

Summary of RFI Responses for Data Sharing Strategies in Environmental Health Sciences Research:

A RFI (request for information), entitled “Input on Strategies to Encourage Broad Data Sharing in Environmental Health Sciences Research”, was released in the NIH guide on June 3, 2011. The purpose of this request was to gather information and recommendations from the environmental health science research community regarding successful approaches and strategies that allow broad data sharing in the field of environmental health sciences in human population studies. NIEHS recognizes that environmental health science research is becoming increasingly complex and multidisciplinary. Therefore, the broad sharing of data generated from epidemiological studies is highly desirable to leverage the NIH investment in these studies and advance the field of environmental health sciences as a whole. NIEHS was particularly interested in: what unique considerations exist for data sharing for studies with environmental exposure data, what challenges or barriers exist for researchers wishing to more broadly share their data with others, and what additional tools or resources do researchers believe will allow more efficient and effective data sharing in the environmental health science community.

The following report attempts to capture the key suggestions and concerns of the researchers and other community stakeholders in environmental health sciences that responded to this RFI. The recommendations and comments addressed in the responses received to this RFI were classified into nine broad themes or categories for the purpose of summarizing these ideas.

Protection of Privacy/Confidentiality Issues:

Data sharing in the context of environmental health sciences may present some unique challenges related to the protection and confidentiality of the study participants (ex. environmental monitoring, disclosure of exposure information at the county and neighborhood level, etc.). Several stakeholders conveyed the importance of community-based participatory research with thorough community consultation and consent for all use and disclosure of data. This can be particularly important in vulnerable communities where more environmental health research related to some environmental exposures is likely to reside. Providing access to individual data to other research groups may be especially problematic in small or uniquely exposed populations because only one or few individual pieces of personal information will be enough to positively identify a research participant. 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 (ex. reporting of lead paint exposures by specific locations to departments of health). Researchers suggested various solutions to the general problem of how to “anonymize” the environmental exposure datasets yet allow enough of the data to be shared to allow for useful, appropriate secondary data analyses while protecting individual identifiers and confidentiality. Some options include allowing a research participant or patient to be

identified across projects and databases without identifying any personal information with a “unique subject identifier” or archiving data to a separate external repository with personal identifying data stored separately and requiring special permits to access. The use of homomorphic cryptography (particularly with respect to geospatial data and place of residence information) was also mentioned as a newly evolving technology that may be used to protect patient/participant confidentiality while sharing environmental data more broadly.

Many researchers stressed the importance of greater security measures to protect participant/patient confidentiality as well. It was pointed out that the possibility of re-identification of participants is entirely possible in many studies given that traditional approaches for protection of research subjects are inadequate for online databases. One stakeholder suggested that the many risks related to breaches in privacy and confidentiality from re-identification are in fact not well understood for environmental exposure data, and this is an area that should receive further research in itself. Many suggested that the security of analysis and data platforms, transmission procedures, and the role of firewalls will need to be examined much more carefully. Several research groups also mentioned that more emphasis should be placed on adequate training of personnel, especially for those individuals who will be the gatekeepers or managers of databases populated with environmental health science data that will be shared across investigators.

Institutional Review Board (IRB) issues:

The lack of continuity or consistency across IRBs for issues related to participant consent and sharing of data was emphasized by many research groups as a disincentive for attempting to more broadly share their data with others. Researchers noted the lack of clarity in the IRB oversight and consent processes. There was a general sense that this ambiguity impedes researchers from rethinking informed consent models (exploring open consent or marketplace consent models, etc.) that might allow data results to be compared, pooled, or analyzed more broadly among research teams. It was also pointed out that individual level data may not be available to be released under current IRB consenting guidelines, and some medical data especially, protected under HIPPA (Health Insurance Portability and Accountability Act) requirements, may be restricted from sharing. Currently each individual study must strictly adhere to the guidelines of the requirements of their specific IRB. With clinical trials, researchers must follow the data safety monitoring boards’ recommendations under the guidelines from the FDA as well. In addition, many requirements lead to removal of all personal data identifiers to the extent that secondary data analysis is no longer feasible for other researchers. One leader of a community based organization stated that community-based organizations may be disadvantaged in accessing data because they lack their own IRBs or appropriate representation on existing IRBs. Finally, several investigators pointed out that multiple IRBs may be involved for data sharing of a study to occur, and IRBs may not accept each other’s decisions regarding which data could be shared and how.

Legal Issues:

Many particular concerns related to legal (with underlying financial and social) implications of environmental exposure data were expressed as an impediment to broad data sharing. Legal issues related to intellectual property handling, licensing, as well as nondisclosure, proprietary, and disclosure agreements were mentioned as general concerns related to broad data sharing in the biomedical community. However, most of the emphasis in this topic area highlighted some of the unique legal concerns that may occur related to sharing of environmental health sciences data. Regulatory reporting or remediation requirements related to data containing specific environmental exposures in human population datasets appears to be a particular issue. Exposure data will continue to be of high interest to regulatory agencies with respect to the evaluation of the health implications of chemicals. Several NIH-funded researchers expressed concerns that industry or private groups could spend considerable effort and resources to reanalyze and reinterpret data obtained by NIEHS grantees in an effort to delay regulatory reform. Industry or other privately funded studies would not fall under the same guidelines of publicly releasing and sharing data, which may make “a very uneven playing field”. One suggestion on how to address this issue would be to require that investigators funded by industry or private studies need to make their data available to the scientific community if requesting similar data from an NIH-funded investigator. 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 what is now present with genomic data under GINA (Genetic Information Nondiscrimination Act) legislation, was also pointed out as a unique concern when considering broadly sharing datasets containing information on unique exposures.

NIH Programmatic Considerations:

Many investigators weighed in on possible ways that NIH might jumpstart data sharing possibilities in the environmental health sciences communities; the creation of searchable data websites, databases, data and sample repositories, and/or registries were a top suggestion of many. Many research groups would like to see NIEHS funded investigators submit their primary data to a searchable centralized repository that everyone can access (ex. dbGAP) with a description of main factors measured and biologic samples collected. Researchers also suggested the creation of sample repositories (listing available plasmids, cell lines and tissues, animal lines and model organisms) and chemical repositories of environmental toxicants that researchers in the health sciences community can freely access. This seemed to particularly resonate with young investigators in the environmental health sciences community who would like access to epidemiologic cohort/case-control study datasets and know which studies are out there and which samples were collected in those studies. Several investigators noted that NIH does not have many mechanisms available to fund researchers to develop extensive databases that can be non-trivial to construct and maintain but incredibly useful to the scientific community as a resource. Several investigators also

expressed the desire for the development of registries for incident cases of diseases/disorders other than cancer, such as Parkinson's, autism, autoimmune disorders, reproductive disorders, etc. in which environmental exposures may be key risk factors (with the national Alzheimer's Coordinating Center suggested as a nice model for this). One researcher recommended that the ideal scenario might be the establishment of regional NIEHS centers that focus on these disorders, from which qualified investigators might obtain access to clinically diagnosed cases and controls without the costly and time-consuming burden of new human subject recruitment, clinical diagnoses, and confirmation in individual studies. The NCI SEER (Surveillance Epidemiology and End Results) System was mentioned as an excellent resource example that provides confirmed diagnostic information on all incident cancer cases as well as identification of some environmental exposures and other information from study participants and the capability of access to numerous cancer registries for multi-site studies, collaborations, and pooling efforts. 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 which does not depend on short term grant funding (an example perhaps being the Coriell Institute for DNA and cell lines) would go a long way to further advancing data sharing.

NIEHS was urged to further advance data sharing efforts by requiring the establishment of data sharing centers for many large research efforts and encouraging investigators to include data sharing costs, plans for sharing of data, and data management efforts into their grant proposals from their inception. Many saw as essential the requirement of a central coordinating center or data management center for larger programs that focus on a specific disease or scientific area. Investigators funded under these programs would be required to release their primary and secondary data into a centralized web-based database that would allow consistent database management across many institutions and agencies for collaborative projects and allow uniformly collect pooled datasets to be securely accessed to many users with different levels of access permission to different subsets of data (an example being the National Database for Autism Research, NDAR). Several responders thought that the lack of incentives for more broadly sharing data was something that NIH should address, particularly for young investigators just starting out. Several investigators suggested that NIH address data sharing costs of projects in funding announcements or initiatives and include these requirements in the grant award. The procedures being put in place for data sharing and management of the National Children's Study was proposed as an example to follow. Other standard data sharing/data management templates that were suggested as exemplary examples include those from NSF and CUAHSI (Consortium of Universities for the Advancement of Hydrologic Research, Inc). Research groups also commented on the extensive data management support that is needed to properly harmonize exposures and/or phenotypes/outcomes, as well as methods of exposure measurements or modeling across studies or cohorts. One researcher cautioned that "issues related to harmonizing the methods used for data collection and modeling should not be underestimated". In his own experience of data pooling, he has invested

much more time and resources for data harmonization than in any other stage of the project.

Computational Challenges:

Many concerns regarding sufficient hardware, software, and general cyber-infrastructure resources to handle an unprecedented volume and complexity of data were stated as well. To quote one researcher, "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 be unable to scale to the level of complexity needed for large scale biological network interactions. 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. More multidisciplinary, cross-trained researchers in computer science, bioinformatics, engineering, epidemiology, and environmental health sciences will be needed to support these efforts. One researcher stated that establishing and maintaining the cyber-infrastructure will require a fundamental paradigm shift in the way scientists think about study design, collaborations and data pooling, analysis, and archiving. Several investigators emphasized the importance of massively parallel data analysis tools that depend on distributed data sharing networks as well as cloud (or grid) computing cyber-infrastructure as emerging ideal systems to work towards. Several biomedical cyberinfrastructure efforts that were hailed as examples include caBIG and the CDC's National Environmental Public Health Tracking System. More specifically, the caBIG cyberinfrastructure has allowed the sharing of patient data in partnerships among academic institutions, contractors, industry, and government agencies in order to integrate extensive individual patient data. More complex cyber-infrastructure will aid in the development of both new spatial pattern identification tools and more efficient study designs and analysis with improved visualization.

Many responders to this RFI also stressed the importance of integrating environmental health data into data-sharing platforms in other science fields and with other diverse or disparate datasets. This was seen as critical for environmental data to be used most effectively across many scientific disciplines. Since an integrated cyberinfrastructure incorporating common data types across many diverse datasets will be key, environmental health scientists are encouraged to work closely with computer and bioinformatics scientists, social scientists, legal and ethical experts, and community stakeholders to ensure that study designs and data sharing platforms in diverse science fields are designed to address the unique problems of environmental health science from the beginning. One stakeholder suggested that data integration efforts could also be built around specific challenges (ex. NIEHS Bioassay Network) or the data integration and visualization platform could be built in the context of a specific disaster response (ex. data portal created for Gulf Coast's Katrina-Rita Hurricane). In addition, integrated environmental health data in epidemiological and clinical studies need to be collected using comparable or standardized methods, protocols, vocabularies, and

measures to allow the combined data to be useful. The importance of standardized measures was emphasized as especially important with respect to environmental health data in order to allow larger populations to be pooled or utilized for meta-analysis which will allow more subtle interactions to be identified and substantially increase statistical power for large scale G x E interactions. A useful paper outlining the features for successful translational cyber-infrastructure, "Building a Biomedical Cyber-infrastructure for Collaborative Research", included many of the points expressed by other researchers responding to this RFI and illustrates many examples of ongoing research in this area. Two well-established examples of open-source systems establishing standard measures and systems that use electronic medical records include PhenX (consensus measures for phenotypes and exposures), and the Open Geospatial Consortium (OGC). The NIH-supported PhenX Toolkit is a catalog of standard measures for large-scale genomic research efforts and the RTI Spatial Impact Factor Database (SIFD) is an extensive repository of geo-referenced data that conforms to OGC standards. Examples of other collaborative projects that have successfully attempted to standardize measures for electronic medical records and/or epidemiology or clinical data include: the Public Population Project in Genomics (P3G), the Genomics and Randomized Trials Network (GARNET), Gene Environment Association Studies (GENEVA), electronic Medical Records and Genomics (eMERGE) Network, Patient-Reported Outcomes Measurement Information System (PROMIS), and the Grid-Enabled Measures (GEM) database. One cautionary note mentioned regarding data platform standardization is that in some scientific fields, the technology is rapidly evolving and changing such that it may not be prudent to expend large amounts of energy into standardizing platforms across studies if one predicts rapid technological advances. An example of this is in the field of epigenomics where, due to rapid turnover in technological advances and platforms, data generated just a year or so ago might be hard to pool and analyze with present data.

Standard Operating Procedures for Data Sharing:

Many researchers encouraged the use of standard guidelines regarding the standard operating procedures and protocols that would be needed to broadly share environmental health data. These include the establishment at the beginning of studies of all rules for data access, release, publication embargos, data validation, and quality control issues that all need to be carefully controlled and executed. There was a strong consensus that a multi-disciplinary ancillary study committee needs to carefully oversee access to datasets and biospecimens. There needs to be consistent curation of data and plans for long-term storage and archiving. The importance of full disclosure of all data that is necessary for replication was emphasized (including all secondary or ancillary data). Special plans may be necessary from the beginning of environmental health studies to ensure that data can be pooled or shared, such as consistent collection of environmental samples under similar conditions and the appropriate handling of biohazards for some environmental exposure measurements. In a few cases where there are valid reasons that original datasets cannot be released in a timely manner, one prominent biostatistician suggested that researchers provide a

pseudodataset obtained by sampling with replacement from the original dataset to allow simulations and other valuable research related to analytical approaches for G x E interaction discovery to still occur.

International Study Data Sharing Issues:

A number of comments related to the particular challenges or additional hurdles to successfully sharing environmental datasets when international collaborations or foreign countries are involved. Researchers warned that international collaborations involving multiple foreign institutions can be incredibly complicated and a “one size fits all” recommendation or guideline from NIH regarding data sharing could be quite counter-productive. Researchers cautioned that if NIH regulations force the sharing of data as part of an NIH grant, some research involving foreign countries or agencies may not occur because the foreign country may not be willing to collaborate or allow the study to be initiated due to concerns that they may lose control over data generated from the study later. In some countries, submission and access of patient datasets can be governed by national legislation. In addition, IRB rules have not allowed some patient data collected in international studies to be used beyond the original study, even if properly de-identified. One approach that has been used to partially circumvent this problem is the development of an on-line analytical tool with standard measures and vocabularies that is being used to combine datasets across a number of different countries with vast National Health Registries, including Denmark, Sweden, Israel, Finland, and Australia.

Key Areas of Science Where Data Sharing Is Particularly Useful:

Gene-Environment interactions: Due to statistical power issues, researchers pointed out that most large scale G x E interaction studies will need to rely on pooling of smaller studies or meta-analysis of cohorts. A prospective study of G X E interaction with an “omic” discovery phase was promoted by one researcher as the ideal way to discover new genes related to exposure and can only be done with extensive data sharing: “NIEHS has funded several birth cohorts with similar phenotype measures in children, DNA collection, and exposure data. Genotyping methods and designs in omics are largely worked out. Two stage designs with a discovery and multiple replication populations are standard in genomics and should be applied in G X E interactions. The critical issues with respect to G X E interaction relate actually primarily to exposure and are a) exposure data need to be prospective and therefore predate the phenotype, and 2) the exposure data need to reflect exposures during critical developmental windows rather than cumulative data or cross-sectional data, which is all that can be accomplished with the standard case control design. A planned pooling of birth cohorts funded by NIEHS specifically for G X E interaction could be a major advance in the field. Phenotypes that could be studied in such an approach include asthma, neurodevelopment, and birth outcomes among others. Exposures could include lead, Hg, pesticides, BPA/Phthalates among others. Many cohorts have archived urine or blood and can add exposure data from critical developmental windows to existing phenotype data, thereby joining the pool. Some cohorts are already funded to do GWAS

and could cost effectively serve as a discovery cohort for G X E interactions. This would greatly reduce costs and also avoid duplication of effort...This approach would like advance the field rapidly and in a cost efficient manner.”

Medical Health Records and Datasets: A number of stakeholders and environmental health scientists expressed the opinion that electronic health registry information linked to disease outcomes was a greatly underutilized resource that could advance many areas of environmental health science. Specifically, recommendations that there be better utilization of Kaiser Permanente datasets and direct to consumer genomic testing companies’ population datasets were cited. Of particular interest is Kaiser Permanente's research program on Genes, Environment, and Health, which has a planned release of patient electronic health records (containing genetic, medical, and environmental data) on 100,000 people in the upcoming year. This resource and others like it can provide researchers with ready access to well-characterized large populations that would be prohibitively expensive to develop from scratch today. Direct-to-consumer genetic companies are also a greatly underutilized valuable resource for researchers wanting genomic and phenotypic data on extensive populations. As an example, one researcher mentioned that “23andMe” is using its database to do research studies on several diseases and is partnering with Genentech to use its patient population data to advance Alzheimer's research. Collaborations between academic institutions and large pharmacy companies and/or managed care organizations are becoming more accepted and feasible as well, and several researchers felt that the environmental health science community had yet to fully embrace these opportunities.

Mixtures: Mixtures of chemicals or exposures is a particularly vexing environmental health problem that might be addressed as well by carefully planned sharing of datasets or pooling of cohorts. The timing and doses over time in large population studies will be needed to study mixture effects for disease outcomes and phenotypes.

Occupational Studies or Uniquely Exposed Populations: In some cases, a higher exposure of a particular toxicant is only observed in isolated populations or occupational studies which can be too small in size alone to generate statistically significant results. The importance of properly sharing and pooling data in these studies and to allow G x E interactions to be confirmed and replicated will continue to be an important area that needs attention.

Scientific Integrity/Ethical Issues:

The most often stated concern for broad data sharing related to scientific integrity is the potential for subjective re-analysis of epidemiological datasets to “prove” specific hypotheses or inappropriate use of datasets by investigators unfamiliar with the details of the population study. Many datasets may be intentionally or non-intentionally used inappropriately without the full knowledge of how the data were collected, collated, and initially analyzed. If data is released publicly and there is no active collaboration and discussion of detailed nuances of the study population with any of the original

researchers of the study, misleading secondary analyses can be performed and published. One researcher put it more strongly as a cautionary tale against “Wikidata” whereby broad and loose data sharing guidelines can allow lots of valuable epidemiological data to be misused and falsely interpreted to the public. Other general concerns relate to the necessary blinding of datasets to remain unbiased and how this might be compromised by early data sharing requirements.

In Summary:

The majority of researchers, community participants, and other stakeholders were overwhelmingly positive about the possibilities and opportunities that broader data sharing for environmental health science data might bring. There seemed to be a general feeling that there are untapped potentials in environmental and epidemiological datasets that could be more thoroughly exploited through rapid data sharing. NIEHS specifically asked what works in addition to what doesn't work with this request for information, and many researchers gave specific examples from their own research experiences of models of what made data sharing easier, more efficient, or timelier. A strong theme was the leveraging of existing datasets and models, particularly to expand environmental epidemiology studies which collect exposure information but not detailed clinical information and genetics, and vice-versa. The promotion and support of collaborative networks of researchers that are open to new technologies, methodologies, and resources that could be applied to specific datasets was also emphasized.