requires a patent, we will ensure that the technology (materials and data) remains widely available to the research community in accordance with the NIH Principles and Guidelines document.

1.5 Sharing Research Resources

Investigators conducting biomedical research frequently develop unique research resources. NIH considers the sharing of such unique research resources (also called research tools) an important means to enhance the value of NIH-sponsored research. Restricting the availability of unique resources can impede the advancement of further research. Therefore, when these resources are developed with NIH funds and the associated research findings have been published or after they have been provided to NIH, it is important that they be made readily available for research purposes to qualified individuals within the scientific community. At the same time NIH recognizes the rights of grantees and contractors to elect and retain title to subject inventions developed with federal funding pursuant to the Bayh Dole Act. See the NIH Grants Policy Statement, and the Office of Extramural Research, Division of Extramural Inventions & Technology Resources (DEITR), Intellectual Property Policy page: <http://inventions.nih.gov>.

The adequacy of resource sharing plans are considered by reviewers when a competing application is evaluated. Reviewers are asked to describe their assessment of the sharing plan in an administrative note, and will not normally include their assessment in the overall priority score. Program staff are responsible for overseeing resource sharing policies and for assessing the appropriateness and adequacy of any proposed resource sharing plans.

1.5.1 Data Sharing Policy

All investigator-initiated applications with direct costs of \$500,000 or greater in any single year are expected to address data-sharing in their application. Applicants are encouraged to discuss data-sharing plans with their program contact at the time they negotiate an agreement with the Institute/Center (IC) staff to accept assignment of their application as described at <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-004.html>.

Applicants are reminded that agreement to accept assignment of applications \$500,000 or greater must be obtained at least six weeks in advance of the anticipated submission date. Instructions related to the data-sharing policy as it is applied to applications and proposals responding to a specific Request for Application (RFA) or Request for Proposals (RFP) will be described in the specific solicitation. In some cases, other Funding Opportunity Announcements (FOAs) may request data-sharing plans for applications that are less than \$500,000 direct costs in any single year.

NIH recognizes that in some cases data-sharing may be complicated or limited by institutional policies, local IRB rules, as well as local, state and Federal laws and regulations, including the HIPAA Privacy Rule. The rights and privacy of individuals who participate in NIH-sponsored research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects. When data-sharing is limited, applicants should explain such limitations in their data-sharing plans.

For more information on data-sharing, please see http://grants.nih.gov/grants/policy/data_sharing and the NIH Final Policy on Sharing Research Data, <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>.

1.5.2 Sharing Model Organisms

All applications where the development of model organisms is anticipated are expected to include a description of a specific plan for sharing and distributing unique model organism research

PHS SF424 (R&R) Adobe Pilot Forms Application Guide
Section III: Policies, Assurances, Definitions, and Other Information

resources generated using NIH funding so that other researchers can benefit from these resources, or state appropriate reasons why such sharing is restricted or not possible. Model organisms include but are not restricted to mammalian models, such as the mouse and rat; and non-mammalian models, such as budding yeast, social amoebae, round worm, fruit fly, zebra fish, and frog. Research resources to be shared include genetically modified or mutant organisms, sperm, embryos, protocols for genetic and phenotypic screens, mutagenesis protocols, and genetic and phenotypic data for all mutant strains.

This expectation is for **all** applications where the development of model organisms is anticipated, regardless of funding amount.

For additional information on this policy, see the NIH Model Organism for Biomedical Research Website at: <http://www.nih.gov/science/models/> and NIH Guide Notices OD-04-042: <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-042.html>, and OD-04-066: <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-066.html>.

1.5.3 Policy for Genome-Wide Association Studies (GWAS)

NIH is interested in advancing genome-wide association studies (GWAS) to identify common genetic factors that influence health and disease through a centralized GWAS data repository. For the purposes of this policy, a genome-wide association study is 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.

All applications, regardless of the amount requested, proposing a genome-wide association study are expected to provide a plan for submission of GWAS data to the NIH-designated GWAS data repository, or provide an appropriate explanation why submission to the repository is not possible. Data repository management (submission and access) is governed by the Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies, NIH Guide NOT-OD-07-088. For additional information see: <http://grants.nih.gov/grants/gwas/>.

1.6 Inventions and Patents

According to NIH Grants Policy and Federal law, NIH recipient organizations must promptly report all inventions that are either conceived or first actually reduced to practice using NIH funding. Invention reporting compliance is described at <http://www.iedison.gov>. Grantees are encouraged to submit reports electronically using Interagency Edison (<http://www.iedison.gov>). Information from these reports is retained by the NIH as confidential and submission does not constitute any public disclosure. Failure to report as described at 37 CFR Section 401.14 is a violation of 35 U.S.C. 202 and may result in loss of the rights of the recipient organization. Inquiries or correspondence should be directed to: **Division of Extramural Inventions and Technology Resources, Office of Policy for Extramural Research Administration, OER, NIH, 6705 Rockledge Drive, Suite 310, MSC 7980, Bethesda, MD 20892-7980, Telephone: (301) 435-1986.**

DATA SHARING WORKBOOK

- [Introduction](#)
- [Protecting the Rights and Privacy of Human Subjects](#)
- [Protecting Proprietary Data](#)
- [Examples of Data Sharing](#)
 - [Data Archives](#)
 - [Federated Data Systems](#)
 - [Data Enclaves](#)
 - [Mixed Mode Sharing](#)

INTRODUCTION

Scientists working in many different areas are already sharing their data through a variety of mechanisms. However, some disciplines are less familiar with this process and associated practices. **The goal of this workbook is to show how investigators working in a variety of scientific areas have shared their data.** To highlight the benefits of data sharing, we have included testimonials from investigators who are already sharing their data.

NIH supports a wide range of scientific research. Some studies, such as small laboratory-based projects, make raw data available in publications. These studies generally are based on small numbers of laboratory animals, specimens, or clinical subjects. Publishing the raw data constitutes an acceptable mechanism for sharing data, provided that privacy of human subjects is protected. However, raw data from large studies are not amenable to sharing through publication. Such studies can make data available through data archives or enclaves. For example, X-ray crystallography, gene mapping, and survey data are available from data archives or repositories, some with sophisticated Web interfaces. Data from human subjects present special concerns regarding data sharing. The rights and privacy of individuals who participate in NIH-sponsored research must be protected at all times, and patentable and other proprietary data should also be protected.

In summary, all data should be considered for sharing. Data that constitute "unique resources" especially should be shared unless there is a strong reason not to. Such data are difficult if not impossible to replicate because of cost (e.g., large national longitudinal surveys), special circumstances (e.g., health effects associated with a natural disaster), or rare population (e.g., a sample of centenarians). Less likely candidates for sharing are data from small studies involving research procedures that are easily replicated or data from human subjects that might identify them.

PROTECTING THE RIGHTS AND PRIVACY OF HUMAN SUBJECTS

An important issue associated with the sharing of all data derived from human subjects is the protection of research participants' identities. The rights and privacy of people

who participate in NIH-sponsored research must be protected at all times. Sensitive data raise special concerns about confidentiality and the protection of subjects' privacy because of a greater possibility of harmful social, economic, or legal consequences if released. However, the collection of sensitive data does not preclude sharing. Indeed, some of the examples of sharing highlighted below include items on highly sensitive and, sometimes, illegal behaviors. But sensitive data call for a higher level of security during collection, analysis, and storage and special consideration when preparing datasets for broader use.

What constitutes "sensitive" data varies by context, population, and time. Illegal and sexual behaviors are almost always considered sensitive. Measures of alcohol use are less sensitive among adults than underage adolescents. Many diseases and medical conditions, such as bipolar illness or HIV infection, could be considered sensitive information. Because access to health insurance and employment can be affected by pre-existing conditions or even risk for certain diseases, information about medical conditions as well as genetic markers and family history, which may be used as indicators of predisposition, are also considered to be sensitive information.

There are two basic tools to protect from disclosure of sensitive data and subjects' identities: Restricting information in the dataset, and restricting access to the data. Thus, data intended for broader use should be free of identifiers that would permit linkages to the research participants and free of content that would create unacceptably high risks of subject identification.

Stripping a dataset of items that could identify individual participants is referred to by several different terms, such as data redaction, de-identification of data¹, and anonymizing data. It is rarely sufficient to simply remove names, addresses, telephone numbers, Social Security Numbers, and the like. Deductive disclosure of individual subjects becomes more likely when there are unusual characteristics or the joint occurrence of several unusual variables. Samples drawn from small geographic areas, rare populations, and linked datasets can present particular challenges to the protection of subjects' identities.

¹ Under the HIPAA Privacy Rule, de-identification of a dataset means removing the following variables: names; geographic information (including city, state, and zipcode); elements of dates such as those for birth, hospital admission and discharge, death; telephone numbers; fax numbers; electronic mail addresses; Social Security Number; medical record and prescription numbers; health plan beneficiary number; account numbers; certificate or license number; any vehicle identifier or serial number, including license plate number; any device identifier or serial number; Web Universal Resource Locator (URL); Internet Protocol (IP) address number; any biometric identifiers, including finger or voice prints; full face photographic images or any comparable images; and any other unique identifying number, characteristic, or code consisting of any segments of the previously listed identifiers.

There are many other methods currently used to anonymize data. Some investigators withhold parts of the sample; others block access to specific variables, especially items with low prevalence rates that make it easier to identify participants with unusual characteristics. Scientific interest in protecting subject identity is growing, and new methods are actively being developed. For example, investigators are creating synthetic datasets that mimic the characteristics of the original dataset without risking the identification of individual participants. It is beyond the scope of this document to instruct investigators about methods used to protect the identity of research subjects. However, several references are provided below. Investigators should also consult with statisticians to determine the best plan for data redaction and test the redaction process prior to the release of data.

Measures used to minimize the risk of breaching the confidentiality of data include the following:

- Mandatory agreements to maintain confidentiality
- Data encryption
- Electronic firewalls and locked storage facilities,
- Password authentication of users
- Audit trails
- Disaster prevention and recovery plans
- Security measures for backup tapes.

Institutions and investigators should work closely to develop and update plans and procedures to protect the security of data.

Data-use sharing agreements put limitations on who can use the data and how they are to be used. (These documents are also known by other names, such as license agreements, data-distribution agreements, and data-sharing agreements.) Such agreements contain different types of requirements, including those to protect the privacy of subjects and the confidentiality of the data. These documents can incorporate confidentiality standards to ensure data security at the recipient site and prohibit manipulation of data for the purposes of identifying subjects. They can stipulate that the recipient not transfer the data to other users, that the data are only to be used for research purposes, that the proposed research using the data will be reviewed by an IRB, and the like. Penalties for violating terms of the agreement are generally specified in these agreements. Below we describe some of the terms included in data-use sharing agreements used by archives and other entities that have shared data.

PROTECTING PROPRIETARY DATA

NIH encourages sharing of data generated with its support for further research, development, and application in the expectation that this will lead to products and knowledge of benefit to the public. However, NIH recognizes the need to protect patentable and other proprietary data and the restrictions on the sharing of data that may be imposed by agreements with third parties. In this regard, note that under the Bayh-Dole Act, grantees have the right to elect and retain title to subject inventions developed with Federal funding. Indeed, for inventions developed in its intramural

program, NIH does file patent applications in accord with a set of policies that are described at <http://www.nih.gov/od/ott/200po6.htm>. It is not the intent of the NIH statement on data sharing to discourage, impede, or prohibit the development of commercial products from federally funded research. However, it should be noted that, in general, NIH does not support the production of data that cannot be shared. If patent protection is being sought, data still can be shared in a timely manner.

EXAMPLES OF DATA SHARING

Data Archives

There are many archives for data. Many data archives facilitate the sharing of data using Web-based platforms. A searchable list of Websites for archives is available through the University of California at San Diego at <http://odwin.ucsd.edu/idata/>.

Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication. The **National Center for Biotechnology Information (NCBI)**, National Library of Medicine (NLM), NIH, was established in 1988 as a national resource for molecular biology information. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information with the goal of improving understanding of molecular processes affecting human health and disease. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit. For more information about submitting and downloading data, see the NCBI Website at <http://www.ncbi.nlm.nih.gov/Genbank/index.html>

The National Center for Chronic Disease Prevention and Health Promotion at CDC operates the **Youth Risk Behavior Surveillance System (YRBSS)**. This system provides data on six health risk behaviors among youth: unintentional injuries and violence, tobacco use, alcohol and other drug use, sexual behaviors, dietary behaviors, and physical activity. The YRBSS is composed of several surveys of different populations of youth, but focuses on national, State, and local school-based surveys of students in grades 9 through 12.

National YRBSS data are available directly through the Internet at the CDC Website. See <http://www.cdc.gov/nccdphp/dash/yrbs/>. Data from each wave of the national survey can be downloaded along with documentation files at <http://www.cdc.gov/nccdphp/dash/yrbs/data/index.htm> The data files for the most recent wave are available in ASCII, SAS, and SPSS. Documentation files are in PDF. User service is available by telephone and email. The YRBSS Website also contains a copy of the current questionnaire, item rationale, and results from previous waves. In addition, users can request a free CD-ROM with 6 years of compiled YRBSS data from the national, State, territorial, and local school-based surveys.

The CDC minimizes the risk of inadvertent disclosure of subjects by collecting the data anonymously. Participants complete a self-administered questionnaire in their regular classroom settings. Only four demographic variables are measured: Age, grade, race/ethnicity, and gender. School and classroom codes are not included in the datasets, so it is not possible to determine the school in which a student was enrolled.

Another example is **Sociometrics Corporation** (<http://www.socio.com/>), which maintains over 300 datasets from 200 different studies in seven topical areas: AIDS and other sexually transmitted diseases, disability, the American family, aging, adolescent pregnancy and pregnancy prevention, maternal drug abuse, and contextual influences on behavior. This archive has been in operation for more than 15 years. An expert panel selects the datasets included in Sociometric's library on the basis of scientific merit, substantive utility, technical quality, and potential for secondary analyses. For the cost of the data (approximately \$100 to \$225 per dataset if purchased individually, less if the entire collection is purchased), Sociometrics provides a complete data file in both CD-ROM and Internet formats, SPSS and SAS program statements, search and retrieval software, data summaries, detailed users' guides, and technical assistance.

The purchaser of data from Sociometrics can be either an individual or an institution. If the purchaser is an institution, an institutional representative must sign a license agreement certifying that only faculty, students, and staff can use the data. The license agreement further stipulates that neither printed nor electronic data may be copied or otherwise shared. Use of the data is restricted to statistical reporting, analysis, and teaching. The agreement prohibits the user from making any efforts to identify individual cases and prohibits linking data from this archive with individually identifiable data from other datasets. Violation of the license agreement carries civil liability.

The **Inter-University Consortium for Political and Social Research at the University of Michigan** has prepared an excellent set of guidelines for preparing data for archiving. While these guidelines were written with social science data in mind, they are broadly applicable. For further information, see <http://www.icpsr.umich.edu/>

Federated Data Systems

The **Biomedical Informatics Research Network (BIRN)** (<http://www.nbirn.net/>) is a National Institutes of Health (NIH) - National Center for Research Resources (NCRR)-sponsored initiative that fosters large-scale biomedical science collaborations by utilizing emerging cyberinfrastructure (high speed networks, distributed high-performance computing and the necessary software and data integration capabilities). The BIRN currently involves a consortium of 12 universities and 16 research groups participating in three "test bed" projects centered around the brain imaging of human neurological disease and associated animal models. Groups are working on large scale, cross-institutional imaging studies on Alzheimer's disease, depression, and schizophrenia using structural and functional magnetic resonance imaging (MRI). Others are studying animal models relevant to multiple sclerosis, attention deficit

disorder, and Parkinson's disease through MRI, whole brain histology, and high resolution light and electron microscopy. These studies are being used to drive the definition, construction, and daily use of a "federated data system." Federation presents biological data held at geographically-separated sites to appear as a single, unified and persistent data archive. Data is securely accessed across institutional boundaries to address issues of data privacy and automatic translation of data formats. Most of the groups participating in the BIRN have traditionally conducted independent investigations on relatively small populations, using site-specific software tools. The promise of the BIRN is the ability to test new hypotheses through the analysis of larger patient populations and unique multi-resolution views of animal models through data sharing and the integration of site independent resources for collaborative data refinement. This evolving "cyberinfrastructure" will enable researchers throughout the United States to collaborate on large-scale studies of human disease with unique, multi-resolution tools.

A Carnegie-Mellon project on parallel simulation of large scale neuronal models funded by the National Institute on Mental Health (NIMH) evolved to include methods for sharing computational models that is now at the heart of the Axiopie data sharing approach. The **Axiopie** project is based in the School of Informatics at the University of Edinburgh and is developing flexible tools for managing large volumes of metadata about local raw data (images, binary files, data on CDs, etc). Their goal is to allow scientists not only to organize their research data, but also to share data more easily with collaborators. Commercial data management and data sharing solutions incorporating ideas from the Axiopie project are now available from www.axiopie.com

Data Enclaves

Some data can be shared only under the most controlled conditions. If, for example, there is any risk of subject identification, the investigator may ask that users submit requests for specific analyses or come to the investigator's site to run analyses under supervision. Data enclaves were designed to deal with such situations.

One such enclave is the **Research Data Center at the CDC's National Center for Health Statistics (NCHS)**. The Research Data Center supports use of several NCHS restricted-use datasets through the Internet and within the Data Center itself. Additional information on the Research Data Center is available at <http://www.cdc.gov/nchs/r&d/rdc.htm>.

One of the datasets that can be used at the NCHS Research Data Center is a periodic survey called the National Survey of Family Growth (NSFG). Data from this survey provide an accurate statistical picture of family life, marriage and divorce, contraception, sexual experience, pregnancy, and infertility. Information concerning the NSFG is available at the NCSH Website at <http://www.cdc.gov/nchs/nsfg.htm>.

NCHS encourages the use of data from the NSFG, which are available through a variety of mechanisms. Public use datasets of the NSFG and other NCHS datasets are available free or at minimal cost after signing a use agreement. However, some NSFG data are not released in order to protect participants' identities. The restricted data, which are referred to as the NSFG contextual data file, do not include direct identifiers, such as name or social security numbers, but they may contain codes for small geographic units, such as blocks or census tracts. Thus, the restricted contextual dataset is only available to approved researchers via a remote access procedure or for analysis at the NCHS facility in Hyattsville, Maryland.

In order to gain access to restricted data, researchers must submit a detailed description of their projects. The proposal must include personal identification and institutional affiliation, a current resume, dates of proposed tenure at the data enclave, source of funding, a detailed summary of the proposed research including a statement of why publicly available data are insufficient, and a complete list of data requested, including data system, files, years, variables, and the like. NCHS staff are available for consultation on the proposal development. A committee consisting of NCHS staff, including the Confidentiality Officer, reviews all proposals. This review addresses the following critical questions: Does the proposed activity constitute statistical research or an illegal attempt to identify respondents? If it is research, is there any risk that respondents will be identified inadvertently?

All applicants are also required to sign an agreement of confidentiality. This agreement prohibits copying files or portions of files, keeping restricted materials, attempting to learn the identity of participants, removing any printouts, electronic files, or other documents from the enclave unless authorized by NCHS staff. In addition all papers or reports submitted for publication must first be submitted to NCHS for disclosure limitation review.

The fee charged for work at the data enclave (\$200 per day or \$1,000 per week) includes space, equipment, staff time for supervision and disclosure limitation review, and the creation and maintenance of data files required by the researcher. All work must be completed within the confines of the enclave. No electronic or hard copies of data can leave the facility unless they are submitted to a disclosure limitation review. In addition, researchers must work under the supervision of NCHS staff during normal working hours.

It should be noted that data collected by NCHS are protected by the Public Health Service Act (Section 308(d)). Under this section, identifying data can be disclosed or used for a purpose other than that for which it was supplied only if the person or establishment identified has consented.

Mixed Mode Sharing

The **National Institute of Mental Health (NIMH) Human Genetics Initiative** collects and distributes family data on schizophrenia, bipolar disorder, Alzheimer's disease, and

other mental disorders. Through the Initiative, qualified investigators can request clinical data, DNA samples, cell line cultures, and data derived from genotyping and other genetic analyses. Information on this initiative is available at http://zork.wustl.edu/nimh/NIMH_initiative/NIMH_initiative_link.html

Researchers can gain access to these data by successfully competing for an NIMH award specifically to analyze these data or by submitting an access request if they have no such award. Access certification is made on the basis of the experience and the scientific qualifications of the investigator. Requests must be submitted in writing on the letterhead of the sponsoring institution at which the research will be conducted and should include identifying information about the Principal Investigator and Coinvestigators, including curricula vitae. If access certification is obtained, biomaterials (DNA or cell lines) can be obtained at cost from the NIMH Center for Genetic Studies (<http://zork.wustl.edu/nimh/>). The Center serves as a data repository and management facility maintained under an NIMH contract.

All investigators must complete a Distribution Agreement, which includes a description of the research project to be conducted. The PI and an authorized representative of the recipient institution must sign the Distribution Agreement. The Agreement specifies that the investigator will only use the data and biomaterials for the specific project as described. The Agreement is not transferable to another recipient or facility, and biomaterials and data may only be shared with others by obtaining them directly through the NIMH center. The recipient must also agree to not attempt to establish the individual identities of subjects who provided the data or biomaterials.

When an access certification has been approved or a grant awarded, pedigree drawings are sent to the PI. Electronic files of clinical and genetic data and other information are available through password protected Websites. The investigator specifies which biomaterials are desired and sends this list to NIMH, which forwards it to the NIMH Center for Genetic Studies. The Center then provides shipping and payment instructions. All biomaterials and clinical data are stripped of personal identifiers. (No personal identifiers are ever received or handled by the Center.) The Center also provides periodic updates of data. Recipients share with the Center all genetic analysis data that they generate within 12 months of receipt of biomaterials or upon publication of research findings, whichever comes first. Upon completion of the project, the recipient must return all biomaterials as well as clinical and genetic data received from the Center or certify that the clinical and genetic data were destroyed in accordance with applicable laws and safety procedures.

The National Longitudinal Study of Adolescent Health (Add Health) is a very large national survey of students in grades 7 through 12. Data for this longitudinal study were collected in three waves. The first wave included questionnaires completed by 90,000 students and in-depth interviews with 20,745 adolescents. The resulting dataset includes items on a range of characteristics and behaviors of adolescents, including sensitive behaviors, such as alcohol use and sex.

From the outset of this study, the investigators planned to share the data. Critical to the protection of subjects is the separation of identities from the data, which occurs immediately after data collection. Only a Security Manager can link the name and address of respondent to interview data. The investigators also asked for and received a Certificate of Confidentiality from DHHS to protect subjects' identities. All Add Health staff are required to take training in data confidentiality and security issues. Individuals and institutions seeking to obtain the Add Health data are encouraged to develop and implement a similar training program. Details about accessing the data are available on the study's Website at <http://www.cpc.unc.edu/addhealth/>. Only certified researchers are permitted access to Add Health data. Thus, the informed consent document notes that the study "is helping researchers understand the health of young adults and the behaviors that affect their health."

Because of the extremely sensitive nature of some of the information collected, the investigators have made data or portions of the data available in three ways: (1) a public use dataset that can be accessed through a data archive; (2) a restricted access contractual dataset, and (3) access at a data enclave at the Add Health facility under the supervision of staff.

The public use dataset includes only a subset of respondents to protect the identity of participants. The investigators found that even with more than 90,000 cases, a cross-tabulation of 5 variables could distinguish an individual record. Therefore, half of the sample was chosen at random for the public use dataset with an oversample of minority adolescents. CD-ROMs are distributed through a data archive run by Sociometrics Corporation (<http://www.sociometrics.com/>). The data are in ASCII format and can be analyzed with several standard statistical packages.

The restricted access dataset is available only to certified researchers who provide a nonrefundable fee to cover administrative handling charges and user support. Add Health investigators have embedded a hidden signature identifying the purchaser in each electronic file, so that unauthorized copies can be traced. All users must sign an agreement to maintain privacy of subjects and confidentiality of the data. In addition, users must certify that they have complied with a set of security requirements covering how the data are handled and stored. These requirements are updated periodically to reflect changes in computer technology. Applicants are also required to submit letters from their IRBs verifying and approving plans for data security and for minimizing risks of deductive disclosure. The staff from Add Health conducts site visits to monitor the use of these data at outside institutions. The user fee covers the cost of these visits.

Researchers requesting use of data that cannot be shared through contractual agreements must come to the Add Health site at the University of North Carolina in Chapel Hill to conduct analyses under the supervision of Add Health staff. Again, these data are only available to certified researchers.