

**National Institutes of Health
National Institute on Aging
Behavioral and Social Research Program**

**Data Sharing Workshop for Behavioral and Social Studies
That Collect Genetic Data**

**August 2–3, 2006
Bethesda, Maryland**

Workshop Highlights*

I. Background and Purpose

The National Institute on Aging (NIA) Behavioral and Social Research (BSR) Program is developing genetic and genomics research directions on ongoing and new studies within the BSR portfolio of research activities. For some studies, DNA already is being collected, and for others, plans for collection have been established. Because procedures surrounding sharing of DNA samples are new to many BSR-supported researchers, developing efficient and feasible data sharing plans is imperative. Using already established data sharing plans associated with ongoing biobank projects and initiatives is a logical first step. However, the nature of behavioral and social sciences data may raise new or unique issues (e.g., deductive disclosure, confidentiality, and privacy) that must be carefully considered. For example, there is great interest in linking survey data to administrative records (e.g. Social Security, Medicare records). Linking across data sets should result in raised confidentiality and security concerns and a more restricted data access model.

To explore issues surrounding data sharing plans for behavioral and social research studies that collect human specimens, the NIA/BSR convened a workshop from August 2 to 3, 2006, to provide a forum for input by representatives of major NIA/BSR-funded surveys that are either engaged in or considering the collection of human specimens, as well as selected investigators of longitudinal studies with existing repositories. The workshop agenda and list of attendees are included as Attachments 1 and 2. Prior to the workshop, NIA staff held initial discussions with several NIA/BSR-funded principal investigators (PIs) to discuss the relevance of this effort for their research, identify special areas of concern, and ascertain specific information to be collected from ongoing studies regarding their data sharing procedures. Subsequently, interviews were conducted with principals of the following six studies: The National Longitudinal Study of Adolescent Health (Add Health); the Framingham Heart Study; the Study of Women's Health Across the Nation (SWAN); the National Health and Nutrition Examination Survey (NHANES); the Dynamics of Health, Aging, and Body Composition (Health ABC) study; and the NIA

* This document was prepared by Mariana González del Riego, Rose Maria Li, and Virginia Lerch, Rose Li and Associates, Inc., under contract to the National Institute on Aging (263-MD-516853 and 263-MD-417248). The information contained in this report reflects the views expressed by workshop participants and are not necessarily those of the National Institute on Aging, National Institutes of Health. Comments provided on earlier drafts by John Haaga are gratefully acknowledged.

Alzheimer's Initiative.¹ In advance of the meeting, workshop participants received a detailed table with the following key elements from existing data sharing plans (Attachment 3):

- Description of biological samples and data that are shared,
- Sensitivity of and access to phenotypic and genotypic data,
- Safeguards to ensure confidentiality,
- Data sharing procedures,
- Informed consent, and
- Institutional review board (IRB) considerations.

The workshop presentations and discussions were intended to inform the development of an NIA/BSR data sharing plan. Workshop organizers asked participants to share (1) their particular concerns and potential solutions for sharing data within their own studies, and (2) how elements of already existing data sharing plans would be adequate/useful or problematic for addressing their concerns.

Dr. Jerome Reiter (Duke University) provided an overview of key disclosure risk types relating to biomedical and genetic data in social science research, and he described tiered access to confidential data as a potential solution to such risks (Attachment 4). The four levels of access Dr. Reiter discussed are as follows:

- (1) General public-use data—Data are altered before dissemination. For example, variables are coarsened; small amounts of data are exchanged between records (swapping) to introduce uncertainty while preserving the marginal relationships within the data set; noise is added to continuous variables; or synthetic data derived from a probability model replace the original values, but relationships in the original data are preserved.
- (2) Licensing researchers—Researchers are required to be licensed to obtain more data. This approach is commonly used by several agencies, such as the U.S. Bureau of Labor Statistics and the National Science Foundation. While practical, this approach is not always appropriate for sensitive data. Most violations result from failure to follow protocols. Although substantial penalties are involved in some cases (e.g., investigator and institution forfeit funding from agency sponsoring grant), enforcement mechanisms are weak. Nevertheless, there has been no evidence of confidentiality breaches.
- (3) Remote access systems—A computer provides results of analyses without revealing an individual's data to the investigator. These systems are used by agencies such as the National Center for Health Statistics and the U.S. Census Bureau. Although secure computation techniques are applied, disclosure risks can occur as a result of judicious queries.
- (4) Data enclaves—Research is conducted in a secure data enclave. Data enclaves are challenging to establish and maintain, especially for individual researchers, and less convenient for researchers.

Even with tiered access, heightened risk of disclosure is possible depending on the context of the data. For example, geographic information may be known, as in the Framingham Heart Study. In

¹ The National Social Life, Health, and Aging Project (NSHAP) currently collects specimens but was not included in the interviews because it does not have long-term experience implementing data sharing plans. Dr. Stacey Lindau, NSHAP Co-PI, was invited but was unable to attend the workshop.

Add Health, knowledge of individuals in the study cohort may be widespread (e.g., an entire school is enrolled in a cohort). Individual-level data could be identifying when genetic information is known by others, and data that are secure in one data set prior to linkage may no longer be secure after linkage.

As disclosures of biomedical/genetic information may cause serious harm, altering or restricting genetic and biomedical data to protect confidentiality—which may compromise data quality—is often necessary. At the same time, shared data must produce reasonably similar analysis results to that of the original data. Consequently, data sharing plan recommendations should emphasize the need to consider the quality of the data to be released and the need for its documentation.

Dr. MaryFran Sowers (University of Michigan) provided an overview of the advantages and disadvantages of a centralized versus distributed biorepository (Attachment 5). Strengths of centrally located repositories include their central accountability, defined organizational responsibility, available resources and specialized services, and effective information technology systems. The efficiency of this model stems from development of a single protocol, centralized training, and an appropriate information technology system. The central biorepository configuration does have a number of limitations. These systems often will not work in settings with already existing structures, the trust or rapport that distributed biorepositories may develop may not be engendered by this configuration, and IRBs and U.S. academic systems through which material transfer agreements (MTAs) are negotiated are not centrally located. Strengths of distributed configurations include their ability to adapt to multiple systems through which IRBs and MTAs are administered in the United States and their flexibility to bring selective resources to bear if central resources or structures are absent. This model works best if resources are available as incentives to collaboration. However, it is frequently cumbersome and inefficient due to greater managerial complexities and uneven levels of technical proficiency across sites.

Traditional repository functions have been limited to archiving specimens and storing selective information about the samples. The increasing expectations for data sharing reflect a new paradigm with funding and management implications. The realities of IRBs, intellectual property considerations, MTAs, and existing study protocols frequently require more costly, less efficient biorepository configurations. In the case of SWAN, nearly 12 percent of the total study funding has been invested in the biorepository. In general, biorepository costs associated with utilization (such as data sharing, the review process, and reutilization procedures) are very difficult to determine and should be separated into aspects of data sharing and storage.

II. Meeting Discussion Highlights

Two prominent themes emerged from the workshop: The uniqueness of social and behavioral research studies resulting from the merging of rich phenotypic and biologic data and the need to consider this distinctiveness during the development of an NIA/BSR data sharing plan. Topics of greatest interest to workshop participants included sensitivity of and access to phenotypic and genotypic data, safeguards to ensure confidentiality, and sharing procedures. Topics of concern to participants included familial data and the lack of third-party consent, transfer of international samples and timely return of unused samples, and funding mechanisms for data and specimen repositories beyond the life of an initial sponsored research project.

In general, NIA/BSR studies have complex, deeply described phenotypes with disclosure risks. In turn, genetic data collected by these studies are more likely to represent the variation in the population than biologically centered studies. Once genetic information becomes available in conjunction with these rich phenotypes, some third parties might consider snooping on these data a worthwhile activity. Even without the genetics component, the manner in which rich phenotypic, sensitive data are shared responsibly is still an important issue. Data sharing policies and procedures for NIA/BSR-sponsored studies must be considered in this broader context. A more thorough understanding of the rates of data utilization and the findings made possible through data sharing should be developed.

Highlights of the workshop discussion are summarized below.

Sensitivity of and Access to Phenotypic and Genotypic Data

Sensitivity is a dynamic characteristic that depends on (1) whether or not a record can be identified; (2) the sample size, for which large and small samples have different determinants of sensitivity; and (3) the availability of other data with which variables can be combined. In deciding which variables should be restricted to reduce the likelihood of disclosure (i.e., which variables are sensitive), a dichotomy between identifying variables (e.g., age and geography) and sensitive variables (e.g., disease status which, when revealed, could cause embarrassment or harm) should be established. Then, a decision should be made regarding whether the concern is identification disclosure, attribute disclosure, or a combination thereof. Assuming the main concern is identification disclosure, access to age and geography data should be restricted. Because it is difficult to anticipate what other information a potential snooper may be able to access, the type of snooper the data must be protected against should be determined before making data restriction decisions. Sample size is not critical to the assessment of disclosure risk; more critical is the information in the data file available to the potential snooper.

Disclosure risk and data sensitivity are separate but related concerns. Variables such as drug use and disease status constitute sensitive data because they could cause harm to an individual if they were revealed. At the same time, because these characteristics might occur in limited numbers within a population, disclosure risk is elevated when such a variable is both known about an individual and observed in a study sample. Although the probability of disclosure is used to assess data sensitivity levels, the concern is not the probability of identification but rather the

expected harm from identification. Difficulty in measuring the latter has led to the general use of the probability of disclosure as a proxy for data sensitivity.

Important issues to consider when establishing a data sharing plan include (1) whether sharing of sensitive data is worthwhile for any particular study and (2) who will use the data and for what purpose. For example, geneticists seeking to analyze only DNA would not require access to other variables. Therefore, access to a minimum data set would be sufficient. Although Neuroscience and Neuropsychology of Aging (NNA)–funded investigators are required to share a predetermined minimum data set, this is not the case in many of the studies under discussion where nonaffiliated investigators are requesting access to numerous variables. Among the six studies that were surveyed prior to the workshop, a range of approaches was observed; some did not classify data as sensitive or risky while others did. None of the studies reported a disclosure breach, which raises the question of whether deductive disclosure constitutes a true risk.

More Empirical Evidence Needed in Assessment of Disclosure Risks

Currently, the guidelines for determining levels of variable sensitivity rest on nonempirical evidence. In the absence of extensive empirical research on disclosure risks, organizations such as the Inter-University Consortium for Political and Social Research (ICPSR), which provides public access to a vast archive of social science data without a review process, have little to guide decisions about the type of data that should not be shared.

Develop Guidance for Investigators To Ensure Appropriate Use

Genetic components are being added to studies across the NIA, and there is great variability in the richness of phenotypic data between and within data sets. Therefore, the NIA/BSR should provide suggestions on how certain phenotypes should be approached for public data sets. This would help ensure that the data are used appropriately. The NIA/BSR also should develop guidance about how to evaluate a proposed data set for public release and should establish collaborations with viable partners to generate formal recommendations to help guard against misuse of data. Partners in this effort could include principals of the Health and Retirement Study, Add Health, and NHANES, all of which move data sets from public to restricted access once linkage to administrative data takes place. Study investigators, outside experts who do not have a vested interest in the study, and community members should play an important role in this process as well.

Investigate Appropriateness of Existing Methods for Genetic Data Release

Methods of coarsening, adding noise, or using synthetic values in publicly released data sets (particularly genetic data sets) should be investigated. To this end, methods utilized by the Genetic Association Information Network (GAIN) initiative, a public-private partnership of the Foundation for NIH, should be explored as a useful model for disseminating data from NIA/BSR studies safely, rapidly, and equitably. However, it is unknown whether the dissemination procedures for GAIN were determined through disclosure risk analysis or mere intuition.

Sharing Procedures for Genetic Samples and Phenotypic Data

Harmonization Encouraged With NIH and Other Efforts

As experience with the NIH Data Sharing Policy increases, the NIA/BSR should be cognizant of new or evolving considerations, which may ultimately impact the data sharing plans to be established at the Institute and Center (IC) level. For instance, the NIH currently is considering how to manage the enormous amount of data generated by large studies. One of the main concerns is that there is not a good understanding of where the data are being housed. A possibility under review is the establishment of data warehouses. Additional issues under consideration include informatics needs and infrastructure, identification of data gatekeepers, and IRB implications. As the NIH Data Sharing Policy evolves, the NIH also is considering whether to incorporate data sharing into the scoring of research applications.

Existing data sharing enforcement mechanisms and their effectiveness should be well understood. For instance, the NIA relies on its ability to restrict grant funding in cases where investigators withhold samples and/or findings. In contrast, the National Institute of Mental Health (NIMH) uses the contract mechanism to enforce data sharing; PIs are required to submit their data to an NIMH database with access available only to grantees supported by that Institute.

The NIA/BSR should consider unique aspects of biorepository rules established by non-NIH resources. For instance, some NIA/BSR studies may be located at the U.S. Department of Veterans Affairs (VA) facility, in which case VA repository approvals will be required to extract and deposit samples.

MTAs Must Be Explicit in Types of Genotypes for Analyses

To avoid situations where researchers conduct further genotyping beyond that specified in the approved research application or proposal, MTAs must specify which genotypes will be analyzed. In the case of funding agencies, withholding funding is an effective way to enforce the submission of research results and related reports. For studies that grant external researchers access to their data, the NIA study PI should contact his/her Program Officer to gain leverage when dealing with noncompliant external investigators. If these steps fail to resolve the situation, the Program Officer or the PI should contact the Office of Extramural Research to report the incident. Based on the Alzheimer's Initiative experience, MTAs should not be negotiated centrally but rather at the level of the institution.

Centralized Data Sharing Functions and Services Favored

In developing a data sharing plan, the NIA/BSR should consider establishing centralized functions and services such as standard methods and expert advisory panels. For example, the NIA/BSR could support multiple high-quality data archiving and dissemination centers to provide the services and expertise required by individual studies. These services should be regularly competed or reviewed to maintain up-to-date processes associated with data sharing. Centralization of data sharing services would represent significant cost savings and efficiencies to the Institute as institutions would not need to be funded to maintain such expertise individually.

Scientific Review of Applications Key to Appropriate Sample and Data Access

The scientific review of an application for data access will certify that:

- (1) The appropriate expertise is incorporated into the application review process to ensure that returned data are of the highest quality possible, particularly for samples that are irreplaceable;
- (2) Due to the nature of phenotypic data collected in NIA/BSR-funded studies, data input from the original PI is provided to ancillary studies to avoid data misinterpretation;
- (3) Proposed genetic analyses to be conducted on publicly available samples and data are appropriate; and
- (4) Ethical issues are considered adequately.

If an investigator has IRB approval and is funded to conduct the proposed research, a review by the core study advisory committee should focus less on scientific merit and more on ethical issues regarding the analyses. However, if the samples requested are in limited supply, additional evaluation for scientific merit or research significance may be appropriate. Having mechanisms in place to ensure that samples in limited supply are used for the maximum benefit of the entire scientific enterprise should be part of a data sharing plan. For example, rare samples should not be used for assay development.

Determining when assays are ready to be tested in a nonreplenishable population resource is critical, particularly when conflicting information exists about such assays. Rather than expending limited resources, the NIA/BSR should encourage investigators to resolve such conflicting information by (1) reviewing the literature to determine why conflicting information exists about those assays and (2) conducting methods research in a different, dispensable and presumably lower quality population resource to determine which approach is best.

It is important to establish communication and collaborative efforts among investigators to minimize unnecessary duplication of effort that wastes nonreplenishable material. This also helps to ensure that appropriate replication and comparisons are made to assess the reliability of reported findings (e.g., analyzing the same gene on an Affimetrix versus Illumina system).

When an investigator with banked DNA decides, after the original application has been funded, to undertake previously unspecified analyses of samples or to make samples available to the research community, the NIA should require scrutiny of the research plan through a peer review process to ensure that the requisite expertise is recruited to adequately conduct the analyses.

Safeguards To Ensure Confidentiality

Due to the expectation that behavioral and social research data will be used increasingly by researchers in other fields (e.g., genetics), the NIA/BSR should identify and establish appropriate safeguards for ensuring confidentiality and security and, in particular, server security while data sharing guidelines are under development.

Consider Legislative Protections

Because safeguards to control disclosure do not provide absolute protection to research participants, the need for legislative protections that would provide legal repercussions against data abusers was considered. However, establishing legislative protections could create

significant moral dilemmas for researchers (e.g., the inability to release genetic findings to family members seeking the information).

Minimize Risk of Familial Data Disclosure

Many behavioral and social research studies contain family structures in their data. Therefore, the possibility of exposing information about family members not consented for the original study is another important topic worthy of consideration. Investigators will need to reassure study participants that researchers using restricted access data abide by ethical and moral uses of data and face strict penalties for any breaches. The NIA/BSR should approve data sharing plans only in an environment with effective strategies to evaluate public release safety. In the interim, all data should be restricted until mechanisms are developed to allow public release of data sets that are adequately screened for minimal risk of disclosure, including familial data disclosure.

Employ Tiered Access for Public Use Files

A tiered access system will reduce and alleviate concerns about deductive disclosure by restricting and supervising data access. The system offers flexibility that will make as much of the data as possible available to the public. It will accommodate differences in study design by making certain components of the study (e.g., a subsample that is located in one city) more or less restricted. A tiered system also will allow for the restriction of particular variables and combinations of variables, such as linkage to administrative data. The NIA/BSR should give serious consideration to the following possibilities: (1) Matching of variables between early and subsequent waves that would identify an individual, and (2) difficulty in restricting use once the data are publicly released.

In core studies where PIs oversee the data access process, a complicated tiered data access system can make it difficult for researchers with legitimate research questions to access the data. Therefore, a balance between disclosure prevention and “data hoarding” must be reached. The application of a centralized, third-party model, where the individual responsible for managing the data access process does not have a research agenda, can help achieve this balance. To ensure that such a third party continuously applies new best practices, the NIA/BSR should consider creating a sense of competition via distributed centralization.

Although establishing different tiers of access for shared data likely will be necessary for NIA/BSR purposes, decisions about the types of data assigned to each tier remain to be made. A research agenda on disclosure risk analysis should be pursued to progress from intuitive answers to research-based approaches for handling such risks. The NIA/BSR should establish separate grantee agreements for disclosure risk analysis versus specialized biorepository support.

Informed Consent

Greater Specificity, Consistency, and Clarity Needed

The increasing complexity and length of consent forms are counterproductive to an informed consent process. Although studies with genetic components require additional explanation, standard consent language does not exist, and the manner in which the topic is addressed in consent forms is inconsistent. Studies vary extensively with respect to how prescriptive their informed consents are about the future use of samples and/or data. The NIA/BSR should

recommend language that specifically addresses these issues and should provide informed consent form templates that can be adapted to conform to State laws and local policies and that are likely to be approved by the appropriate IRB. To inform these approaches, the NIA/BSR should support empirical research to (1) determine participant comprehension of consents (i.e., whether current forms convey the intended message), (2) understand participants' desires regarding sample uses, and (3) assess differences among racial and ethnic groups.

Most people are willing to share their specimens and data for research purposes as long as they are informed about present and future use. However, the vast majority of informed consent forms are silent on possible subsequent sharing of samples and data outside the original research group. Two distinct interpretations of this are possible: Either any amount of sharing is feasible, or no amount of sharing is feasible. To avoid this ambiguity, consent forms must describe sample and data sharing explicitly so study participants understand the process and the potential ramifications. Participants should be given the opportunity to opt out from sharing their samples and data with investigators not affiliated with the study (i.e., pharmaceutical industry) for which they are being consented.

Sharing for Commercialization Purposes Creates Sensitivity Issue

The possibility of commercializing a product based on samples and/or data provided by a study is usually addressed in consent forms. This possibility should be considered a sensitivity issue since study participants may be more easily identified when they share certain levels of data for commercialization purposes. The NIH Data Sharing Plan does not restrict the use of data from or by commercial entities as it would be counterproductive, particularly in the case of Small Business Innovation Research Awards. Automatically including such a restriction in informed consent forms would be questionable. However, providing a restricted, minimum data set to commercial companies would be a potential approach. Implementation of any NIH-wide product commercialization policy is expected to be made at the IC level, much like the implementation of the NIH Data Sharing Policy.

IRB Considerations

IRB demands for investigators to address biospecimen storage locations and types of users are likely to increase with the linkage of genetic data to rich phenotypic data. This is because of the greater risk of identification, particularly for minority and vulnerable populations. This risk will have to be incorporated into informed consent forms and, ultimately, will make data sharing much more difficult. Although no NIA/BSR-supported study has reported a disclosure breach, it is conceivable that IRBs may require the establishment of a protocol to deal with such breaches should they occur. Therefore, the NIA/BSR should adopt as a best practice appropriate disclosure safeguards and a standardized response to breaches.

Proactive Collaboration With IRBs Critical

Based on existing and anticipated IRB challenges, the NIA/BSR should encourage more effective approaches to providing guidance to IRBs on how to establish successful data sharing plans. The NIA/BSR also should build on efforts of the IRB outreach group—established by the U.S. Department of Health and Human Services to analyze repository policies and biospecimen use—to educate IRBs on these issues. Journals such as the *Journal of Empirical Research on*

Human Research Ethics also could be used for IRB education purposes. Finally, harmonization initiatives within the NIH focusing on biorepositories to promote uniform practices for repository development and operation (<http://www.capconcorp.com/crpac2006/>) should be exploited. Taken together, these strategies can help promote uniformity across IRBs with respect to levels of sharing and agreement on the appropriate use of samples.

Need for Investigator Support

Time and resource demands relating to data sharing are significant. For example, SWAN has two full-time employees who process applications, prepare data sets, and answer questions, and the study uses a commercial firm to manage the biorepositories and fill sample requests. The English Longitudinal Study of Ageing also has contracted a commercial firm to perform similar duties. Investigators should not be required to spend time and effort applying for additional support for the required maintenance and sharing of a data set. Such costs should be anticipated and incorporated into the original research grant application or into subsequent agreements possibly cofunded through public-private partnerships.

Consider Centralizing Some Support Functions

A plausible, alternate approach to an NIA/BSR funding “set-aside” may be to recruit specialists to handle data sharing considerations for multiple studies in a centralized manner, perhaps by establishing an ICPSR-like system, as opposed to having each study attempt to recruit such proficiency individually. A similar source of expertise may need to be sought for studies that collect both biospecimens and data. In fact, commercial firms have evolved over the past several years to deliver a dynamic product rather than a static “freezer” product.

The NIA/BSR should play a role in helping behavioral and social scientists identify and access appropriate literature, expertise, and other resources that can facilitate their pursuit of complementary biologic research. This would ensure that the research conducted is of the highest quality, data are reproducible, and the appropriate standards are being used. The possibility of providing the infrastructure to archive and curate such specimens (e.g., establishing central biorepositories) should be revisited because creating such a resource is typically not among the main goals of many NIA/BSR-funded research projects.

Establish Reliable Funding Mechanisms To Sustain Sample and Data Sharing

A funding mechanism that extends beyond the life of the original study grant (i.e., R01 or P01) is needed to ensure that data sharing is continued. Although PIs are responsible for keeping data available for 3 years after the end of the original study award, this period of time is inadequate for some of the social and behavioral studies in question. The NIA/BSR can award R03 grants for data dissemination and archiving, and these grants can be awarded to individuals other than the original study PIs. Other approaches for supporting sustained data dissemination and archiving should be investigated.

Improve Quality of Returned Data

It is important to conceptually distinguish between (1) primary or “raw” data generated by core studies such as Add Health that are then shared for analysis and (2) nonprimary data or high-throughput data (e.g., genotype data) generated by a “broker” based on primary data.

Nonprimary data may be returned to the core study to be made broadly available to the research community. Returned data typically are perceived as being of lower quality than those data generated by the core study because of issues of quality assurance (QA) or quality control (QC), laboratory variation, and related factors over which the core study has no control. In the case of SWAN, returned data are made available to the research community after being labeled as such. Documentation of the returned data, including how the analyses were conducted and what QA/QC procedures were performed, also are documented. This information is made available to researchers, but the process requires a significant amount of time, energy, and informatics support.

Other Considerations

Wholesale Adoption of Existing Data Sharing Models Not a Viable Solution

The NIA/BSR should not adopt any existing data sharing model in its entirety. Several of the models discussed during the workshop have specific characteristics that are in place due to the nature of the original study. Although certain elements of these data sharing plans could be applied to a broad range of NIA/BSR-funded studies, they should not be adopted *in toto*.

Existing Documents Will Better Inform the Development of an NIA/BSR Data Sharing Plan

The International Society for Biological and Environmental Repositories (ISBER) “Best Practices for Biorepositories I” (March 2005) reflects the collective experience of its members to provide repository professionals with a comprehensive foundation for the guidance of repository activities (<http://www.isber.org/ibc.html>). In April 2006, the Office of Biorepositories and Biospecimen Research, National Cancer Institute (NCI), issued First-Generation Guidelines for NCI-Supported Biorepositories to accompany the ISBER Best Practices guide (http://biospecimens.cancer.gov/biorepositories/guidelines_full_formatted.asp). Participants highly recommended that the NIA/BSR review these documents prior to the development of a data sharing plan.

Building Public Trust Critical

Steps taken by other ICs and NIA divisions for obtaining buy-in from the community can provide working models for the NIA/BSR. For instance, components of the NIA/NNA model include a communications group at the Institute level and an Alzheimer’s Disease Education and Referral Center. These groups have conducted focus groups in the Alzheimer’s community aimed at determining how to present the genetics initiative. Such focus groups, in addition to providing news outlets with information relating to Alzheimer’s disease, have worked extremely well to raise awareness about the disease and what families can do to help (i.e., participating in the genetics initiative and understanding the benefits of data sharing). Applying similar approaches to communities targeted by NIA/BSR-funded studies would help build public trust, which in turn would enhance the sample and data sharing process and advance behavioral and social science research.

Building public trust in minority and vulnerable populations is a priority because consent rates are known to vary greatly across these communities and tend to be significantly lower than those for the majority population. The NIA/BSR should consider support of multifaceted interventions

including communication and outreach as well as empirical research to measure the effect of unspecified use of samples and data on the willingness of minorities to consent to studies.

III. Next Steps

Based on the workshop discussion, existing data sharing plans will be drawn upon to shape the development of an NIA/BSR data sharing plan. The NIA/BSR's objective is to generate first-generation sample and data sharing guidelines that will undergo multiple iterations based on internal and external feedback. Ultimately, the goal is to remove the burden associated with the implementation of a sample and data sharing plan for individual investigators, provide them with useful information and services, and maximize researcher access to NIA/BSR-supported data sets.

Studies in the NIA/BSR research portfolio present a unique challenge due to the volume and richness of phenotypic data collected. The implications of linking genetic data to such information have unknown consequences; some variables currently not considered to be sensitive may become sensitive in future combinations. Thus, a tiered access protocol is critical for public use data files. Likewise, robust protection systems that obligate investigators to confidentiality and security are needed. Secure data enclaves, while effective, are widely underused, and access systems that allow researchers to prepare synthetic data remotely and submit them to a secure site for the actual analyses should be considered.

Attachment 1

**National Institutes of Health
National Institute on Aging (NIA)
Behavioral and Social Research (BSR) Program**

**Data Sharing Workshop for Behavioral and Social Studies that Collect
Genetic Data**

**Gateway Building 5th Floor Conference Room
7201 Wisconsin Avenue
Bethesda, MD**

August 2-3, 2006

AGENDA

Wednesday, August 2, 2006

- | | |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1:00 p.m. – 1:15 p.m. | <i>Welcome and Introductions</i>
Jennifer Harris |
| 1:15 p.m. – 3:15 p.m. | <i>Review and Discussion of Current Approaches</i>
(Review of findings from Add Health, Alzheimer's Initiative, Framingham Heart Study, Health ABC, NHANES, and SWAN as background for workshop discussions and points for formulating a BSR data sharing plan).
Jennifer Harris and Teresa Seeman <ul style="list-style-type: none">• <i>Biological Samples and Data Collected</i>• <i>Sensitivity of and Access to Phenotypic and Genetic Data</i>• <i>Sharing Procedures for Genetic Samples and Phenotypic Data</i>• <i>Safeguards to Ensure Confidentiality</i>• <i>Informed Consent</i>• <i>IRB Considerations</i> |
| 3:15 p.m. – 3:30 p.m. | <i>BREAK</i> |
| 3:30 p.m. – 4:00 p.m. | <i>Identification Disclosure and Social Science Research</i>
Jerome Reiter |
| 4:00 p.m. – 4:15 p.m. | Discussion |
| 4:15 p.m. – 5:00 p.m. | <i>NIA BSR Principal Investigator Perspectives and Concerns</i>
Round table discussion |

5:00 p.m. *ADJOURN*

6:30 p.m. – 8:00 p.m. *GROUP DINNER*

Thursday, August 3, 2006

7:30 a.m. – 8:00 a.m. *CONTINENTAL BREAKFAST*

8:00 a.m. – 8:15 a.m. *Recap of NIA BSR Principal Investigator Perspectives and Concerns*
Teresa Seeman

8:15 a.m. – 8:45 a.m. *Centralized Versus Distributed Repository Models: Who Has Control?*
MaryFran Sowers

8:45 a.m. – 9:00 a.m. Discussion

9:00 a.m. – 1:45 p.m. *Proposal of Principles and Approaches for Setting BSR Data Sharing Guidelines and Procedures*
Chairs: Jennifer Harris and Teresa Seeman
Workshop Participants

10:00 a.m. – 10:30 a.m. *BREAK*

12:00 p.m. – 12:40 p.m. *WORKING LUNCH*

1:45 p.m. – 2:00 p.m. *Wrap-Up and Closing Comments*
Jennifer Harris and Teresa Seeman

2:00 p.m. *ADJOURN*

Attachment 2

**NATIONAL INSTITUTE ON AGING
BEHAVIORAL AND SOCIAL RESEARCH PROGRAM**

**DATA SHARING WORKSHOP FOR BEHAVIORAL AND SOCIAL STUDIES
THAT COLLECT GENETIC DATA**

Bethesda, Maryland
August 2-3, 2006

Attendees

Peg (Marguerite) Barratt, Ph.D.

Deputy Director
Clinical Research Policy Analysis and
Coordination (CRpac)
Office of Science Policy
Office of the Director
The National Institutes of Health

Constance Citro, Ph.D.

Director
Committee on National Statistics
The National Academy of Sciences

Richard Kulka, Ph.D.

Senior Vice President of Strategic Business
Development
Abt Associates, Inc.

Meena Kumari, Ph.D.

Senior Research Fellow
Department of Epidemiology and Public Health
University College London

Jerome Reiter, Ph.D.

Assistant Professor of the Practice of Statistics
and Decision Sciences
Institute of Statistics and Decision Sciences
Duke University

Carol Ryff, Ph.D.

Professor
Department of Psychology
University of Wisconsin-Madison

Eleanor Singer, Ph.D.

Research Professor
Survey Research Center
Institute for Social Research
University of Michigan

MaryFran Sowers, Ph.D.

Professor
Department of Epidemiology
University of Michigan

Miron Straf, Ph.D.

Deputy Director
Executive Office
Division of Behavioral and Social Sciences and
Education
The National Academy of Sciences

Tony Tse, Ph.D.

Program Analyst
CRpac
Office of Science Policy
Office of the Director
The National Institutes of Health

Robert Wallace, M.D., M.Sc.

Professor
Departments of Epidemiology and Internal
Medicine
The University of Iowa

Maxine Weinstein, Ph.D.

Professor
Center for Population and Health
Georgetown University

NIA Staff and Contractors:

Dallas Anderson, Ph.D.

Program Director
Population Studies
Dementias of Aging Branch
Neuroscience and Neuropsychology of
Aging Program

Angie Chon-Lee, M.P.H.

Health Program Specialist
Behavioral and Social Research Program

Mariana González del Riego, Sc.M.

Project Manager
Rose Li and Associates, Inc.

John Haaga, Ph.D.

Deputy Director
Behavioral and Social Research Program

Jennifer Harris, Ph.D.

Senior Researcher
Department of Genes and Environment
Division of Epidemiology
The Norwegian Institute of Public Health

Tamara Harris, M.D., M.S.

Chief
Geriatric Epidemiology Section
Laboratory of Epidemiology, Demography, and
Biometry

Virginia Lerch

Project Coordinator
Rose Li and Associates, Inc.

Marilyn Miller, Ph.D.

Program Director
Etiology of Alzheimer's Disease Genetics, Tau,
and Reproductive Hormone Research
Neuroscience and Neuropsychology of Aging
Program

Marcelle Morrison-Bogorad, Ph.D.

Director
Neuroscience and Neuropsychology of Aging
Program

Lisbeth Nielsen, Ph.D.

Program Director
Psychological Development and
Integrative Science Branch
Behavioral and Social Research Program

Georgeanne Patmios, M.A.

Acting Chief
Population and Social Processes Branch
Behavioral and Social Research Program

Anthony Phelps, Ph.D.

Program Director
Alzheimer's Disease Centers
Neuroscience and Neuropsychology of Aging
Program

Teresa Seeman, Ph.D. (IPA)

Professor of Medicine and Epidemiology
Associate Chief for Research
Division of Geriatrics
Geffen School of Medicine
University of California, Los Angeles

Sherry Sherman, Ph.D.

Program Director
Clinical Aging and Reproductive Hormone
Research Program

Nina Silverberg, Ph.D.

Assistant Program Director
Alzheimer's Disease Centers
Neuroscience and Neuropsychology of Aging
Program

Erica Spotts, Ph.D.

Health Scientist Administrator
Behavioral and Social Research Program

Data Sharing Workshop for Behavioral and Social Studies that Collect Genetic Data

The National Institute on Aging
August 2-3, 2006
Bethesda, MD

Background

- Developing genetic and genomics within within BSR portfolios
- DNA already collected or plans for collection
- Important to develop efficient and feasible data sharing plans
- Procedures surrounding sharing of DNA samples are new to BSR many researchers

How to Proceed?

- Can we use already established plans connected to existing biobank projects and initiatives?
- Does data sharing within behavioral and social sciences raise new or unique issues due to the nature of the data?
 - Does the collection of DNA in BSR studies that have a large array of psychosocial & health data and links to administrative data pose new issues regarding deductive disclosure, privacy and confidentiality?
 - How do we determine which data can be safely shared?
 - How will genetic data sharing affect participation and sample attrition?

Data Sharing in the Post-Genomic Era

- Quick release of sequence data
- NIH data sharing requirement (>500 k)
 - Some NIH institutes and programs develop specific policies
 - NIH workgroups (trans-NIH and trans-HHS) to analyze needs, protect human subjects and coordinate policies to facilitate sharing of data and specimens
- Harmonization of biorepository procedures
- Moving target (scientific, analytic, legislative)

Purpose of Workshop

- Explore issues surrounding data sharing plans and policies for BSR studies that collect DNA
- Provide a forum for investigator input on these issues
- Develop preliminary guidelines for data sharing plan

Preliminary Work

- Teleconference with BSR PIs:
 - discuss the relevance of this effort for their research
 - Identify special areas of concern
 - identify specific information to be collected from ongoing regarding their data sharing procedures

Studies Interviewed

- Conducted interviews and compiled information from 6 major studies:
 - Add Health (*Dr. K Mullen Harris*)
 - SWAN (*Dr. MF Sowers*)
 - Framingham heart study (*Dr. A Cupples*)
 - Health ABC (*Dr. T Harris*)
 - NHANES (*Dr. G McQuillan*)
 - NIA Alzheimer's initiative (*Dr. Morrison-Bogorad, Dr. T Phelps & Dr. M Miller*)

Information Collected: Elements of Data Sharing Plans

- 1) Description of biological samples and data that are shared
- 2) Sensitivity of and access to phenotypic and genotypic data
- 3) Safeguards to ensure confidentiality
- 4) Sharing procedures
- 5) Informed consent
- 6) IRB considerations

Biological Samples and Data Collected

(Table 1)

Biological Samples Collected

- Blood (DNA)/Saliva/Urine – MIDUS, Add Health
- Blood (DBS)/Buccal Swab – HRS
- Blood (DNA)/Saliva – HABC
- Blood (DNA)/Urine – SWAN, NHANES
- Blood (DNA)/autopsy tissue – NIA Alzheimer's
- Blood (DNA) – Framingham

DNA Samples

- DNA amplification/cell lines
 - Add Health
 - Framingham
 - NHANES
 - SWAN
- Available DNA – could be amplified
 - HABC
 - MacArthur
 - HRS

Other Samples

- Limited supplies
 - Blood – DBS, plasma, serum
 - Urine
 - Saliva
- Allocation guidelines/priorities
 - Review of applications – written protocols (e.g., SWAN, HABC, Add Health)
 - Reviewers (internal/external)

Commonly Available Data

- Socio-Demographics – e.g., age, sex, ethnicity, education, income, occupation (hx), household/family structure
- Behavioral/Lifestyle – e.g., physical activity, smoking, alcohol consumption
- Physical Health – e.g., self-rated health, self-reported chronic conditions
- Psychosocial (less standardized) – e.g., social ties/support, depression, anxiety

Other Available Data

	Physiol	Genetic	Mental Health
Add Health	X	X	X
Framingham	X	X	?
Health ABC	X	X	X
HRS	2006	APOE	X
NHANES	X	X	X
MIDUS	X	In Future	X
SWAN	X	Encrypted	?

Specialized/Restricted Data

	Geog. area	Other admin data	Admin (w/in study)
Add Health	X	School	Encrypted
Framingham		?	X
HABC		?	Encrypted
HRS		Soc Sec	?
NHANES	X	NDI, CMS, Soc Sec	Encrypted
MIDUS		NDI	?
SWAN		?	X

Sensitivity of and Access to Phenotypic and Genetic Data

Table 2

Sensitivity

- Two studies with categorized levels of sensitivity (AH, NHANES)
- Methods for deciding what is sensitive varies (e.g. statistical disclosure analysis, variable content & intuition)
- Access and procedures to obtain data vary by level of sensitivity

Data Access by Level of Sensitivity

- Non-sensitive & public (no application required) – HRS, NHANES (general data), Taiwan, Add Health (general, 50% subsample), MIDUS (interview only at present)
- Sensitive & Restricted (application required) – NHANES (geog/mortality); HRS (geog linkage; pension/SSN?); Add Health (“partners”, geographic)
- No sensitivity level-application required for any data access – Framingham (NHLBI/Framingham), HABC, SWAN

Safeguards to Ensure Confidentiality

Table 3

Safeguards to Ensure Confidentiality

- Contractual agreements for data handling, storage, access and analysis
- Important distinction between restrictions imposed by security plans versus restrictions on which data can be used
- Generally don't share variables that carry high disclosure risk
- Certain types of data more difficult to de-identify (e.g. bone scans)
- Must pass committee reviews
- Examples of other safeguards:
 - Encryption and password protection
 - An off-site broker with key
 - Limitations on publishable sample size

Procedures to Prevent Deductive Disclosure?

- Affected by study design
- Restricted data but no standard procedures for decisions regarding:
 - Sensitive single variables or specific combinations of variables
 - Trimming/collapsing of extreme/rare data values
 - Restrictions on available sample (e.g. 50% only Add Health)
- Restricted access
- Standards for data security when sharing data (e.g., distribution and confidentiality agreements)

Sharing Procedures for Genetic Samples and Phenotypic Data

Table 4

Procedures for Sharing Samples

- Application process
 - internal review – HABC, MacArthur
 - internal/external (both) - SWAN, Add Health, Framingham, NIA Alz.
- Material transfer agreements
- Data security plan
- Data use agreement
- Confidentiality agreement
- Local IRB approval

Sharing Samples

- Return of results/samples (e.g., after 3 papers, after “x” period of time, contingent upon receipt of genotypes) – specified in Data Use/Material Transfer agreements
- Commercial use (but no proprietary use)
 - Yes (NIA Alz)
 - In principle – Add Health, SWAN
 - No – Framingham, NHANES

Duplication of effort

- Particular concern given high-throughput genotyping
 - Database of published works and funded project titles/abstracts
 - Oversight of research projects
 - Duplication is discouraged during application process
 - Require list of desired SNPs
 - Negotiation towards collaboration when proposals overlap

Issues Related to Scope of Genetic Analyses?

- Problems arising with overlap when multiple researchers want to analyze thousands of SNPs
 - Analyses are limited to initial number of SNPs in approved proposal
 - Attempts to separate phenotypes even if genotypes overlap
 - Collaborations encouraged
- Researchers conduct further genotyping beyond original proposal
- Need a mechanism whereby SNP panels are immediately available
- WGA: researchers have enormous amounts of data
- How do researchers follow up on 'hot' topics?

Problematic Areas Identified with Data Sharing Plans

- Resource demands to establish and maintain sharing plans
- Security issues require expertise and resources
- Possibility to conduct new analyses on borrowed samples are restricted by contract
- No proprietary period for use of samples or data
- Problems in getting data and reports back from researchers
- New issues emerge with advances in genetic analyses

Repositories

- Models for sample management
 - Study Investigator
 - Established independent repository
- Backup locations
 - Framingham
 - NIA Alzheimer's
 - HABC (McKesson mostly; U Vermont – genetic material)

Resource Demands for Data/ Sample Sharing Plans

- Expertise and review panels
- Funded staff (process applications, prepare data sets, answer questions, pull samples etc)
 - Add Health (3 FTE/P01)
 - Framingham (2+/NHLBI)
 - Health ABC (2 FTE at Coordinating Center/NIA contract)
 - SWAN (contracted commercial firm)
- Fees to user's – yes

Informed Consent

Table 5

Informed Consent

- Structure of Consenting for biological samples
 - *Single consent* (Health ABC; NHANES, SWAN [DNA/other biological protocols])
 - *Segmented consent* (e.g., MIDUS, Framingham)
 - *Separate consents* (e.g. HRS; Add Health)
- Certificate of Confidentiality - HRS, ADD Health, SWAN, MIDUS

Informed Consent

- Future removal of stored specimens
 - Possible – SWAN, HRS, NHANES, Framingham
 - Not possible – Add Health
 - Not specified – MIDUS, HABC

Informed Consent

- Initial Samples - Feedback to participants?
 - NHANES – results to participant and MD
 - MIDUS/MacArthur - clinically informative only (e.g., BP, BMI, total & HDL cholesterol, glycosylated hemoglobin)
 - HRS - glucose & cholesterol
 - Add Health - HIV/STD's
 - nothing at Wave IV
 - HABC, SWAN - no results

Informed Consent

- Stored Samples – results to participants?
 - Generally NO – e.g., NHANES, MacArthur
 - SWAN (DNA) – call if interested

Informed Consent

- Future uses of stored samples
 - SWAN – “you will not be able to direct use”
 - HRS/MIDUS/NHANES – general statement of “health-related”
 - Framingham – list w/ broad scope (e.g., heart and blood disease, cancer, memory, bone, other diseases and health conditions)
 - HABC – “related to body composition, disability, weight-related conditions” (will contact if use of other than this)

Informed Consent

- Stipulations regarding possible subsequent sharing of data outside original research group
 - SWAN – outline application and oversight
 - NIA Alzheimer’s-explicitly mentions sharing with other researchers studying Alzheimer’s disease
 - No specification – MIDUS, HABC, HRS, Add Health

Informed Consent

- Feedback/concerns from participants
 - Generally no negative impact of biological/genetic consenting
 - Some additional concerns regarding genetic components
 - Intimidated by length and language of current consents
- No standard “lay language” explanations
- Who does consenting? What kind of training to do staff have?

IRB considerations

Table 6

IRB Considerations

- Access to shared data
 - No redistribution – restricted to the investigator/institution explicitly referenced in “data use/material transfer agreements”
 - Recipient responsible for ensuring adequate security and restricted access at local institution
- Procedures for Inter-institutional transfer
 - Generally not allowed without new agreement/contract with new institution (e.g., UCLA CARDIA “Standard Operating Procedures” – streamline IRB approval)

IRB Considerations

- Challenges
 - Variations in IRB standards/requirements across institutions and over time – lack of common rules/guidelines (prior study’s experience may not “work” in other settings)
 - Frequent need to “educate” IRB regarding specifics of your data and the implications of this for:
 - Consent process – e.g., who does this (MESA/MD vs MIDUS/staff)?
 - Sharing results with participants
 - Sharing data and/or samples with other investigators

BSR Data Sharing Workshop

Day 2

Recap: Elements of Data Sharing Plans

- 1) Description of biological samples and data that are shared
- 2) *Sensitivity of and access to phenotypic and genotypic data*
- 3) *Safeguards to ensure confidentiality*
- 4) *Sharing procedures*
- 5) Informed consent
- 6) IRB considerations

Additional Questions

- Concerns not yet addressed?
 - Are there characteristics of “your” studies that pose other data sharing concerns/issues?
- Shortfalls in models considered?
- What is unique to data sharing in context of likely BSR-funded studies (merging of social science & genetic/biological data) – different from six studies reviewed?

Critical Topics for Discussion

- Data Access (esp. issues of sensitivity, deductive disclosure)
 - Tiered System – best option?
 - Accommodating different levels of data sensitivity
 - Reduce/alleviate concerns about “deductive disclosure” by restricted and supervised data access?
 - Flexibility – need to accommodate differences in:
 - study design (e.g., Add Health vs. national samples [HRS/MIDUS]),
 - data elements – individual variables and special combinations (e.g., via linkages)

Critical Discussion Topics

- Assigning “sensitivity-level” of variables – who should be involved (e.g., study investigators, outside ‘experts’)?
- Data/Sample Sharing system(s)
 - Centralized and/or Distributed
 - one or more “BSR-supported data sharing centers” providing services and expertise vs. individual studies?
 - what services best centralized (all/some)?
 - Accessing “restricted” data
 - Licensing?
 - Remote Data Centers?
 - Data enclaves?



Identification Disclosure and Social Science Research

Jerry Reiter
Institute of Statistics and Decision Sciences
Duke University, Durham NC, USA
jerry@stat.duke.edu



Setting for problem

- Data producer collects data on individuals, including biomedical/genetic variables.
- Data producer seeks to share collected data after removing obvious identifiers.
- Data producer concerned about risk of deductive disclosures.



Types of disclosure risks

- *Identification disclosure*
Match record in released data with target.
- *Attribute disclosure*
Learn value of sensitive variable for target.
- *Perceived identification disclosure*
Match record in released data with incorrect target.
- *Inferential disclosure*
Closely estimate value of sensitive variable for target.



Identification disclosure

- For particular target, snooper knows and matches on values of key variables that are in shared data, such as:

geography,
race, sex, marital status, age,
occupation, housing, income, family,
medical/genetic profiles



It may be riskier than you think...

- Sweeney (2000) estimates that 87% of U.S. population is uniquely identified by combination of gender, birth date, and 5-digit zip code.


Linked MA voting records to insurance data containing these variables to learn health data for MA governor.



And riskier still...

- Web sites like Choice Point allow purchase of massive data sets containing:

geography, race, sex, marital status, age, housing data, employment histories, lifestyle choices, ...




Measuring identification disclosure risk

- Attempt to determine if record is unique in population on available keys.

(Duplicates also at risk.)


- Attempt to match released records to other databases.

(Match to existing data set if snooper knows target is in released data.)



Additional concerns for biomedical/genetic data

- Geographic information may be known even when not released (e.g., Framingham study).
- Knowledge of who is in the data may be widespread.
- Genetic information could be identifying when known by others.
- Linked data: safe in one data set may not be safe after linkage.
- Potentially serious harm from disclosures of biomedical/genetic information.



Data quality: The other part of the story

- Usefulness of sharing data measured by benefit to society of dissemination.
- Attainable: compare analyses from released and original data.
- Altering/restricting data to protect confidentiality worsens data quality.



Potential solutions: Tiered access

- General public use:
Data altered before broad dissemination.
- Licensing to researchers:
More detail in exchange for guarantees.
- Remote access systems:
Computer provides results of analyses without revealing individuals' data.
- Research data center:
Research done in secure data enclave.



Broad dissemination strategies

- Not release geographic identifiers below certain geography (e.g., 250,000).
- Coarsen ages or outlying demographic variables (e.g., income).
- Swap small amounts of data between records (e.g., swap sexes).
- Add noise to continuous values.
- Simulate values at risk (synthetic data).



Licensing

- Used by some agencies, such as NCES, BLS, and NSF.
- Violations mostly not following protocols, but no evidence of confidentiality breaches.
- Strong penalties in some cases (e.g., investigator and institution forfeit funding from agency sponsoring grant).
- Enforcement mechanisms seen as weak.



Remote access

- Used by some agencies, such as NCHS and Census Bureau.
- Allows analyses to be done on (nearly) actual data.
- Mostly for tables rather than microdata, but this is changing.
- Techniques from secure computation.
- Disclosure risks from judicious queries.




Research data centers

- Allows access to (nearly) actual data.
- Challenging to establish and maintain, especially for individual researchers.
- Inconvenient for researchers who do not live near a center.



Concluding remarks

- Need to decide whether identification disclosures are not problematic (maybe only attribute disclosures are harmful?)
- Data for broad dissemination can be coarsened/simulated, but some analyses are distorted or impossible.
- Data licensing with strong penalties offers possibilities for access to detailed data.



Additional concerns for biomedical/genetic data

- Geographic information may be known even when not released (e.g., Framingham study).
- Knowledge of who is in the data may be widespread.
- Genetic information is identifying when available to other researchers.
- Linked data: safe in one dataset may not be safe after linkage.
- Potentially serious harm from disclosures of biomedical/genetic information.

“Repository” Configurations: Central vs. Distributed

**MaryFran Sowers, Ph.D.
University of Michigan**



Charge

To present the advantages and disadvantages of models that use a centralized versus distributed repository as it relates to issues of control and other aspects of data sharing (e.g. confidentiality, costs and resources needed, ease of sharing, return of results to the repository archive etc).



Primary Repository functions: Archive **plus** data sharing



traditional Repository activity is/was primarily archival

- “cold refrigerator” or “freezer” mentality for specimens
- Two general types of configurations:
 - Commercial firms or informal govt organizations that would provide unique specimens with selective documentation about the specimen
 - Study-specific specimen storage under study control

Primary Repository functions: Archive **plus** data sharing



New permutation:

- universal data sharing
- Shift in time frames for availability
 - Immediately
 - Concurrent with ongoing data collection
- Associated with unprecedented data scope
 - Longitudinal data
 - Health, social, behavioral, family-linked data
- Associated with rapidly developing information systems
 - Critically important, costly, complex, evolving

Primary Repository functions: Archive plus data sharing



- **Consent, data and specimen collection**
- **Specimen receipt and quality checking, (may include processing steps), assignment to storage, and establishment of procedures for inventory**
- **Administrative procedures for appropriate review and implementation of release of specimens and data to users**
- **Mechanism to monitoring use, assure return of specimens and/or data, reutilization of data and/or specimens**

Archival and data sharing activities require vastly different skill sets



- **Data and specimen collection (consent)**
- **Demands of data and specimen processing (representative data samples, cell immortalization)**
- **Data and specimen storage and processing – with a particular focus on IT**
- **Review and use processes and management**

Complexity when Repository is afterthought



- Study originators have no experience with archive + data sharing
- Repository/data sharing is not study priority (scientific question orientation vs. service orientation to provide resources to other investigators)
- Repository functions and the data sharing elements are grafted on an existing structure and/or organization.
- Those who mandated configuration frequently have less (or no) experience than would be desirable.
- Substantial negotiations

Configurations can be centrally located or distributed



Central Hybrid Distributed

Collection
Archive
Processing
Utilization

SWAN Repository configuration



	Central	Hybrid	Distributed
Collection			X
Archive	X		
Processing		X	
Utilization		X	

Central Configuration: Strengths



- Central accountability: Program officer knows who to talk to
- Organizational relationship has been defined and implemented to allow for accountability for provisions of services
- Has resources and can enlist specialized services as needed
- Effective IT system is implemented that integrates all activities with effective relational data base
- Can clearly define goals and objectives or is entrusted with mechanism to redirect goals and objectives
- Efficient
 - developing single protocol
 - central training
 - Developing appropriate IT system

Central Configuration: Limitations



- Power may not travel with responsibility
- Will not work in settings that have an existing structure that must be accommodated
- Centralized organizations may not engender the trust or rapport (or actual protection of confidentiality) that distributed organizations may be able to develop.
- IRB's are not centrally located, they are distributed
- US academic (commercial?) system(s) through which material transfer agreements are implemented are distributed
- "cooperating" units make decisions with direct bearing on the Repository, but fail to collaborate or even consult with Repository

Distributed Configuration: Strengths



- Adapts to the realities of multiple systems through which material transfer agreements and IRBs are administered in US
- Have flexibility to bring selective resources to bear if central resources/organizational structure are absent
- Works best if resources are available as incentives to collaboration

Distributed configuration: Limitations



- **Frequently cumbersome and inefficient**
- **More costly**
 - Multiple levels of administrative costs
 - Accommodate costs of non-standardized approaches
- **Uneven levels of technical proficiency across all levels of collaboration**
- **Much more complex managerial job**
 - Need the skills of a union negotiator, the patience of Job, the perseverance of a rat terrier, a very thick skin, and self-motivation to do a thankless job

Summary



- **Need to have a more clear understanding of the “new” paradigm of Repository archival and data sharing and implications for the nature of funding, management, and collaboration.**
- **Understanding strengths and limitations of organizational configuration is absolutely critical at level of funding agency, managerial implementation and collaborators.**
- **Realities of IRB, intellectual property, material transfer agreement, and existing “parent” study configurations frequently mitigate against less costly, more efficient configurations**