Principles and Obstacles for Sharing Data from Environmental Health Research—Workshop in Brief

On March 19, 2014, the Roundtable on Environmental Health Sciences, Research, and Medicine held a 1-day workshop titled “Principles and Best Practices for Sharing Data from Environmental Health Research.” The purpose of the workshop was to explore key concerns, principles, and best practices to foster the responsible sharing of data used in support of environmental health research and decision making, while understanding how to protect the privacy of participants and addressing the concerns of the research community. The workshop was divided into the following areas, which included brief presentations from invited speakers followed by in-depth discussions among the speakers and individual audience members:

- Overview and brief legal history,
- Benefits and importance of data sharing,
- Challenges associated with data sharing,
- Potential ways forward, and
- Reflections on the workshop.

This brief summary of the workshop highlights the dialogue that emerged from the individual speakers’ presentations and discussion sessions that followed, and it should not be seen as conclusions or recommendations from the workshop. Statements, recommendations, and opinions expressed are those of individual presenters and participants and are not necessarily endorsed or verified by the Institute of Medicine (IOM) or the Roundtable, and they should not be construed as reflecting any group consensus. A full summary of the workshop that provides more detail of the presentations and discussion that ensued will be available in 2015.

Overview and Brief Legal History

In thinking over her career as an epidemiologist, Lynn Goldman, Dean of the Milken Institute School of Public Health at the George Washington University, briefly described two instances where her data had been requested. She explained that, regarding her experience, sharing data has not been a particularly difficult issue but one that can be quite contentious when occurring in the context of litigation. She noted that the Shelby Amendment, passed in 1998, added a requirement that federally funded research data be made available to the public under the Freedom of Information Act (FOIA) in response to concerns over the Harvard Six Cities Study (a prospective cohort study investigating the health effects of air pollution). Congress further acted in 2001, she said, with the Information Quality Act that resulted in guidelines by the Office of Management and Budget (OMB) for how agencies respond to requests for data.
Paul Verkuil, Chairman of the Administrative Conference of the United States, provided information on the federal rule-making process (where federal agencies write regulations to implement or interpret law or policy) from a legal perspective. After publication of a final rule the legality of the rule can be challenged in court. He explained that the courts do not review a rule as if they were scientists because of limitations in how they can analyze the data, but they do try to ensure agencies have behaved rationally in light of the available data. This is especially important when reviewing the science involved in considering identified uncertainties and alternative courses of action with a rule-making decision. He explained that the more important a study is to a particular decision from an agency, the more important it is to provide the underlying data.

George Gray, Professor in the Department of Environmental and Occupational Health at the Milken Institute School of Public Health at the George Washington University, described policies put forward by the executive branch related to data sharing. He explained that Circular A-110 (the circular that provides standards for federal agencies in the administration of grants and agreements with institutions of higher education, hospitals, and other nonprofit organizations) was amended by OMB in response to the Shelby Amendment to obligate agencies to share research data from scientific assessments resulting from grants and agreements when used by agencies in support of action that has the force and effect of law. (The question of “what the science is ultimately used for” is important, he said.) Gray stated that the place where he sees the greatest challenge in sharing data is when trying to understand how information influences decisions, how decisions are made, and how alternative assessments may look.

During the discussion session, Ellen Silbergeld, Professor at the Johns Hopkins Bloomberg School of Public Health, asked the speakers, “what do you mean by ‘data’?” She noted that there needs to be more consideration of the word “data” to understand at what level to start sharing (when something is measured, when something is analyzed, etc.). Goldman also noted the need to have an agreed-upon understanding and use of the terms reanalysis, replication, and reproducibility when discussing data sharing.

**Benefits and Importance of Data Sharing**

Francesca Dominici, Professor of Biostatistics and Senior Associate Dean for Research at Harvard University School of Public Health, explained that a spectrum exists with sharing data: Publication is at the low end, publication of the underlying codes and data are in the middle, and full replication of a study is at the highest part of the spectrum. She noted that it is extremely hard to achieve full replication in environmental health research—large cohorts are difficult to replicate because of time and expense, and much of the raw data are confidential. But reproducible research—reproducing the most important data—is achievable and an area where tremendous progress can be made.

Louis Anthony (Tony) Cox, Chief Science Officer at NextHealth Technologies, stated that there are two main benefits to increased data sharing: (1) a greater trustworthiness in published results and in actions taken based on them, and (2) more frequent follow-up to increase the value from investments already made to assemble good datasets. He noted that making key data used to support decision making widely available has important potential benefits in improving both original research and follow-up research.

However, there are questions around whether data may be re-identifiable, as highlighted by Julia Brody, Executive Director of Silent Spring Institute, especially when linking household air, dust, or biomonitoring data and personal information such as home characteristics and consumer product purchases. Brody noted the benefits of sharing data in a U.S. Environmental Protection Agency (EPA) online exposure database (ExpoCast) but listed three problems that her institute is trying to solve in order to make decisions about sharing environmental data: (1) what is the empirical knowledge specifically relevant to re-identification in environmental health datasets?; (2)
what are the potential technical solutions to masking data, synthesizing data, or creating server systems that would address re-identification; and (3) what can and should investigators promise to study participants in informed consent?

The IOM Committee on Strategies for Responsible Sharing of Clinical Trial Data\(^2\) is tasked with investigating responsible sharing of clinical trials data, explained Bernard Lo, chair of the committee and President and CEO of the Greenwall Foundation. He noted that findings, conclusions, and recommendations will come in the final report where the charge is to analyze the benefits, challenges, and risks of various models of data sharing and to make recommendations to enhance the responsible sharing of clinical trials data.

During the discussion session, the speakers elaborated on the benefits that would result from data sharing. Cox stated that the greatest benefits from data sharing are likely societal benefits and their externalities, where one dataset can be analyzed many times for new purposes or unexpected insight. Dominici noted that one of the main benefits of reproducible research is that it elevates the scientific rigor of the discussions around the research question. Another benefit, she said, is the speed of scientific discovery in allowing one investigator to start where another left off.

The individual speakers also highlighted some difficulties associated with data sharing. Gwen Collman, Director of the Division of Extramural Research and Training at the National Institute of Environmental Health Sciences, noted that it can be challenging to communicate all the details and nuances of how epidemiologic or clinical studies were done and how the data variables were created. Dominici agreed and explained that much improvement can be made to the methods sections of published studies (even if the information is provided as supplemental material). In response to comments on informed consent, Lo noted that the old model of consent does not work today and possibly never worked. In looking back at consent forms from drug clinical trials, he said, some trials gave broad consent in terms of sharing data with other researchers but participants may not have known what this meant and what was covered. Lo stated that consent and privacy may bear too much weight in research studies—by implying that researchers will only share what a participant consents to, and that privacy will be protected in strict terms, may be promising too much and may be unrealistic.

Challenges Associated with Data Sharing

The challenge of sharing de-identified data and possible re-identification of study participants was elaborated on during this session. Daniel Barth-Jones, Assistant Professor of Clinical Epidemiology at Columbia University, noted that de-identified data provides invaluable public good and is an essential tool in supporting scientific innovation and research. But, when data are de-identified they are degraded in different ways, either through overt information loss (like restricting dates that are more specific than a year) or through the tradeoff of grouping people together (so they are not observable as specific individuals). He stated that the competing goals of privacy protection and preserving the utility and statistical accuracy of de-identified data needs to be balanced.

Latanya Sweeney, Professor of Government and Technology in Residence at Harvard University, noted that data de-identification is a necessary condition for improving privacy protection. However, real world demonstrations of actual vulnerabilities and risks of re-identification are needed to develop policy framed around scientific improvement in privacy enhancing technologies. She stated that more needs to be learned to improve the privacy protections appropriately, whether they are made weaker or stronger, but more likely they will be totally different than what is known today.

Kevin Casey, Associate Vice President for Public Affairs and Communications at Harvard University, noted that when data intersect public policy it can become a complex controversial issue for not only the researchers but for the study participants, policymakers, and politicians. He stated that a number of things make sharing data from the Harvard Six Cities Study particularly complex. First, he said, the disclosure agreement gives specific assurance that the participants’ identities and relationship to any information will be kept confidential. Second, three sets of data are involved in the study: (1) the participants’ histories and what happens to their respiratory systems,

---

\(^2\) More information on the IOM Committee on Strategies for Responsible Sharing of Clinical Trial Data is available at http://www.iom.edu/Activities/Research/SharingClinicalTrialData.aspx.
(2) their location, and (3) their death records. He explained that this combination of data makes it quite easy to identify any participant despite de-identification efforts.

Turning to challenges that industry faces, Greg Bond, Dow Masters Fellowship Program, explained that a fundamental administrative challenge is who owns the data. He noted that this is especially true when companies are absorbed into others via mergers and acquisitions or when companies go out of business. The financial challenge associated with protecting data of commercial value is another administrative issue that industry considers. Despite these challenges, Bond stated that industry does believe that all relevant health and environmental information—not necessarily the raw data, but the information to take prudent action—should be publicly accessible.

Glenn Paulson, Science Advisor in the Office of the Administrator at EPA, noted that Bond drew the distinction between data and information and asked him to share any goals or objectives that industry is working on for information sharing. Bond noted that a large amount of case-by-case decision making will have to be made, but a set of industry principles could help provide guidance. Bond explained that sharing information at a sufficient level for people to take action (product labels, material safety data sheets, etc.) should be improved. He explained that this is information as opposed to the underlying data on which the information was based. Linda Birnbaum, Director of the National Institute of Environmental Health Sciences, noted that EPA and the European Union REACH program (the regulation on Registration, Evaluation, Authorisation, and Restriction of Chemicals) are dealing with challenges because they cannot regulate on the basis of robust summaries and the material safety data sheets, which often do not have all the information that people need or want. Bond stated that there are probably ways of finding common ground and agreed that the material safety data sheets and product labels need improvement.

John Howard, Director of the National Institute for Occupational Safety and Health (NIOSH), stated that a number of lessons around data sharing have been learned at NIOSH. First, he said, researchers need to think about the optimal data-sharing practices at the study concept stage to balance the research mission and the ability to share data for reanalysis. Second, research budgets should prospectively include sufficient resources for the investigators to implement robust data-sharing plans. Third, it is difficult to enforce specified restrictions on data use, and recognizing these limitations, data use agreements should clearly state what happens in the case of non-adherence. Fourth, one way to protect confidentiality while providing access for analysis of highly sensitive and potentially identifiable datasets is to host the datasets in secure enclaves where only non-identifying aggregate analyses can be removed. Fifth, reanalysis should be based on a strong, reproducible foundation, and only verified and finalized datasets from completed and published studies should be shared. Sixth, specific communications about reanalysis should take place with study participants at the time of enrollment so they are aware of the possible disclosure parameters in the study.

Paulson asked Howard to elaborate on the secure enclave model he proposed to provide privacy protections. Howard stated that NIOSH is working with some pioneering models but there are few practices in this area and much remains to be seen. Daniel Greenbaum, President of the Health Effects Institute, and Al McGartland, Office Director of the National Center for Environmental Economics at EPA, noted that the National Center for Health Statistics has protocols and procedures for protecting the identity of data, including secure mechanisms for analyses of National Health and Nutrition Examination Survey (NHANES) data.

Stacy-ann Allen-Ramdial, Intern with the House Committee on Science, Space, and Technology, asked the speakers to comment on the risk of losing study participants when providing increased transparency on how their information could be used by the federal government for regulation. Casey stated that this is the balance researchers are trying to strike, but there could be a chilling effect on enrollment when personal information might become part of the public record. Howard noted that in order to get to this stage of enrollment an education process would have to occur for scientists to be able to make the appropriate promise to a potential study participant based on today’s world. Casey agreed and stated that the scientific community needs to come together and begin the

---

3 The main aims of REACH are to ensure a high level of protection of human health and the environment from the risks that can be posed by chemicals, promotion of alternative test methods, free circulation of substances on the internal market, and enhancing competitiveness and innovation. More information is available at http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm.

4 NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. More information about NHANES data is available at http://www.cdc.gov/nchs/nhanes.htm.
conversation of how to enroll people going forward. Goldman elaborated on this point and questioned how lack of confidentiality will impact collaborations with researchers around the world who may be hesitant to sign data-sharing agreements knowing that their data could be made freely available if U.S. federal funding is involved.

**Potential Ways Forward**

Edward Sondik, former Director of the National Center for Health Statistics, explained that the agency has two main roles. The first is to provide information for policy and research on the health care system and the health of people in this country as compared with other countries. The second is to ensure and achieve widespread dissemination and access to the wide variety of surveys the agency coordinates (e.g., NHANES and the Health Interview Survey). The agency prohibits the release of potentially identifiable data under the Confidential Information Protection and Statistical Efficiency Act, he said, and there needs to be a balance between access to the information and the issue of preserving confidentiality. The agency creates public use datasets, which are de-identified, to safeguard the identity of the people who supply the data. The agency also created a research data center where people can rework the datasets. He stated that estimating the probability of disclosure remains an unresolved issue and federal agencies should work to understand public expectations in today’s world.

Daniel Greenbaum, President of the Health Effects Institute, stated that he would place many of the issues discussed throughout the workshop in a slightly different context that he calls the “adversarial context.” He explained that there is significant science and policy interest in sharing underlying data and cited examples of sharing data from detailed exposure experiments. However, there is an inherent tension between collaborative scientific data sharing and adversarial science that is different from independent reanalysis, he said. He explained that researchers are sometimes engaged by advocates who are not really trying to advance scientific knowledge and whose primary purpose is to undermine the original study. He asked, is there a way to facilitate a true dialogue between those who fund the adversarial reanalysis and the scientific community to set a foundation for more thoughtful and independent testing of key studies? There are a lot of challenges around the technical and privacy issues, but this is really the biggest issue, he said.

Linda Birnbaum, Director of the National Institute of Environmental Health Sciences, stated that a language standard is needed for environmental health research. She noted that the workshop started with “what do you mean by ‘data’?” and “what do you mean by ‘reanalysis’ or ‘replication’ or ‘reproducible’?” Common language for environmental health would foster the interoperability of databases and promote sharing and reanalysis of data to hopefully accelerate the pace of discovery, she said. The National Toxicology Program (NTP) began an effort to develop a new approach to systematic review based on data streams from environmental health studies (including observational human studies, animal data, and toxicology information) that is fully transparent so anyone can follow the processes involved. She noted that two of the key remaining issues are people not meaning the same things when using the same words and how to extract all the data so everyone can have access to it. She explained that NTP is developing a database format where data abstraction files can be made publically available to increase transparency and allow for supplemental or updated analyses of the data. The creation of data repositories of abstracted data is a huge effort, she said, requiring a great amount of time and money to realize.

George Daston, Victor Mills Society Research Fellow at Proctor & Gamble, noted that this is a systems biology era, which promises to challenge environmental health and toxicology arenas with its computational power. But systems biology approaches or computational approaches are incredibly hungry for data, he said. For example, there is a wealth of information on toxicity end points of various chemical structures or substructures but the boundaries could be expanded to better understand chemical structural features and putative modes of action for toxicity of reproductive or developmental toxicants. He explained that by compiling these ontologies for modes of action of toxicity, the field could be organized in a way to allow regulatory agencies to understand more about the tens of thousands of chemicals for which there is not much information. He stated that from his perspective, sharing of data is more valuable than retaining the data, and, as others noted, the most important question is, “what
do you mean by ‘data’?” As Bond noted earlier, Daston agreed that improved data summary information is needed. He noted that researchers should start with sharing the most refined and processed data they use. There may be instances were raw data is also valuable to share (e.g., genomic analyses) but this should not exclude sharing the refined data, he said.

Frank Loy, Chair of the Roundtable, asked the speakers if they have thought about a system where there would be a way of adjudicating the adversarial approach that Greenbaum discussed and the interests of research and privacy. Greenbaum noted that there is much attention to certain studies, usually because of the policy and economic stakes involved in decision making based on those studies. He stated that it would be helpful to construct rules of engagement that promote a level of civil discourse to enable people to produce quality science, have it challenged by scientists in a scientific manner regardless of who they work for, have an opportunity for open dialogue, and then in the end know something more as a result of that process.

**Reflections on the Workshop**

Linda Birnbaum, Director of the National Institute of Environmental Health Sciences, highlighted some of the main points she heard throughout the workshop. First, she said, there is a wide array of data, including personal exposure measurements and geographic data, that is essential to environmental health research and potentially problematic in terms of re-identification of participants. Second, the costs and efforts to reanalyze datasets are large and, rather than spending resources on reanalysis, it may be better to have more reproduction and replication. Third, there is a real concern for how data sharing will impact participant enrollment in environmental health studies, especially those with smaller cohorts that are followed over time. This could also affect international scientific collaboration if any research conducted with federal funding has to be made publicly available.

Jerry Blancato, Director of the Office of Science and Information Management at EPA, noted that the workshop discussions covered three overlapping spheres. The first aspect relates to ethics and privacy and how to protect participant confidentiality, he said. The second aspect is governance, which involves making data publicly available and the challenges associated with managing both current and archived datasets. The third aspect is technology and identifying technical or business solutions to share the data publicly.

Joseph Rodricks, Principal at ENVIRON, stated that there was much discussion around protecting privacy and suggestions on how to maximize the release of data while minimizing the risk of re-identification of participants. He noted that it would be helpful to review those suggested models and determine if there is a systematic way to look at all those opportunities. He also suggested that government agencies that are sponsoring research on environmental matters work to develop guidelines on best practices for sharing data that can be built into the initial planning stages of any study that may be shared. This guidance would help organize the data management process so that at the end of the study the data are ready to go where needed.

Silbergeld noted that she found it extraordinary at the beginning of the workshop that nobody defined “what do you mean by ‘data’?”. Before moving forward, she said, it would be helpful to define that term with some specificity or at least provide a range of what it could mean. She stated that coming away from this workshop she may never do another environmental epidemiological study because of the inability to guarantee confidentiality to the people who participate in these studies, and she will think twice about engaging colleagues in other countries in joint ventures in which their data could become accessible to adversarial proceedings in the United States. Looking at the bigger picture, she said, environmental science has always been about ensuring that the highest-quality data go into making decisions that have both economic and public health impacts. She noted that ensuring complete access to datasets may not be a way forward to reach that goal most consistently and expeditiously. To improve decision making, she said, the question may be “how much information do you need?” The framework for systematic review may help address these questions.

Frank Loy highlighted a series of points identified throughout the workshop. First, he said, there is a lot at stake here. The ability to do quality research in a way that protects individuals is a challenging issue. “What can be done to make that easier rather than harder?” He stated that the scientific community should try to change the relationship between transparency and privacy and make it more of a science issue than a regulation issue.
Second, he said, the privacy claims need to be as limited as possible. He noted that different levels of burdens were presented and one ought to be careful in how broadly one defines privacy and how broadly or not broadly one defines the ways in which privacy is protected. Third, he said, there are conflicts among various interests in this discussion and the question is “how is that to be resolved?” He closed by stating that the way data sharing is addressed and the way it is framed will be quite influential in moving the discussion forward.
PLANNING COMMITTEE FOR THE WORKSHOP ON PRINCIPLES AND BEST PRACTICES FOR SHARING DATA FROM ENVIRONMENTAL HEALTH RESEARCH*

Ellen Wright Clayton, Vanderbilt University, Nashville, TN; Dennis J. Devlin, ExxonMobil Corporation, Irving, TX; Lynn R. Goldman, George Washington University, Washington, DC; Bernard D. Goldstein, University of Pittsburgh, PA; Suzette M. Kimball, U.S. Geological Survey, Reston, VA; David Korn, Harvard University, Cambridge, MA; Alan B. Morrison, George Washington University, Washington, DC; Jerome P. Reiter, Duke University, Durham, NC; Joseph V. Rodricks, ENVIRON, Arlington, VA; Susan L. Santos, Rutgers School of Public Health, Piscataway, NJ; and Kimberly Thigpen Tart, National Institute of Environmental Health Sciences, Research Triangle Park, NC.

* IOM planning committees are solely responsible for organizing the workshop, identifying topics, and choosing speakers. The responsibility for the published Workshop in Brief rests with the institution.

ACKNOWLEDGMENT: The planning committee benefited from the support of members and staff from the Committee on Science, Technology, and Law of the National Research Council.

DISCLAIMER: This workshop in brief has been prepared by Erin Rusch, rapporteur, as a factual summary of what occurred at the meeting. The statements made are those of the authors or individual meeting participants and do not necessarily represent the views of all meeting participants, the planning committee, or the National Academies.

REVIEW: To ensure that it meets institutional standards for quality and objectivity, this workshop in brief was reviewed by Julia G. Brody, Silent Spring Institute, and Kimberly Thigpen Tart, National Institute of Environmental Health Sciences. Chelsea Frakes, Institute of Medicine, served as review coordinator.

SPONSORS: This workshop was partially supported by the National Institute of Environmental Health Sciences (HHSN263201200074I, Order #33), The Kresge Foundation, Colgate-Palmolive Company, ExxonMobil Foundation, and Royal Dutch Shell. The views presented in this publication do not necessarily reflect the views of the organizations or agencies that provided support for the activity.

For additional information regarding the workshop, visit www.iom.edu/Activities/Environment/EnvironmentalHealthRT/2014-MAR-19.aspx.