

Data Sharing Strategies for Environmental Health Science

Executive Summary

As an initial step in further exploring data sharing efforts for environmental health science researchers, NIEHS put a Request for Information (RFI) into the NIH Guide in the summer of 2011 entitled “Input on Strategies to Encourage Broad Data Sharing in Environmental Health Sciences Research.” Researchers and other community stakeholders in the environmental health sciences provided suggestions and concerns regarding potential approaches and strategies that would allow broad data sharing in environmental health human population studies. Based on this input, NIEHS convened a workshop in February 2012 to explore some of the issues and challenges related to sharing environmental health data. This report highlights the meeting topics and speaker presentations and summarizes key recommendations arising from both the RFI and the workshop.

The first session of the workshop set the framework for considering data sharing guidelines for environmental exposure data in the context of the broader NIH data sharing guidelines and policies, including the GWAS data sharing policies, and included relevant legal aspects to consider with data sharing. Presentations in the second session addressed research participants’ perspectives related to the sharing of environmental exposure data. This included ethical concerns such as confidentiality and privacy issues, protections from discrimination, the lack of clarity in the IRB process and inconsistencies in IRB approvals in multi-site studies, as well as more positive perspectives on data sharing. Following the presentations, a panel discussion explored some of the challenges and obstacles, as well as opportunities and successful strategies, utilized in sharing environmental exposure data in a variety of human population studies. An additional session explored what should be the minimal data expectations to include in a data sharing plan to maximize data sharing and collaborations among researchers. Several examples highlighted data sharing “success stories” and included applications for best leveraging unique environmental exposure datasets. Efforts to establish environmental exposure measurement standards and common vocabularies that could facilitate cross-study comparisons were also presented. The final session sought consensus among meeting participants on specific recommendations for successful implementation of a data sharing strategy.

Session I: Opening Remarks and Introduction

Linda Birnbaum, Ph.D. and Gwen Collman, Ph.D. provided opening remarks on the purpose and goals of the workshop, which were to:

- clarify the scientific needs, importance, and goals for environmental data sharing,
- identify specific challenges related to the sharing of environmental health data,
- identify best practices and successful models of data sharing that are applicable to environmental health data sharing, and
- explore the technological considerations for harmonizing phenotypic data and merging diverse datasets.

Both Birnbaum and Collman noted that NIEHS has funded some valuable and unique studies of exposed populations, and it is increasingly important to leverage these existing data resources in the context of ongoing budget constraints. They also noted that performing secondary data analysis and meta-analysis from previous investments is strongly encouraged at NIH. In addition, Collman mentioned some of the unique aspects and challenges related to data sharing in the environmental health sciences field that need to be addressed, including:

- heterogeneity of environmental and biological measurements,
- potential to identify individuals based on the association of environmental exposures with geographical data,
- increased interest on the part of research participants in return of individual or community-level research results from environmental research,
- regulatory implications of the use of environmental exposure and health data in developing national research policies, and
- unique concerns of vulnerable populations who are disproportionately impacted by environmental exposures.

Kim McAllister, Ph.D. summarized the comments and suggestions from the RFI and presented the key questions that NIEHS wanted the workshop participants to consider:

- What can NIEHS do to facilitate data sharing efforts?
- Is there a need to identify “best practices” for sharing environmental health data?
- If NIEHS adopts data sharing guidelines, what are the minimal data requirements that should be included?

In the keynote presentation, Bruce Lanphear, Ph.D. pointed out that many researchers rely on data collected by someone else and that many published papers in the environmental health sciences field rely on national survey data, shared datasets, or pooled data. Data sharing can maximize public investment and public benefit. However, there are costs and issues associated with sharing data and pooled analysis as well as restrictions on how some data can be shared, and oversight for appropriate data usage is necessary.

Session II: Legal and Policy Considerations of Data Sharing

Data Sharing: Taking Research Further – J.P. Kim

Many NIH policies support data sharing efforts to enable full exploration of important research topics and leverage existing investments. NIH policies and guidance are available online (<http://grants.nih.gov/grants/sharing.htm>). Key NIH policies related to data sharing include:

- expectation of a data sharing plan for applications of \$500,000 or more in direct costs,
- NIH public access policy for publications from NIH-funded research, which requires peer-reviewed published papers funded by NIH to be submitted to PubMed Central,
- model organism sharing policy statement, and
- GWAS data sharing policy.

Kim noted the NIH website will be adding more information about NIH sharing plans and policies, as well as education about the benefits of data sharing and technical assistance for accessing NIH datasets. More specific policies may be needed for data sharing of environmental exposure datasets, but the existing NIH data sharing policies establish general guidelines and serve as a starting point for NIEHS researchers.

Sharing Genomic Data: NIH Policies Past, Present, and Future – Laura Lyman Rodriguez

Rodriguez' presentation focused on genomic data sharing. She noted genomics research has a strong culture of rapid and broad data release. Due to the effort, cost, and the frequent consortium nature of genome-wide association studies, NIH decided GWAS data was a "community resource" that should be founded on the principle of no-cost and rapid release of data for use by investigators throughout the global scientific community. The GWAS data sharing plan may help provide context for other types of data sharing, including environmental exposure data. The trans-NIH GWAS policy allows rapid release of data within the context of expressed data use limitations based on informed consent of individuals whose data is being shared, IRB restrictions, and legitimate concerns from the investigators. Genomic data from NIH funded GWAS studies is submitted to the Database for Genotypes and Phenotypes (dbGaP) after an IRB has reviewed the submission plans from the investigators to ensure that data use limitations are consistent with existing informed consents and other restrictions. The investigator is required to remove Health Insurance Portability and Accountability Act of 1996 (HIPAA) identifiers before deposition and retain the keycode to the data prior to data submission. NIH Data Access Committees (DACs) review all requests from the research community for access to dbGaP data to ensure the proposed research use is consistent with any data use limitations for the dataset. A new trans-NIH genomic data sharing policy is being developed based on rapid advancements in the field of genomics that will also include whole genome sequencing data and other types of genomic data.

Session III: Ethical Concerns

Ripple Effects of Data Sharing: Ethical Concerns - Richard Sharp

In the shift from individual research to collaborative research projects, there is a greater expectation of data sharing, particularly among patient advocacy groups who desire greater impact on disease outcomes. But, little research has been done to examine public views of data sharing. When considering participant perspectives, the viewpoint of most patients about their data being made available to more researchers is unknown. Bioethics research may help with understanding the adequacy of informed consent when the secondary uses of data (including use of stored biological samples) are still undetermined. More research is also needed to understand participants' perspectives on potential future risks and harms with sharing their data in different contexts. Preliminary research suggests participants may be open to more uses of their data and biologic materials than researchers and NIH previously believed. These studies suggest that most people are mainly concerned about extensive follow-up or have privacy concerns. When research participants have a trust relationship with the clinicians/researchers and institutions, it is more likely that they will give consent for secondary uses of data. However, consent for unknown end users might be problematic, and the process of data "anonymization" makes it difficult to go back to patients for secondary analysis permission. Involving more community members or advocates who speak on behalf of patients in the development of data sharing policies and complete transparency with research participants will encourage the public to participate in research and allow broad use of their data.

Examples from the Field - Julia Brody

Shared data sets with demographics like date of birth, gender, and zip code can be linked to public registries, such as a voter list, to re-identify people by name and gain sufficient information to contact the subject of the data. Some examples of what makes environmental data identifiable include: information in published findings (e.g., locations where PCBs were found), observable data (e.g., house has wood stove), linkable data (e.g., a building permit or purchase data from stores), and stakeholder knowledge (e.g., information about spouses or employers). Potential harms from individuals being linked to exposure data may include: insurance premium increases, employment status, property value fluctuation, illegal behaviors that could be revealed, and liability triggers for litigation based on reporting of or remediation for a regulated substance. Technological solutions for these issues may include computer tools to predict the identifiability of environmental data and tests of data-masking and auditing.

Researchers are encouraged to share personal exposure data with study participants (report back) because individual participants want their results and frequent interactions also increases the trust between the researcher and participants and increases pride in having participated and contributed to science. However, participants' frustration may grow as information is gained about

exposures that affect them (toxic trespass). Researchers and IRBs would benefit from additional guidelines related to the development of streamlined and standardized informed consents, data-user agreements, and individual report-back responsibilities as well as new legal protections to prevent forced disclosure or undue liability.

Data Sharing in Broader Context – Wael Al-Delaimy

Data sharing presents many ethical risk-benefit balances that need to be considered. Environmental justice issues in communities and specific community concerns may yield additional sensitivities. For example, a stigma (low-income and higher air pollution) can be attached to an entire geographic area involved in an environmental study or indigenous populations may have special restrictions they want placed on their data. As a recent example of the potential for stigma, the Havasupai Indians consented to genetic analyses related to diabetes in their community but were later outraged when their data were also used to explore conditions related to mental health, an area for which they had not consented. Al-Delaimy recommended that a community IRB or IRB subcommittee review ethical aspects in community studies. This subcommittee could facilitate local IRB approval, address conflicts of interest, help synchronize community interests and concerns with the original data design and consent process, provide a risk stratification checklist, and provide guidance for research ethics training.

Session IV: Environmental Health Sciences Data Sharing Strategies - Panel Discussion

A diverse panel shared their successes and frustrations/concerns with sharing exposure data. Some lessons learned included:

- ❖ A consortium model that may work well is one in which each researcher keeps his/her own data and runs their own analysis based on their local cohort. A meta-analysis could then be performed for multiple studies in the consortium to address particular scientific questions. Pooling raw data (GWAS or other) can be much less efficient because of the difficulty in understanding other investigators' datasets. In addition, the time-consuming nature of cleaning data and making it consistent across studies should not be underestimated.
- ❖ Sending data or analysis code to data owners and getting them to do the actual analysis while sending only the results back to the researchers may work when health departments cannot or are reluctant to release raw data.
- ❖ Studies with government-managed cohorts are constantly under pressure to release data; however, data about exposures may have economic implications and potentially harm communities. Findings may be interpreted in misleading ways if data are released too soon or analyses are incomplete.
- ❖ A restrictive protective order in court to guard research work against litigation may be an option if there are many FOIA requests or if subpoenas for data disclosure become extremely cumbersome.

- ❖ Separation of data collection and cleaning from data analysis is desirable to allow unbiased secondary analysis of complicated datasets.
- ❖ IRB discrepancies across institutions should be clarified; there is a need to educate IRBs about feasibility of simpler consent forms and data sharing plans and NIH should take the lead in this effort.
- ❖ Data sharing requires oversight, limits, and guidelines; FOIA requests have been abused and a data-sharing plan that doesn't consider FOIA may undermine the researchers' efforts. A mechanism for an independent reanalysis could reduce the burden of FOIA requests.
- ❖ Data sharing plans should address: data cleaning and coding clarification, uniform sample storage, consent forms, strategies for sharing with different research groups (ensuring that investigators are qualified and have no financial conflicts of interest), and approach for report-back to participants.

Session V: Implementation of Data Sharing Strategies

Are You Ready for Data Sharing? Lessons Learned from the Fernald Community Cohort - Susan Pinney

Fernald community cohort data is shared by using an exposure metric for uranium (air, water, and organ doses), which precludes the need to distribute geocodes. Developing an exposure metric and sharing the metric only is one way to solve the problem of sharing identifiable data. The researchers involved in this cohort plan and prepare for data sharing by documenting and coding all data, using standardized derived variables for consistency in analysis, providing the software code, standardizing missing data rules, establishing a data dictionary that allows variables to be linked across years, and demonstrating that the cohort has statistical power to address potential research questions. Fernald researchers have a biospecimens sharing policy and a policy to sanction other researchers who are not compliant with approved data usage. They maintain a website to disseminate up-to-date information about the study.

The PhenX Toolkit: Make Data Sharing Easier - Carol M. Hamilton

The PhenX Toolkit contains standard measures for phenotypes and exposures to facilitate data sharing among researchers who incorporate PhenX measures into their studies. Inclusion of some standard PhenX measures, in addition to the specific measures needed to address the research question, will aid study compatibility for future studies. The Toolkit includes 15 measures or characteristics for each of 21 research domains (including cancer, physical activity, diabetes, demographics, anthropometrics, and environmental exposures) and additional collections of measures specific to substance abuse and addiction. All of the measures in the Toolkit were selected by working groups of domain experts, using a consensus process. Detailed protocols are provided so that investigators can consistently collect data associated with PhenX measures. This toolkit provides a common currency of measures that can facilitate validation studies and increase statistical power for cross-study analysis and meta-analysis. The

idea is that studies asking different but related research questions, when combined, may yield new (unexpected) insights into risk factors, disease susceptibility, and progression. Researchers can visit the Toolkit to add standard measures to ongoing studies, consider PhenX measures when planning new studies, and easily obtain high quality measures outside of their primary area of expertise. Some in-depth environmental epidemiology studies may need more specific measures to share in that research community. NIDA sponsored a project to develop PhenX measures for the substance abuse and addiction research community. This project added 44 new measures to the Toolkit; these measures provide depth and specificity for NIDA and NIAAA researchers. As substance abuse and addiction investigators begin to share and combine their data, the results of each individual study will have a higher impact. NIEHS could do something similar as one means for promoting broader environmental data sharing.

Sharing Data via Established Policies and Procedures in NHLBI Cardiovascular Cohort Studies: Lessons from MESA for Environmental Health Research - Joel Kaufman

MESA Air, a large federal investment with an obligation to appropriately share information with others, pairs state-of-the-art cardiovascular epidemiology with exposure estimation. Sharing is done in three different ways:

1. data sets, containing limited individual-level covariates, are published and available on a regular schedule,
2. data from investigator-directed scientific collaborations follow an ancillary study process, and
3. investigators deposit and share (access) genomic data in dbGaP or another genomic repository.

Genomics data must be uploaded according to data use agreements and an approved proposal is required for data access and use.

MESA Air encourages extensive collaborations and ancillary studies. Exposure data needs to be well documented and quality assured so that end-users can use it appropriately. One particular issue in this study is that the precise geocoding can identify participants and must be protected like other unique identifying information. MESA Air cannot release participants' geographic data to reduce individual linkage options. Secondary researchers must be satisfied with exposure estimates done by primary researchers and variables of estimates of exposure to particular toxicants or they must go through the IRB process for permission to obtain release of the geocode information.

Data Sharing in an Integrated Health Delivery System - Stephen Van Den Eeden

Kaiser is a strong integrated health system whose resources can be used for a wide range of environmental research needs and questions. It has more than 8 million participants (representing 20% of the U.S. population), each with an

electronic medical record. Some examples of environmental health study collaborations with Kaiser include:

- BEST Study, which concerns bio-monitoring in the Central Valley of California,
- Autism Portfolio, a virtual registry with each HMO's local data standardized to run queries to get the same type of data from each center, and
- Research Program on Genes, Environment, and Health (RPGEH), a gene and environment study linking electronic medical records with a GIS database with many environmental variables.

An access committee reviews and prioritizes structured applications from external scientists for use of their resources, which restricts the time and use of the data and requires submission of any "new" data generated to be shared with others as well. Collaborations with an internal Kaiser investigator are encouraged so the clinical data can be utilized as appropriately.

Models for 'Sharing' Research Data: A Data Coordinating Center Perspective - Howard Andrews

To create guidelines for data sharing, consider the specific types of data sharing, which can be a pre-planned multi-site study (which agrees up front how to collect and measure data), a post-hoc pooled data study (which often finds that elements are measured in different ways and requires a meta-analysis approach), or a study that uses common data elements before data collection begins (like the PhenX concept). The type of data (aggregate, subject level "processed" data, or original "raw" data) and the type of sharing (through investigators, central repository, or public access) can also define the data sharing considerations. Some of the biggest issues yet to be addressed for wide-spread data sharing are "identifiability" problems, difficulty in finding common/standard data elements that work for many researchers, and inherent challenges with pooling or "harmonizing" data.

Some advantages of a central processing and analytic center are minimization of confidentiality concerns when only statistical results are shared, expertise in working with raw data is built, consistent quality control procedures are established, sophisticated statistical designs may be created, and documentation and syntax for derived variables can be managed and modified as needed. In general, oversight provided through a center can enable analyses to be more feasible, efficient, and quality controlled than when analyses is performed by a widely distributed group of independent recipients of shared individual-level data.

Developing a Framework for the Institute: Moving Forward: Final Recommendations

The final comments, suggestions, and recommendations of the researchers and other community stakeholders in environmental health sciences that

responded to the RFI or attended the workshop are summarized below under several broad themes or categories.

Unique Considerations with Exposure Data:

Because environmental health science studies have a strong component of community involvement, communities may be more motivated to participate in environmental health research than other types of research. Conducting community-based participatory research with thorough community consultation and consent for all use and disclosure of data is of utmost importance in establishing trust with the community. Trust is particularly important in vulnerable communities where environmental health research related to multiple environmental exposures is likely to occur. The specific needs and stated preferences of individuals and communities should be incorporated into researchers' data sharing plans. There is an ethical imperative for making aggregated data broadly accessible while also protecting confidentiality. Furthermore, researchers should report back data to communities and disadvantaged groups to the extent possible as many study participants strongly desire outcome information.

Environmental data are also unique because the focus is on exposure more than disease outcome. In some community studies, people may not necessarily have a disease but may have a heightened risk that is under investigation due to an exposure. Confidentiality issues may be different between those with disease and those who are not yet affected. Exposures are in many cases geographically based and researchers should develop approaches to address specific needs for de-identification of geographically based data. Environmental exposure data with GPS information in particular allows specific identification of the sources of exposure and can, and has been, used against communities to discriminate (e.g., reporting lead paint exposures by specific locations to departments of health).

Several successful options for sharing data across environmental projects and databases without identifying subjects' personal information were suggested. Options include using a "unique subject identifier" or archiving data to a separate external repository with the identifying data stored separately and requiring special permission to access. Existing security measures to protect participant/patient confidentiality are often inadequate for online databases. The security of analysis and data platforms, transmission procedures, the role of firewalls, and training of data gatekeepers need to be addressed better to safely allow broader data sharing in environmental health sciences and many other fields.

IRB and Consent Issues:

The lack of continuity, consistency, and clarity across IRBs for issues related to participant consent and sharing of data was emphasized by many research groups in both the RFI and at the workshop as a potential disincentive for attempting to more broadly share their data with others. IRB issues were cited as an impediment to the development of informed consent models that might allow

data results to be compared, pooled, or analyzed more broadly among research teams. Several investigators pointed out that multiple IRBs may be involved for data sharing to occur, and IRBs may not accept each other's decisions regarding which data could be shared and how. Although some restrictions on data sharing (particularly for medical/clinical data) exist, many feel that NIH guidance for IRBs and researchers about data sharing and templates for easy-to-understand consent forms would go a long way to alleviating this issue. Additionally, NIH could provide education and training for IRBs and researchers about how environmental exposures in communities are typically assessed and measured that might improve IRB understanding of the risks and benefits of such studies. One recommendation was that the intent to share data and the approach for doing so should be specified to the IRB from the beginning of the study to avoid unnecessary delays later.

Unique IRB considerations for research studies with tribal populations also arise as many tribes have their own IRBs in addition to an academic IRB review. Tribally affiliated IRBs guard against potential adverse impacts to tribal individuals or governments that may be overlooked by academic IRBs and help ensure data collected by researchers on behalf of the tribe remain tribal property. In addition, informed consent may need to be obtained at both the tribal government and individual level for some tribal organizations.

Researchers agreed that consent forms need to be designed specifically for each individual study that incorporates study participants' and researchers' needs and desires. Language on participant expectations for return of research results, scientific publications, and other forms of dissemination should also be encouraged and incorporated into the informed consent process. NIH could provide general guidance on restructuring consent forms, including boilerplate language about data sharing. Recommended language related to data sharing for consent forms would minimize the need to go back to subjects for further consent. Some model consent forms include those developed for the U.K. Biobank, the Nurses' Health Study, and the U.S. National Children's Study. New models for consents (e.g., digital signatures or videos for children) can help alleviate the burden for both researchers and participants. More open-ended consents for biospecimen sharing may be possible since privacy risks to individuals may be lower in these situations.

Legal and Regulatory Considerations:

A unique parameter of environmental health science research is related to "toxic tort." Exposure data will continue to be of great interest to regulatory agencies with respect to the evaluation of the health implications of chemicals. The reanalysis and/or reinterpretation of environmental health science data in an effort to delay regulatory reform or influence court cases and the general public will always be a cause of concern for environmental health science researchers. Researchers should always anticipate regulatory or industry concerns regarding

the data that will be collected. Utilizing independent analysis may help to prevent some of the delays related to secondary analysis by special interest groups.

Special conditions might also be needed for industry or private users of data given the potential for discrimination. Conditions include discrimination based on genetic information sharing, the potential for negative economic impacts on property values, and employment issues related to exposure disclosure. Because taking data out of context is a concern, researchers should carefully define uses of data and ensure that confidentiality agreements are in place for any data sharing. In addition, complete financial disclosures for all data users could be published online since conflicts of interest apply in all areas of data sharing (voluntary, involuntary, and voluntary non-collaborative). The fact that no legal protections are in place for use of environmental data with respect to decisions on personal health insurance or employment, unlike the legal protections for genomic data under the Genetic Information Nondiscrimination Act of 2008 (GINA) legislation, is a unique concern when considering broadly sharing datasets containing information on unique exposures. The question of whether a comparable law is necessary for environmental data was asked.

Caution should be exercised when using public health surveillance data because it is not the same as environmental health research data. Public health surveillance data are highly regulated and not easily linked with other types of data. Researchers need community consent to do exposure studies dealing with data under state regulations. Federal data sharing should therefore consider any unique state requirements about public health surveillance data as part of data sharing plans.

NIH Programmatic and Logistical Considerations:

Many investigators weighed in on possible ways that NIH might stimulate data sharing possibilities in the environmental health sciences communities. Key recommendations include:

- ❖ **Costs for Data Sharing in Research Plans**
 - Mandates to share data should be appropriately funded regardless of the mechanism.
 - The data sharing costs should be built in during the design of any study or grant request, not just those studies over \$500k or above.
 - Data sharing plans should include costs associated with report-back to participants and costs associated with the development of data variable dictionaries.
 - NIEHS might use administrative supplements to fund data sharing in existing studies.
 - NIH should consider supporting the sustainability of data and projects by funding institutional sharing of data after the project is no longer actively funded.

- ❖ NIH guidance and direction (to the extent possible) on obtaining financial disclosures from secondary data users and carefully defining data uses to minimize legal issues would be useful.
- ❖ NIEHS should facilitate more research on participant perspectives regarding environmental health data sharing since this area has not been thoroughly explored.
- ❖ NIEHS should identify and promote existing data sharing models and resources for data sharing management and storage. For researchers who are already funded, NIEHS could look at who is already sharing data and provide guidance and models based on best practices. Dryad is one potential data-sharing model that uses a repository (<http://datadryad.org>) for storage of data linked to particular publications and where “anonymized” data is freely accessible to other researchers. Membership to Dryad is open to funding agencies, and this repository could easily archive and make publicly available a wide variety of environmental health data.
- ❖ Because there is a lack of awareness of existing environmental health datasets, NIEHS could provide funding for increased use of existing datasets like NHANES and other relevant cohorts, and fund more re-analysis of extant data through various mechanisms.
- ❖ NIH may need to develop specific guidelines for data sharing for collaborations involving multiple foreign institutions. International collaborations can be incredibly complicated and a “one size fits all” recommendation or guideline from NIH regarding data sharing could be counter-productive. Researchers cautioned that if NIH regulations force data sharing as a condition for a grant, some research involving foreign countries or agencies may not occur because of concerns that they may lose control over the data generated from the study.
- ❖ NIH should provide plain language guidance for investigators about expectations for future data sharing of the results of human studies. In addition, it may be helpful for NIEHS to develop guidelines specific to each type of data sharing (voluntary, involuntary, and voluntary non-collaborative). Investigators should include data sharing plans and data management efforts as well as standard environmental measures to facilitate cross-study analyses whenever possible into their grant proposals from the beginning.
- ❖ NIEHS should promote the use of electronic health registry information linked to disease outcomes (e.g., Kaiser Permanente’s electronic registry), which are currently underutilized resources that could enhance certain environmental health science investigations.

- ❖ NIH Institutes and Centers should encourage the incorporation of key standardized elements and measures for environmental exposure data. Further, NIH should encourage investigators to use core elements and standard measures to assess exposures in the same way to allow for data variables to be more sharable. Funding studies that develop a consensus on core elements (e.g., PhenX extension) would allow data sharing and cross-study analyses without implementation of a data-coordinating center for each new consortium or large project. The importance of standardized measures, protocols, and vocabularies was emphasized as particularly important with environmental health data to allow study results from larger population studies to be pooled or utilized for meta-analysis. Consistent core elements may allow identification of subtle interactions and increase statistical power for large scale G x E interactions.
- ❖ NIEHS should place a strong emphasis on cross-training future scientists to promote multidisciplinary research encompassing the fields of computer science, bioinformatics, engineering, epidemiology, and environmental health sciences.
- ❖ NIH was also encouraged to play a bigger role in supporting environmental sample banking, tracking, and long-term storage of biological samples. It was suggested that a support mechanism to allow long-term storage of biosamples that does not depend on short term grant funding (e.g., Coriell Institute's repository for DNA and cell lines) would further advance data sharing.
- ❖ Along with promoting collaborations, NIEHS should curate and distribute a list of ongoing data projects in the environmental health science field for which secondary analysis is possible as well as a list of people who might be interested in conducting such projects.
- ❖ NIEHS should address the fact that longer embargos (delayed release) might be needed with environmental data sharing for studies with extensive community involvement so that results can be presented to a community before they are posted, further shared and reanalyzed, or published.

Computational Challenges:

Although the meeting did not have a session devoted exclusively to computational challenges associated with data sharing, this topic repeatedly came up in discussions. Sufficient hardware, software, and general cyber-infrastructure resources to handle an unprecedented volume and complexity of data was stressed by many RFI responders and workshop participants as an overriding concern for many data types, including environmental health science data. Data

sharing will only increase the complexity and size of datasets. Several researchers stressed that analysis, not data creation, will be the fundamental hurdle preventing further advances in the field of environmental health. Concerns were voiced that even the most popular bioinformatics tools will not be able to scale to the level needed for large biological network interactions. Several investigators emphasized the importance of large parallel data analysis tools that depend on a distributed data sharing networks as well as cloud (or grid) computing cyber-infrastructure as emerging systems to consider. There is a strong need for new, high-performance computational tools and approaches with massive storage capabilities to accommodate the mining, pooling, and analysis of multidisciplinary environmental health science projects. This topic is not unique to NIEHS—NIH has recently been asked to consider cloud computing applications for many databases it supports—and some of the key recommendations related to computational challenges expressed at the workshop and in RFI responses include:

- ❖ The importance of creating searchable data websites, databases, data and sample repositories, and/or registries.
- ❖ NIEHS should require the establishment of data sharing centers for many of the larger research efforts. Many considered essential a new requirement for a central coordinating center for large, multi-site studies. The studies would be required to release their primary and secondary data into a centralized Web-based database, thus allowing consistent database management across institutions and agencies. Establishing dedicated independent centers of analytic and data processing excellence may be preferable to wide distribution of data sets containing subject-level data.
- ❖ NIEHS should encourage efforts that integrate environmental health data into data-sharing platforms in other science fields and with other diverse or disparate datasets (e.g., genomic datasets) with common shared vocabularies or standardized measures as key components. This integration is considered a critical step for more effective and widespread use of environmental data across diverse scientific disciplines.
- ❖ NIEHS should also encourage secured (controlled) access to uniformly collected (using standardized measures) pooled datasets. This would allow a variety of users with different levels of access permission to utilize different subsets of data (e.g., the National Database for Autism Research).

Conclusion

A majority of researchers, community participants, and other stakeholders were positive about the possibilities and opportunities that broader data sharing might bring to environmental health science research. Untapped potential for

investigations using environmental and epidemiological datasets could be developed through rapid and widespread data sharing. Leveraging existing datasets and models was a consistent theme in both the RFI responses and workshop. Examples of successful approaches of data sharing strategies were discussed extensively. The need to promote and support collaborative networks of researchers who are open to new technologies, methodologies, and resources and their application to specific datasets was also emphasized.

NIEHS' recently developed strategic plan incorporates many of the recommendations in this report into future NIEHS initiatives and efforts. Specifically, Goal 7 of the NIEHS strategic plan encourages promoting and maximizing data sharing and collaboration among environmental health scientists and identifying strategies that support greater data sharing while recognizing the unique sensitivities of environmental exposure information. NIEHS is exploring mechanisms and ways to provide infrastructure and increased support for biorepositories, cohorts, and datasets. NIEHS is investigating efforts to further enhance awareness and broader use and utility of environmentally relevant databases, such as the Comparative Toxicogenomics Database (CTD), CEBS, and NDAR. The Institute is also exploring the extension of the PhenX Toolbox exposure measures to improve their use in environmental health sciences and will continue to promote the development and application of common exposure ontologies across databases and datasets. NIEHS is committed to providing guidance and examples of templates of minimal elements for investigators to include in a data sharing plan in their grant applications.